République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Ecole Nationale Polytechnique
Département d'Electronique
Laboratoire Signal et Communications LSC

Projet de Fin d'Etudes

En vue de l'obtention du diplôme d'Ingénieur d'Etat

Thème :

Elaboration d'un Système Orthophonique : Cas de la Fente Palatine

Réalisé par :

Mr. Nouadri Alaeddine

Soutenu Publiquement le 19 Juin 2014 devant le jury composé de :

Présidente :	HAMAMI Latifa	Prof	ENP Alger
Rapporteur :	GUERTI Mhania	Prof	ENP Alger
Examinateurs :	TAGHI Mohammed Oussaid	MAA	ENP Alger

Promotion : Juin 2014

REMERCIEMENTS

En premier lieu je remercie Dieu le tout puissant de m'avoir donné le courage et la force pour réaliser ce travail.

Je tiens à exprimer mes remerciements avec un grand plaisir et un grand respect à mon encadreur Mme M. Guerti qui a bien voulu m'encadrer. Je la remercie pour sa disponibilité, son aide, les précieux conseils qu'elle nous a prodigués, ses critiques constructives, ses explications et suggestions pertinentes.

Je remercie les membres du jury Mme. L. Hamami et Mr. M.O. Taghi respectivement Professeur et MAA à l'ENP, pour l'honneur qu'ils me font de juger mon travail.

Mes remerciements vont également à tous les enseignants de l'Ecole Nationale Polytechnique qui ont contribué à ma formation. Qu'ils trouvent ici l'expression de mon profond respect et ma grande considération.

Je remercie également Mrs M. Kabache de l'ISMAS et A. Abed du Kleaa, pour leur aide.

DEDICACES

A la mémoire de mon grand père et ma grande mère

A mes très chers parents qui ont toujours été là pour moi,

Vous avez fait plus que des parents font pour que leurs enfants suivent le bon chemin dans leur vie et leurs études.

Je vous dédie ce travail en témoignage de mon profond amour.

Puisse Dieu, le tout puissant, vous préserve et vous accorde santé, longue vie et bonheur.

A mon frère Raouf, mes sœur Bouchra, Yousra et Malek

A tous mes amis

Qui m'ont soutenu et encouragé tout au long de ce Projet de Fin d'Etude

J'espère qu'ils trouveront dans ce travail

Toute ma reconnaissance

Et tout mon amour.

Alaeddine

ملخص

في إطار عملنا، نحن مهتمون بأمراض الكلام خاصة الحنك المشقوق.

الهدف من المشروع النهائي للدراسة هو إنشاء نظام تقويمي من اجل إعادة تأهيل المرضى الذين يعانون من الحنك المشقوق بعد الجراحة.

لتحقيق هذا الهدف، قمنا بدراسة مدونة تم تسجيلها من قبل المتكلمين (اثنين عاديين واثنين مرضى). تحليل الاصوات من قاعدة البيانات هذه اجري باستخدام معاملات وسائط مل كلبستر(MFCC)، في حين ان نمذجة إشارة الكلام حققت باستعمال نموذج المتعدد الغوصيات (GMM).

اظهرت النتائج التي تحصلنا عليها وجود معدل الاعتراف بنسبة 90٪ بالنسبة للمتحدثين العاديين و50٪ للمرضى.

كلمات مفتاحية: التعرف الآلي للكلام، امراض الكلام، شق الحنك، وسائط مل كلبستر(MFCC)، نموذج المتعدد الغوصيات (GMM)، معدل الاعتراف.

RESUME

Dans le cadre de notre travail, nous nous sommes intéressés à la pathologie de la parole et au cas particulier de la fente palatine.

Le but ciblé de notre Projet de Fin d'Etude est d'élaborer un système orthophonique afin de rééduquer les patients ayant une fente palatine après une intervention chirurgicale.

Pour atteindre cet objectif, nous avons étudié un corpus. Ce dernier a été enregistré par 4 locuteurs (deux normaux et deux pathologiques). L'analyse acoustique de cette base de données a été effectuée en utilisant les **MFCC** (**M**el **F**requency **C**epstral **C**oefficients), quant à la modélisation du signal de parole nous avons appliqué les **GMM** (**G**auss **M**ixture **M**odel).

Les résultats obtenus indiquent un Taux de Reconnaissance (TR) de 90% pour la Parole Normale et 50% pour la parole pathologique.

Mots-clés : Reconnaissance Automatique de la Parole, Pathologie de la parole, Fente palatine, MFCC, GMM, Taux de Reconnaissance.

ABSTRACT

As part of our work, we are interested in speech pathology case: cleft palate.

The targeted goal of our Final Project Study is to develop a speech-language system to rehabilitate patients with cleft palate after surgery.

To achieve this goal, we studied two corpus, four speakers (two normal and two pathological) recorded this last. The acoustic analysis of this database was performed using the MFCC (Mel Frequency Cepstral Coefficients) to extract the parameters and for the modeling of the speech signal; we applied the **GMM** (**G**auss **M**ixture **M**odel).

The obtained resultants indicate a recognition rate of 90% for the normale speech and 50% for the pathological speech.

Keywords: Automatic Speech Recognition, Speech Pathology, Cleft Palate, MFCC, GMM, recognition rate.

LISTE DES ABREVIATIONS

API	**A**lphabet **P**honétique **I**nternational
BD	**B**ases de **D**onnées
CV	**C**ordes **V**ocales
DE	**D**istance **E**uclidienne
E	**E**nergie
EM	**E**spérance **M**aximale
DTW	**D**ynamic **T**ime **W**arping
FAR	**F**alse **Ac**ceptance **R**ate
FRR	**F**alse **R**ejection **R**ate
FFT	**F**ast **F**ourier **T**ransform
FIR	**F**inite **I**mpulse **R**esponse
GMM	**G**aussian **M**ixture **M**odels
HMM	**H**idden **M**arkov **M**odel
IAL	**I**dentification **A**utomatique du **L**ocuteur
IFFT	**I**nverse **F**ast **F**ourier **T**ransform
LPC	**L**inear **P**redictive **C**oding
LPCC	**L**inear **P**redictive **C**epstral **C**oefficients
MFCC	**M**el **F**requency **C**epstral **C**oefficients
MV	**M**aximum de **V**raisemblance
OE	**O**reille **E**xterne
OI	**O**reille **I**nterne
OM	**O**reille **M**oyenne
PNorm	**P**arole **N**ormale
PPath	**P**arole **P**athologique
QV	**Q**uantification **V**ectorielle
RAL	**R**econnaissance **A**utomatique du **L**ocuteur
RAP	**R**econnaissance **A**utomatique de la **P**arole
TAP	**T**raitement **A**utomatique de la **P**arole
TFD	**T**ransformée de **F**ourier **D**iscrète

LISTE DES FIGURES

		Page
Fig. 1.1	Dispositifs artificiels pour la génération des voyelles orales…..……	02
Fig. 1.2	Modèle simplifié de l'appareil phonatoire humain……	04
Fig. 1.3	Système auditif humain…….	06
Fig. 1.4	Echelle du bruit : de l'audible au seuil de douleur…….	07
Fig. 1.5	Production des sons voisés ……...	10
Fig. 1.6	Production des sons non voisés ……	
Fig. 1.7	Classification des voyelles…….……	11
Fig. 1.8	Organes modifiant le volume et la forme des résonateurs de la voix…	12
Fig. 1.9	Classification des Consonnes …….……..	13
Fig. 1.10	Sonagrammes des occlusives : sourde [t] / sonore [d] en contexte vocalique [a]……..	13
Fig. 1.11	Sonagrammes des fricatives : non voisée [s] / voisée [h] en contexte vocalique [a]……..	14
Fig. 2.1	Algorithme Diagnostic de la Dysphonie ……...	15
Fig. 2.2	Vue postérieure des principaux muscles du palais ……...	19
Fig. 2.3	Malformations faciales (Classification Y)……...……..	21
Fig. 2.4	Fente vélaire partielle (luette bifide) / totale……..	
Fig. 2.5	Fente vélo-palatine……...……..	22
Fig. 2.6	Fente complète unilatérale……...……..	
Fig. 3.1	Organigramme de la procédure d'analyse d'un signal vocal ……	28
Fig. 3.2	Banc de filtres sur l'échelle linéaire……..	31
Fig. 3.3	Banc de filtres sur l'échelle Mel……	31
Fig. 3.4	Schéma de calcul des MFCC ……..……	32
Fig. 3.6	HMM gauche-droite à trois états……...	34
Fig. 3.7	Modèle de GMM……...	35
Fig. 3.8	Exemples d'une distribution Gaussienne à 2 dimensions……..	37
Fig. 3.9	Exemple d'un modèle Gaussien à 2 dimensions…….	37

Fig. 4.1	Studio d'enregistrement…….	43
Fig. 4.2	Organigramme de classification automatique de parole normale / pathologique……..	44
Fig. 4.3	Signal temporel du mot [babobi] après préaccentuation……...……	45
Fig. 4.4	Représentation spectrale de la MFCC du corpus traité…….....	46
Fig. 4.5	Evaluation des taux FAR et FRR……...	47

LISTE DES TABLEAUX

		Page
Tableau 1.1	Nomenclature des lieux d'articulation ……..…….	11
Tableau 1.2	Phonétique et enseignement du Français ……	14
Tableau 2.1	Caractéristiques des occlusives …….…….	24
Tableau 4.1	Corpus enregistrés en phrases et mots isolés …….	41
Tableau 4.2	Locuteurs à tester (2 Normaux et 2 Pathologiques)……	42
Tableau 4.3	Paramètres utilisés pour l'extraction des MFCC avec HTK……...	46

TABLE DES MATIERES

Introduction Générale 1

Chapitre 1
Notions Fondamentales sur la Parole

1.1. Introduction 2
1.2. Historique de la production de la parole 2
1.3. Appareil phonatoire humain 4
1.3.1. Voies aériennes inférieures 5
1.3.2. Larynx et le cordes vocales 5
1.4. Système auditif et perception de la parole 6
1.4.1. Système de transmission 6
1.4.2. Aire de l'audition 7
1.5. Concepts fondamentaux de physico-acoustique et de phonétique 8
1.5.1. Fréquence fondamentale F_0 8
1.5.2. Durée 8
1.5.3. Timbre 8
1.5.4. Intensité 8
1.5.5. Prosodie 9
1.6. Fonctionnement acoustique de l'appareil vocal 9
1.6.1. Voisement 9
1.6.2. Mode d'articulation 10
1.6.3. Lieu d'articulation 10
1.6.4. Voyelles 10
1.6.5. Consonnes 11
1.6.5.1. Articulation occlusive 13
1.6.5.2. Articulation fricative 14
1.7. Conclusion 15

Chapitre 2
Pathologie de la Parole

2.1. Introduction 16
2.2. Pathologies de la parole 16
2.2.1. Dysphonie 16
2.2.2. Dysarthrie 18
2.2.3. Dyslalie 19
2.2.3.1. Bégaiement 19
2.2.3.2. Sigmatisme 20
2.3. Fentes palatines 20
2.3.1. Anatomie du palais 20
2.3.2. Fonction du palais 22
2.3.2. Classification des fentes 22
2.2.3.1. Division simple du voile 23
2.2.3.2. Division du voile et de la voûte palatine 24
2.2.3.3. Division du voile associée à une fente labio-alvéolaire unilatérale .. 24
2.4. Etude de la fente palatine simple 24
2.5. Conclusion 26

Chapitre 3
Analyse de la Parole Pathologique

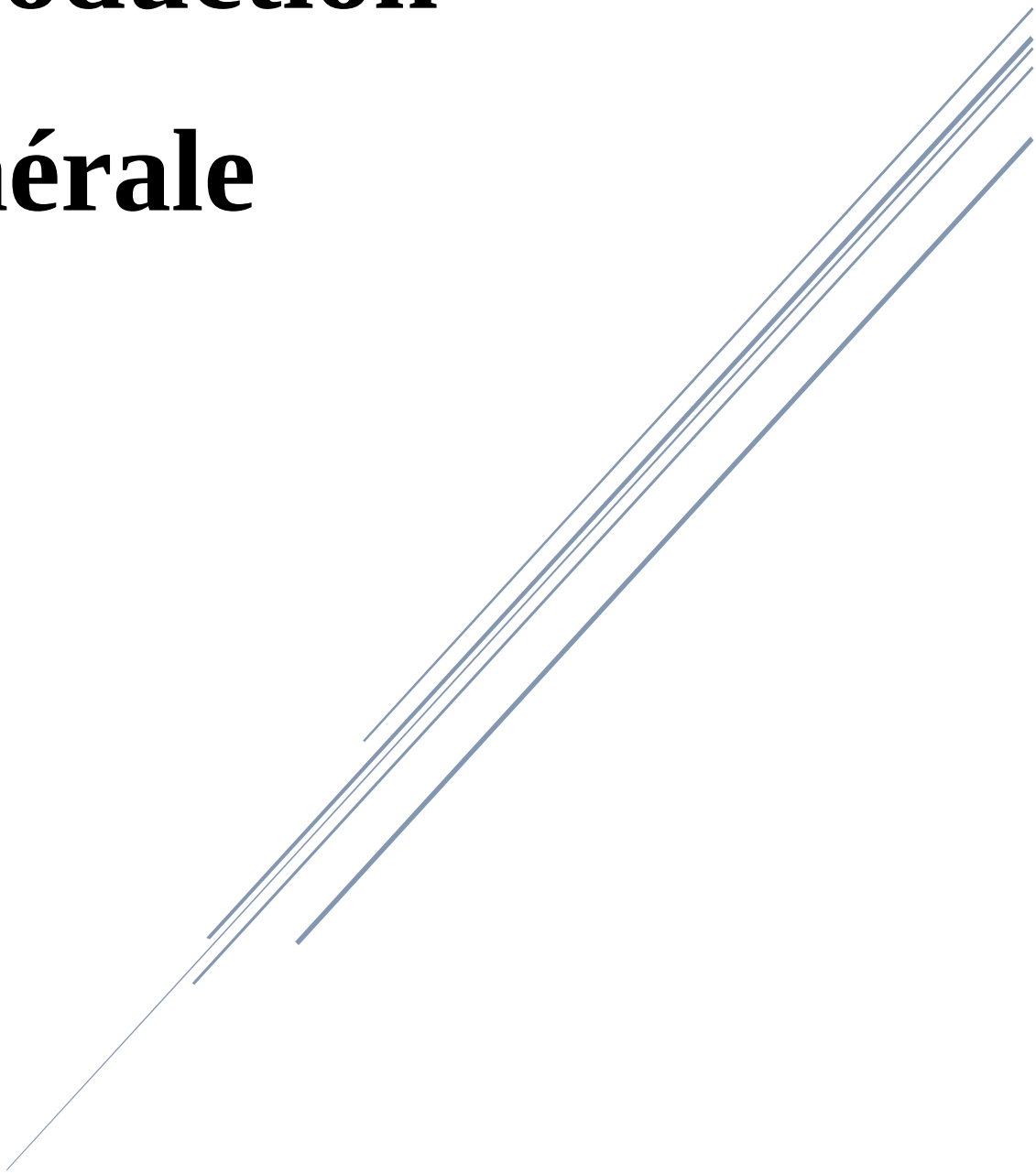
3.1. Introduction 27
3.2. Analyse de la parole pathologique 27
3.2.1. Paramètres acoustiques de la Parole Pathologique 27
3.2.2. Prétraitement acoustique 28
3.2.2.1. Préaccentuation 28
3.2.2.2. Fenêtrage 29
3.2.3. Méthode d'analyse en traitement de parole 29
3.2.3.1. Coefficients de prédiction linéaire LPC 29
3.2.3.2. Paramètres LSP 29
3.2.3.3. Coefficients cepstraux de prédiction linéaire LPCC 30
3.2.3.4. Coefficients MFCC 30
3.3. Modélisation des locuteurs 33
3.3.1. Approche vectorielle 33
3.3.1.1. Reconnaissance du locuteur à base de DTW 33
3.3.1.2. Quantification vectorielle (QV) 33
3.3.2. Approche statistique 34
3.3.2.1. Modèles de Markov cachés HMM 34
3.3.2.2. Les mélanges de gaussiennes GMM 34
3.4. Reconnaissance par le mélange de gaussiennes 35
3.4.1. Estimation des paramètres 38
3.3.1.1. Algorithme des K-moyennes 38
3.3.1.2. Algorithme EM (Expectation Maximisation) 39
3.4.2. Phase d'apprentissage 40
3.4.3. Rapport d'hypothèses Bayésien 40
3.5. Conclusion 40

Chapitre 4
Expériences et Résultats

4.1. Introduction 41
4.2. Description de la base de données 41
4.3. Description de l'application développée 42
4.4. Description de l'interface 42
4.5. Conception du système de classification élaboré 43
4.5.1. Enregistrements du corpus 44
4.5.2. Traitement du signal de parole 45
4.5.3. Extraction des MFCC 46
4.6. Evaluation du système développé 47
4.6.1. Evaluation des performances 47
4.6.2. Choix du seuil de décision 48
4.7. Conclusion 48

Conclusion et Perspectives 49
References bibliographiques 51

Introduction Générale



Les travaux présentés dans notre projet se situent dans le cadre de la RAP (Reconnaissance Automatique de la Parole). Le système développé a pour objectif la rééducation des patients ayant une fente palatine auparavant. Pour ce faire, deux types de traitements sont nécessaires : un traitement acoustique accompli par un modèle de langage.

Notre travail est inspiré de certaines lacunes relevées dans la prise en charge des patients et que nous avons jugé utile à étudier et corriger :

- une absence de formation des orthophonistes dans la manipulation de logiciels d'analyse acoustique ;
- un manque flagrant de coopération entre l'orthophoniste en milieu hospitalier algérien, l'ingénieur, chercheur phonéticien, et acousticien ;
- une utilisation exclusive de l'oreille humaine (ouïe) pour évaluer l'effet de la réhabilitation vocale dans les hôpitaux algériens. L'évaluation de la voix pathologique est principalement basée sur la perception subjective des cliniciens sans aucune analyse acoustique de la PPath.

Dans le cadre de notre travail, nous nous sommes intéressés à la classification automatique de la Parole Pathologique (PPath) par rapport à la Parole Normale (PNorm). Nous avons effectué, dans une première étape, une analyse acoustique portant sur la PPath. Dans une seconde étape, nous avons utilisé les Mélanges de Gaussiennes GMM. Cette approche a fait preuve de fiabilité et d'efficacité. Dans cette dernière, le problème de reconnaissance est divisé en phases d'apprentissage et de test.

Notre projet est divisé en quatre chapitres :

- le premier présente une étude générale sur le rôle des structures périphériques du système de production de la parole, ainsi que les mécanismes de production et de perception de la parole et sa description acoustique ;
- le deuxième est consacré aux généralités et classification de la Pathologie de la Parole, en particulier le cas de la fente palatine ;
- dans le troisième nous exposons les outils de traitement de la parole pathologique, les paramètres les plus efficaces pour le représenter et les différentes approches;
- le dernier chapitre est le noyau de ce projet, il présente la description de notre application : le principe de fonctionnement des différentes fonctions constituant notre Système Orthophonique de la Fente Palatine (SOFP) ainsi que les étapes suivies pendant sa réalisation et son évaluation.

CHAPITRE 1 :

Notions Fondamentales sur la Parole

1.1. Introduction

Dans ce chapitre, nous commencons par un bref historique de la production de la parole, puis nous exposons le rôle prépondérant que peuvent jouer les structures périphériques du système de production de la parole, en liaison avec leurs efficacités perceptives. Il s'agit en réalité d'examiner l'importance des gestes articulatoires pour le système de la communication humaine, et les conséquences de déviances articulatoires éventuelles liées, par exemple, à la parole pathologique.

1.2. Historique de production de la parole

Depuis les temps les plus reculés, les Hommes ont toujours eu l'ambition de faire produire à des dispositifs artificiels des actions d'hommes ou d'animaux.

Nombreuses sont les légendes qui témoignent de la persistance de ce désir. Des figurines animées (à la main) ont été fabriquées dès l'antiquité. Mais c'est vers la fin du XVIII^{ème} siècle qu'a vu naître les premières machines mécaniques capables de simuler les sons vocaux.

En 1779, l'Académie impériale de Saint-Pétersbourg organise un concours scientifique avec deux questions : qu'est ce qui différencie autant les voyelles des autres sons ? Est-il possible de faire prononcer par une machine, les sons de ces voyelles ? Le lauréat est un professeur de l'université de Copenhague, Christian Gottlieb Kratzenstein qui réalise une série de résonateurs acoustiques de dimensions et formes similaires à celles de la bouche humaine, et excités par une anche vibrante simulant le fonctionnement des cordes vocales (figure1.1).

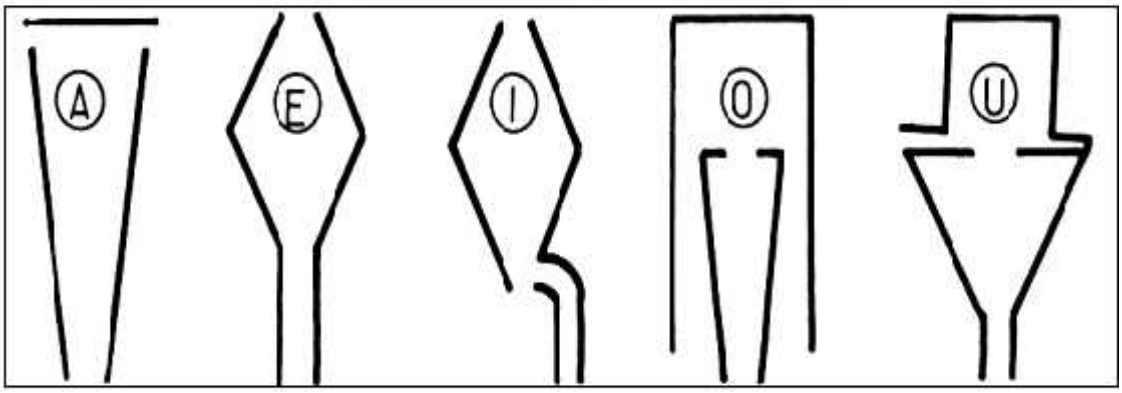


Fig. 1.1 : Dispositifs artificiels pour la génération des voyelles orales [1]

En 1939, Homer Dudley, présente le Voder "**V**oice **O**peration **D**emonstrator" à l'Exposition internationale de New York. Cet appareil est excité soit par un bruit blanc, soit avec un signal très riche en harmoniques, ces sons étant modulés par une boîte de contrôle de résonance le "conduit vocal" contenant un banc de dix filtres passe-bandes répartis entre 300 Hz et 3000 Hz.

Les recherches sur la synthèse de la parole sont motivées par le souci de transmettre la voix avec une plus grande efficacité, c'est-à-dire, en réduisant la largeur de bande nécessaire aux conversations téléphoniques. C'est la raison pour laquelle, très tôt, des recherches sont menées aux laboratoires de Bell Telephone. Ces recherches ont conduit à l'élaboration du Vocoder ou "**Vo**ice Coder" qui permet d'utiliser uniquement une bande passante de 275 Hz au lieu des 3100 Hz requis pour le téléphone.

Les années soixante et soixante dix ont vu apparaître d'autres techniques permettant de synthétiser la parole notamment la synthèse par éléments phonétiques. Cela consiste à reconstituer artificiellement des mots à partir de segments de mots par exemple avec "auto ", "mati","que" on peut reconstituer le mot "automatique". Le découpage peut se faire de façon plus fine encore jusqu'à la plus petite unité phonétique : le phonème. L'assemblage des phonèmes, selon un ensemble de lois particulières à chaque langage permet de reconstituer les mots parlés.

Il existe de nombreuses autres méthodes permettant de générer de façon synthétique un signal de parole telle que la synthèse par formants qui donne de meilleurs résultats mais demande une analyse fine du signal de parole.

Le développement des techniques numériques de traitement du signal a permis l'intégration de cette technique dans un circuit intégré et cela dès 1978 par Texas Instruments.

C'est à ce moment aussi que le **T**raitement **A**utomatique de la **P**arole (**TAP**) proprement dit, c'est-à-dire le traitement de l'information contenue dans le signal vocal a pris un essor considérable. Malgré cela la recherche dans le domaine du TAP est toujours très active dans divers domaines :

- amelioration des codages dans le but de réduire le débit binaire du signal;
- reconnaissance de la parole (Dialogue Homme Machine) ;
- synthèse de la parole à partir d'un texte écrit ;
- identification d'un locuteur ;
- apprentissage de langues étrangères, etc…

En conclusion, ces quelques applications montrent que le traitement de la parole prend une part de plus en plus importante dans notre vie quotidienne. Dans un futur proche on peut parier que tout ou presque se fera à l'aide de la parole et cela est d'autant plus vrai que les microprocesseurs chargés de faire les traitements sont plus rapides et plus petits [1].

1.3. Appareil phonatoire humain

L'appareil phonatoire est un ensemble d'organes d'une grande complexité mécanique. Il se compose de deux parties anatomiquement distinctes. Les poumons et le larynx, partie supérieure de la trachée artère, constituent l'essentiel du générateur sonore. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée artère (figure1.2).

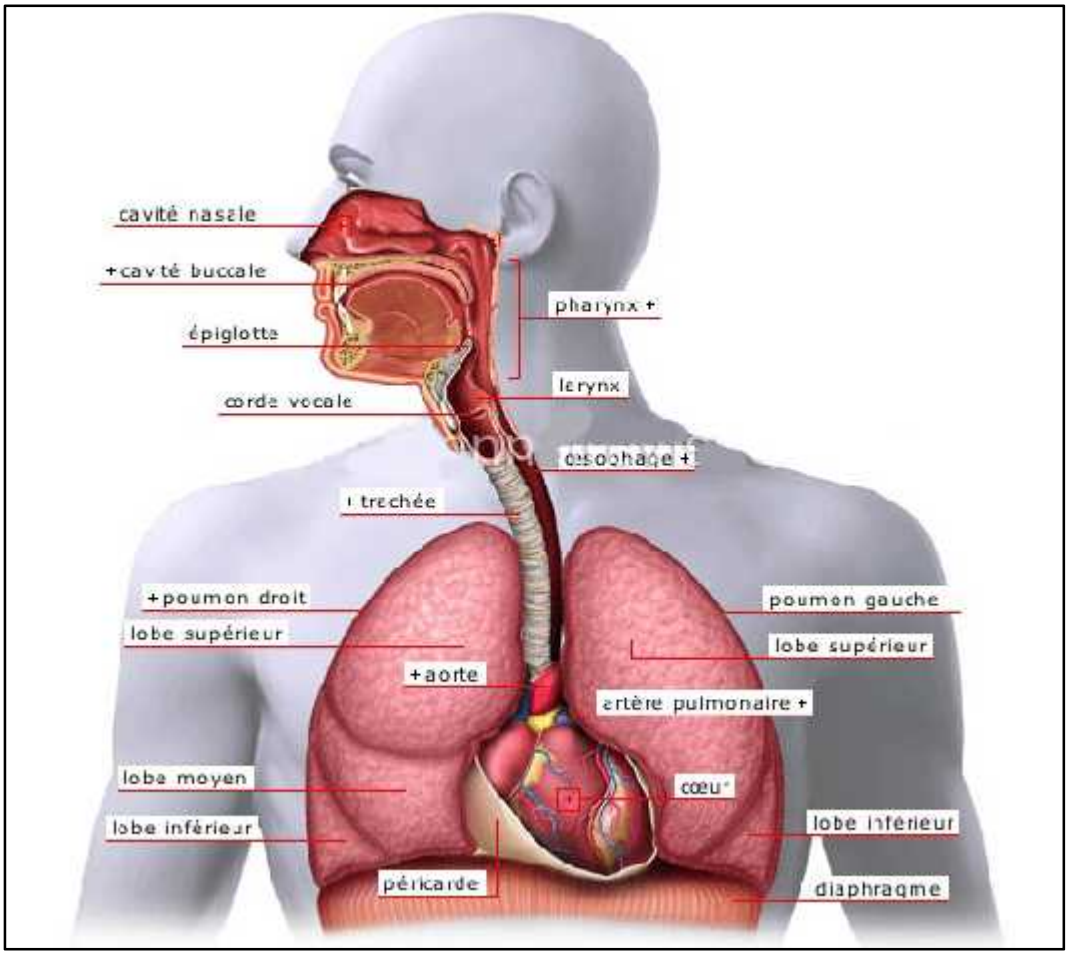


Fig. 1.2 : Modèle simplifié de l'appareil phonatoire [2].

1.3.1. Voies aériennes inférieures

Les voies aériennes inférieures correspondent à la partie de l'appareil phonatoire située

dans le thorax et sont composées de deux poumons reliés à la trachée qui elle-même remonte jusqu'aux voies aériennes supérieures. La fonction première des poumons est d'assurer la fonction de respiration en permettant l'échange d'oxygène et de dioxyde de carbone entre le sang et l'air extérieur. Lors de la phonation, les poumons jouent le rôle de réservoir de pression et permettent de générer l'écoulement d'air à l'origine de la production de sons et notamment des vibrations des **C**ordes **V**ocales (**CV**).

La circulation de l'air entre les poumons et l'extérieur est réalisée grâce aux mouvements du diaphragme et aux contractions et relâchements des muscles de la cage thoracique. Cette ventilation se fait ainsi dans un mouvement de va-et-vient correspondant alternativement à l'inspiration et à l'expiration. Sauf pour des cas atypiques, la phonation intervient durant la phase d'expiration.

1.3.2. Larynx et le conduit vocal

Le larynx a une fonction qui lui est propre: c'est la production des sons, ou "phonation". Les CV sont en fait deux lèvres symétriques placées en travers du larynx. Ces dernières peuvent fermer complètement le larynx et, en s'écartant, déterminer une ouverture triangulaire appelée glotte. Pendant la respiration, l'air y passe librement et aussi pendant la phonation des sons sourds ou non voisés. Les sons voisés résultent au contraire d'une vibration périodique des CV; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal qui s'étend du pharynx jusqu'aux lèvres.

La voix résulte du fonctionnement simultané des poumons, du larynx, de la cavité buccale et nasale qui modifie sa forme et ses dimensions suivant le son émis et qui avec la poitrine, jouent le rôle de caisse de résonance. Toutes les voix se ressembleraient si la voix était seulement laryngée. Or, ce sont les modifications de forme et de dimensions que subissent la bouche, le pharynx pendant l'émission de la voix, qui donnent au contraire à celle-ci un timbre qui est particulier à chacun d'entre nous.

On peut remarquer que, d'après ce que nous avons vu précédemment, nous avons deux générateurs de sons, source excitatrice les CV, quand elles vibrent le son est voisé ou sonore et quand elles ne vibrent pas le son est sourd ou non voisé, et d'un filtre (le conduit vocal) capable d'amplifier ou d'amortir certains sons [3].

1.4. Système auditif et perception de la parole

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production.

Les processus complexes par lesquels un auditeur comprend un message oral émis par un locuteur peuvent être fonctionnellement décomposés en deux grandes phases :

- l'oreille transforme l'information contenue dans le signal acoustique et le transmet ensuite au cerveau par l'intermédiaire du nerf auditif ;
- la reconnaissance du message linguistique par l'interprétation d'indices fournis à l'issu de prétraitement auditif sans référence à la signification puis la réalisation de l'accès au sens.

1.4.1. Système de transmission

L'appareil auditif comprend l'Oreille Externe (OE), l'Oreille Moyenne (OM) et l'Oreille Interne (OI) (Figure 1.3).

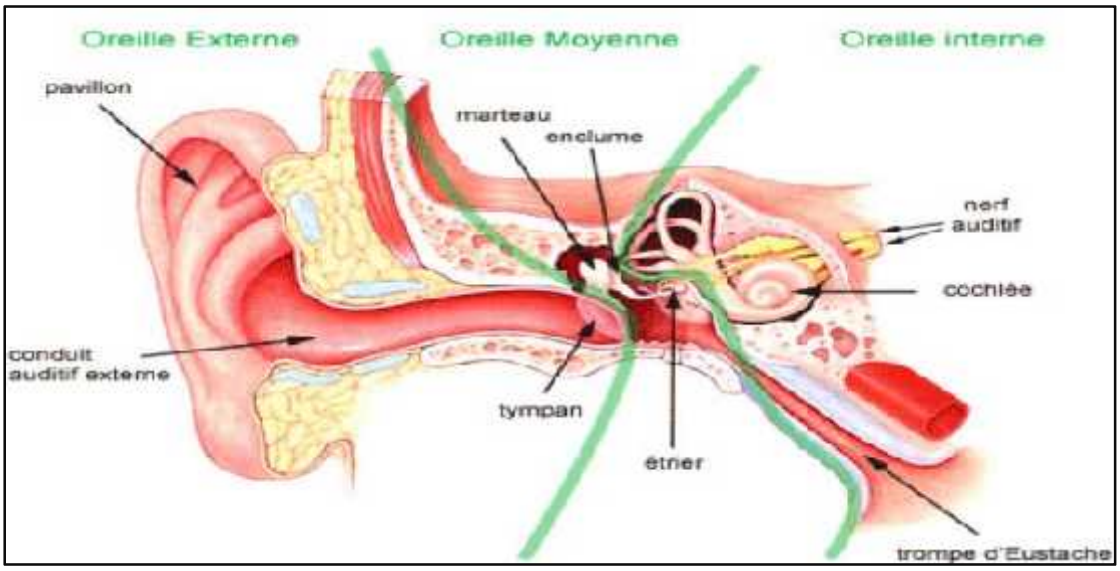


Fig. 1.3 : Le système auditif humain [4]

- l'OE relie le pavillon et conduit auditif permet de recueillir les sons et de les orienter vers l'oreille moyenne ;
- l'OM comprend le tympan et les osselets. Elle assure la fonction de transmission, qui inclut une transformation d'ondes sonores aériennes en ondes liquidiennes. Elle joue aussi un rôle d'accommodation auditive. La position des osselets les uns par rapports aux autres assure l'amplification des sons ;

- le mécanisme composé de marteau, étrier et enclume permet une adaptation d'impédance entre l'air et le milieu liquide de l'OI. Les vibrations de l'étrier sont transmises au liquide de la cochlée qui avec le nerf auditif assurent la fonction de réception [4].

1.4.2. Aire de l'audition

Le système auditif ne répond pas également à toutes les fréquences. Le champ auditif humain est délimité par la courbe de seuil de l'audition et celle du seuil de la douleur (figure 1.4). Sa limite supérieure en puissance (≈80 – 120 dB), au de la sa risque d'endommager son systemes d'une façon irrecursive .

Fig. 1.4 : Echelle du bruit : de l'audible au seuil de douleur

1.5. Concepts fondamentaux de la phonétique

La parole est le seul moyen qui permet de communiquer la pensée par un système de sons articulés. Les humains sont les seuls êtres vivants qui utilisent un tel type des systèmes structurés. Alors avant qu'on entame le traitement automatique de la parole **TAP**, il nous faut d'abord connaître quelque notions fondamentales sur production et la caractèrisation de la parole.

1.5.1. Fréquence fondamentale F_0

La vibration, qui est en fait l'accolement puis la séparation, des cordes vocales portées par le larynx détermine la fréquence fondamentale appelée encore pitch ou F_0. Elle est comprise entre 75 et 150 Hz chez les hommes, 150 et 300 Hz chez les femmes, et est supérieure ou égale à 300 Hz chez les enfants .

1.5.2. Durée

Elle est mesurée en unités de temps, généralement la ms en ce qui concerne l'analyse de la parole. L'étude de ce paramètre peut être corrélée avec le paramètre intensifié ou fréquence dans l'analyse de l'accentuation ou d'autres phénomènes prosodiques tels que le débit.

1.5.3. Timbre

Le timbre correspond aux caractéristiques auditives de la coloration du son d'un individu. Ce timbre est en grande partie dépendant de caractéristiques physiologiques, notamment celles du larynx et des structures supra-laryngales d'un individu. Par exemple, le timbre sera plus ou moins aigu selon la longueur des cordes vocales, leur degré de tension, etc. Le timbre varie donc considérablement d'une personne à l'autre. Naturellement, il y a aussi un fondement acoustique à ce timbre.

Il est caractérisé d'une part par le type d'harmoniques présents dans le son et d'autre part par les amplitudes de ces harmoniques.

1.5.4. Intensité

C'est la qualité du son d'être plus ou moins fort. Elle se mesure en décibels (dB). Ses valeurs les plus élevées correspondent à des accents (accents d'intensité). Elle dépend de l'amplitude des vibrations. La perception se situe entre 0 et 140 dB

1.5.5. Prosodie

La prosodie introduit dans la prononciation d'une phrase des nuances qui, dans la langue écrite, demanderaient des ponctuations ou des énoncés différents. Ce sont les caractéristiques prosodiques qui permettent à un auditeur de suivre une conversation même en milieu défavorable. Les principaux paramètres prosodiques sont l'intonation, l'intensité et la durée [5].

1.6. Fonctionnement acoustique de l'appareil vocal

La parole résulte de l'excitation du conduit vocal par deux types de sources :

- La vibration des cordes vocales qui produit les sons voisés ou sonores. Dans ce cas le son émis est périodique. C'est le cas pour toutes les voyelles ;
- Une source de bruit qui se crée en un point de resserrement du conduit vocal. Le son émis sera apériodique. Les sons non voisés ou sourds se rencontrent uniquement parmi les consonnes.

La combinaison des deux types de sources donne les consonnes voisées.

Tous les phonèmes sont transcrits dans le code de l'**A**lphabet **P**honétique **I**nternational (**API**) [6].

Les sons élémentaires de la parole peuvent être classés en fonction de trois variables essentielles :

- le voisement : activité des cordes vocales ;
- le mode d'articulation : type de mécanisme de production ;
- le lieu d'articulation : endroit de resserrement maximal du conduit vocal.

1.6.1. Le voisement

Tous les sons du langage peuvent être classés, suivant le mode de leur production, en deux catégories. Ils proviennent soit :

- d'une vibration laryngée : "son voisé" ;
- d'un bruit de passage de l'air pulmonaire à travers les organes phonatoires : "son non voisé".

Dans la première catégorie, les mouvements de vibration des cordes vocales sont adductés (fermeture glottique). Le débit glottal émis est alors considéré comme étant la forme de l'onde du signal sonore périodique fourni par la source d'excitation voisée (Fig.1.5).

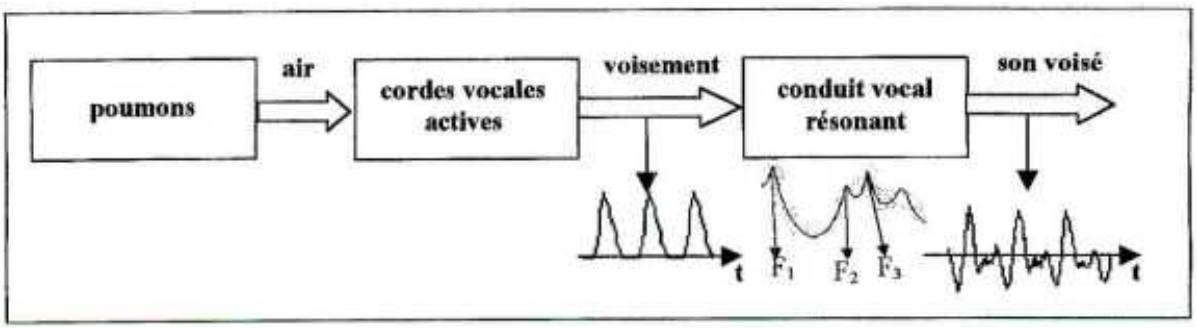


Fig.1.5 : Production des sons voisés [7]

En revanche, dans la deuxième catégorie, les cordes vocales sont en mouvements d'abduction (ouverture glottique) et sont donc tenues à part de manière à ne pas affecter le passage de l'air pulmonaire par les vibrations glottales. Le signal d'excitation ainsi émis est donc de type apériodique (Figure 1.6).

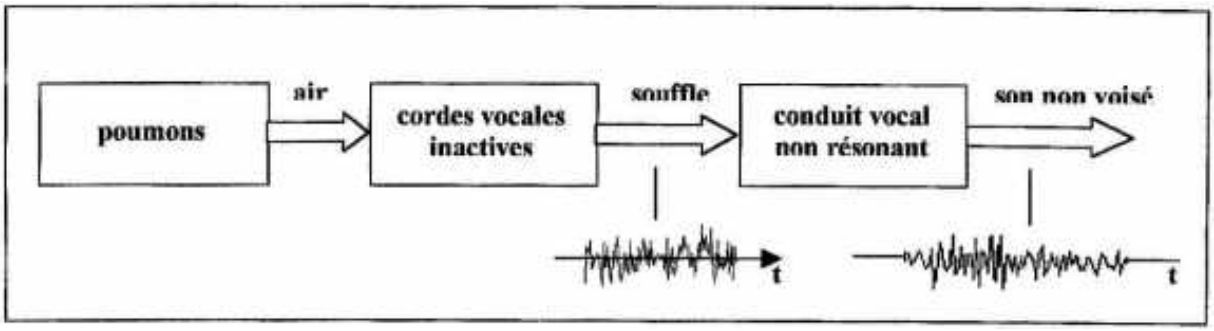


Fig.1.6 : Production des sons non voisés [7]

1.5.2. Le mode d'articulation

Le mode d'articulation est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré. Parmi ces facteurs, nous pouvons citer les plus importants

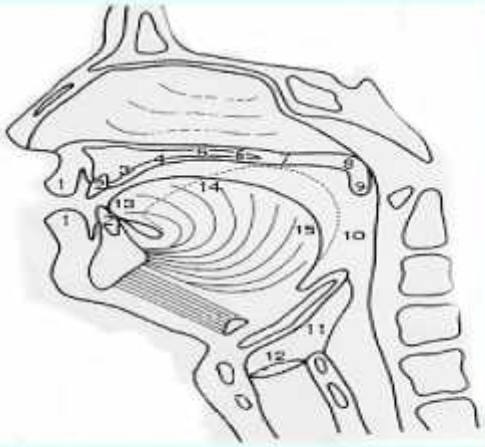
- passage libre ou mise en vibration de l'air au niveau de la glotte (son sourd ou sonore) ;
- passage par une voie unique ou deux voies différentes (son oral ou nasal) ;
- obstruction totale ou partielle du passage de l'air dans un lieu du conduit vocal (son occlusif ou fricatif).

1.5.3. Lieu d'articulation

Le point d'articulation est l'endroit où se trouve, dans la cavité buccale, un obstacle au passage de l'air. De manière générale, on peut dire que le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage du canal d'air [8].

Le lieu d'articulation peut se situer aux endroits cité dans le tableau (Tab 1.1)

Tab 1.1 : Nomenclature des lieux d'articulation.



	Organe anatomique		Nomenclature phonétique correspondante	
1	lèvres		labiales	
2	dents		dentales	
3	alvéoles		alvéolaires	
4	palais dur		pré-palatales	
5			médio-palatales	
6			post-palatales	
7	voile du palais		pré-vélaires	
8			post-vélaires	
9	luette (*uvula*)		uvulaires	
10	pharynx		pharyngales	
11	larynx		laryngales	
12	glotte		glottales	
13	apex	de la langue	apicales (pré-dorsales)	dorsales
14	dos		médio-dorsales	
15	racine		radicales (post-dorsales)	

1.5.4. Voyelles

Dans la description articulatoire des voyelles, on peut distinguer deux dimensions. D'un côté, le mode d'articulation décrit la configuration générale des organes articulatoires dans la production d'une voyelle donnée. D'un autre côté, le lieu d'articulation décrit le point de rétrécissement maximal (c'est-à-dire fermeture) dans la production d'une voyelle (Figure 1.7).

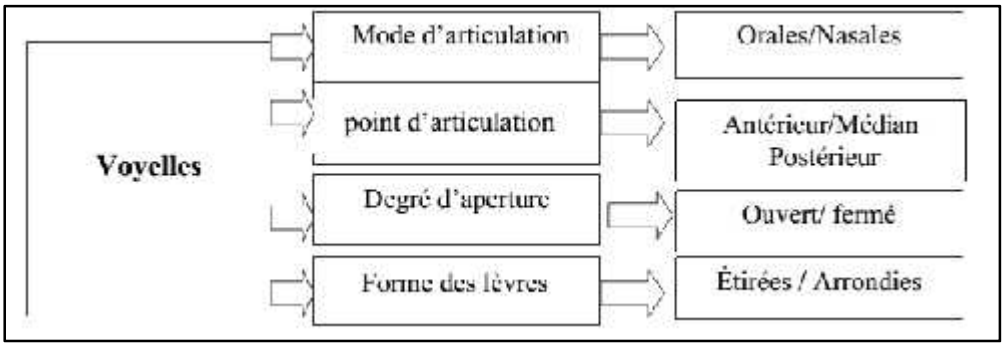


Figure 1.7 : Classification des voyelles.

- en français, le mode d'articulation permet de distinguer quatre grandes classes de voyelles, classes qui s'entrecoupent entre elles ;
- l'un des modes d'articulation dépend de la présence ou absence de nasalité. Les voyelles orales se prononcent avec le voile du palais relevé, ce qui ferme le

passage nasal. Par contre, les voyelles nasales se prononcent avec le voile du palais abaissé, ce qui laisse passer de l'air et par la bouche, et par le nez.

On a déjà dit que, du point de vue articulatoire, les voyelles comportaient toujours, à condition qu'elles ne soient pas chuchotées, des vibrations des cordes vocales ; et un passage libre de l'air ; aucun organe ne fait obstacle au passage de l'air vibrant provenant de la glotte. Aussi au sujet des corrélations entre facteurs articulatoires et facteurs acoustiques : le timbre des voyelles est dû essentiellement à deux formants.

On peut dire en gros que ces deux formants correspondent aux deux résonateurs principaux de l'appareil phonateur, buccale (pour le F_2) et pharynx (pour le F_1), et que ce sont principalement les mouvements de la langue et de la mâchoire inférieure qui permettent de varier l'effet résonateur de ces 2 cavités (Figure 1.8) [9].

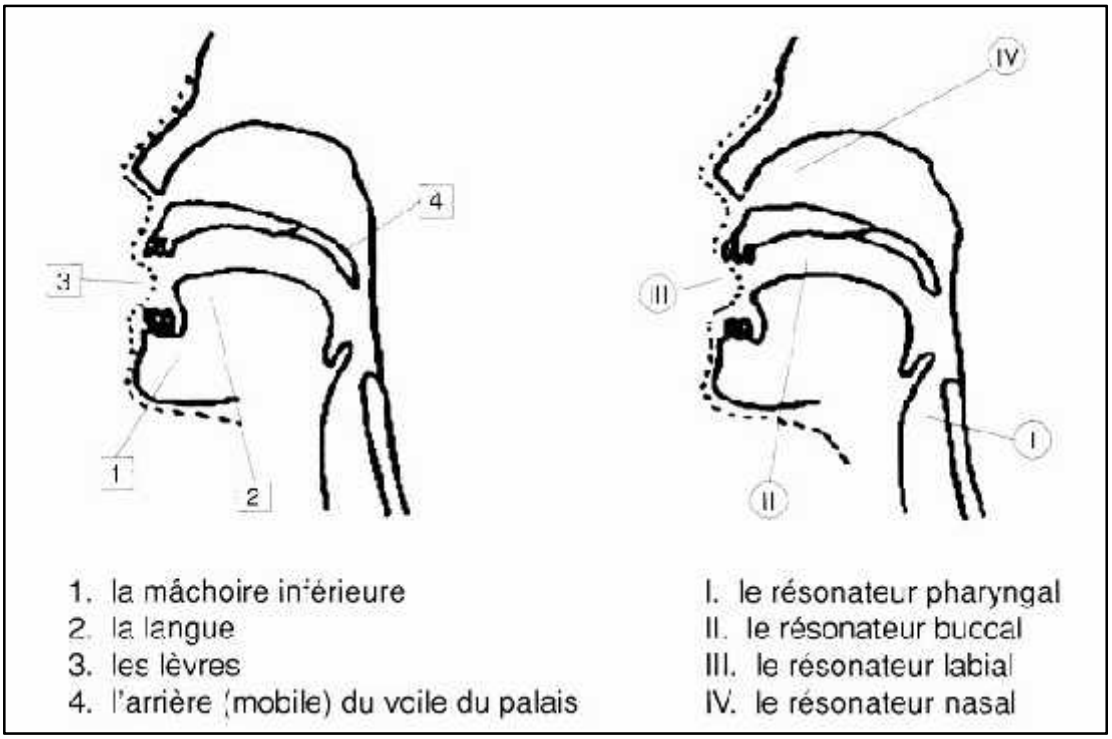


Fig. 1.8 : Organes modifiant le volume et la forme des résonateurs de la voix [9]

1.6.5. Consonnes

Il existe plusieurs classifications pour les consonnes (Figure 1.9), tout dépend de type choisi, suivant le voisement, elles se répartissent en 2 séries : les sonores et les sourdes. Lors de leur production, l'air expiré rencontre un obstacle en un ou plusieurs points de la cavité buccale ce qui provoque un bruit (friction ou impulsion).

Comme on a déjà dit, il est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré :

- intervention des cordes vocales ou mise en vibration : articulation sonore ;
- fermeture momentanée du passage de l'air suivie d'une ouverture brusque (explosion) [10].

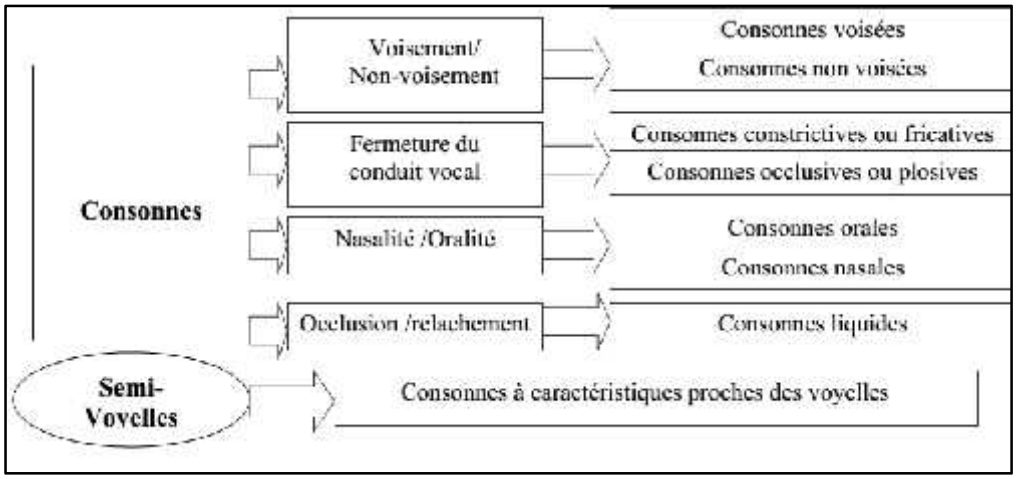


Fig. 1.9 : Classification des Consonnes.

1.6.5.1 Articulation occlusive

Les consonnes occlusives sont produites par obstruction totale du conduit vocal (occlusion) de brève durée empêchant momentanément l'air de sortir (implosion), suivie d'une ouverture articulatoire expirant brutalement l'air emmagasiné dans le conduit vocal (explosion). Ils apparaissent sur le sonagraphe, sous forme d'un silence plus ou moins court correspondant à la phase de la tenue articulatoire de l'occlusion. Lorsque nous n'observons aucune amplitude d'énergie à basses fréquences dans cette zone de silence, la consonne est dite non voisée. Quand cette zone contient de l'énergie à basses fréquences, étalée le long d'une barre horizontale nommée "barre de voisement", la consonne est voisée. Cette durée de la tenue est suivie d'une barre verticale correspondant à la "barre d'explosion" (due au relâchement de l'occlusion) ou "Burst" (Figure 1.10).

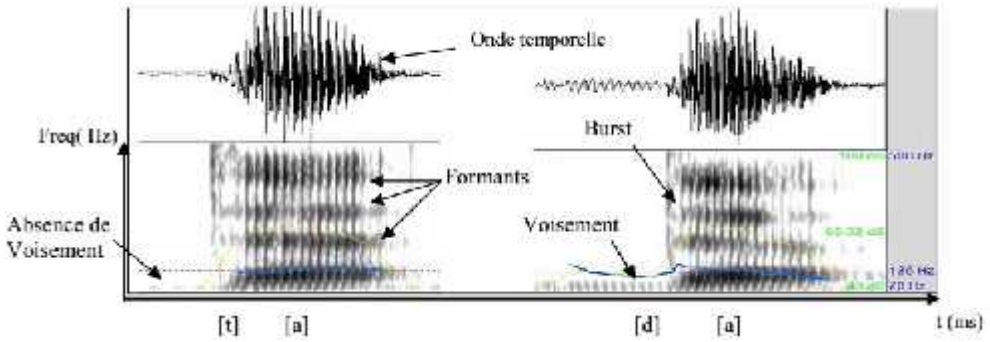


Fig. 1.10 : Sonagrammes des occlusives sourde [t] / sonore [d] en contexte vocalique [a]

1.6.5.2. Articulation fricative

Les consonnes fricatives sont produites par un rétrécissement au lieu d'articulation du conduit vocal, lors du passage de l'air pulmonaire. Sur le sonagramme, elles apparaissent sous forme d'un bruit aléatoire. Les consonnes fricatives peuvent être voisées ou non voisées. Le voisement est caractérisé par une présence d'une barre horizontale d'énergie à basses fréquences. Il est représenté également par une courbe dite de voisement (Figure 1.11). L'absence de cette bande d'énergie correspond au trait sourd (non voisé) correspondant à la non de vibration des cordes vocales.

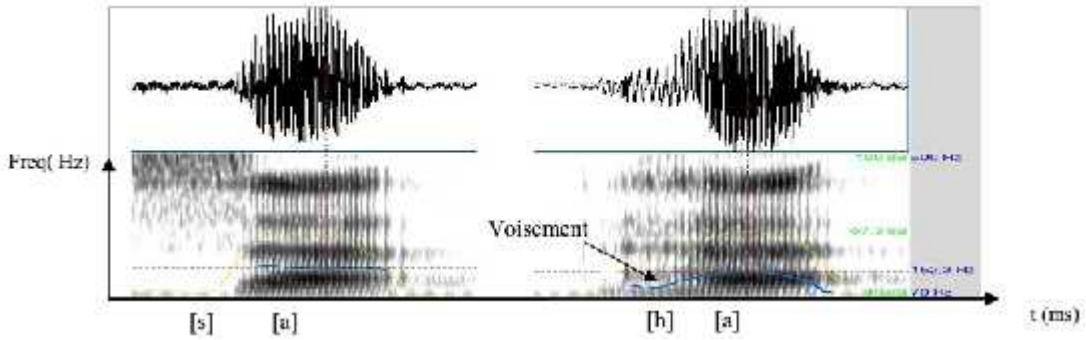


Fig. 1.11 : Sonagrammes des fricatives non voisée [s] / voisée [h] en contexte vocalique [a]

Tab 1.2 : API du Français [11]

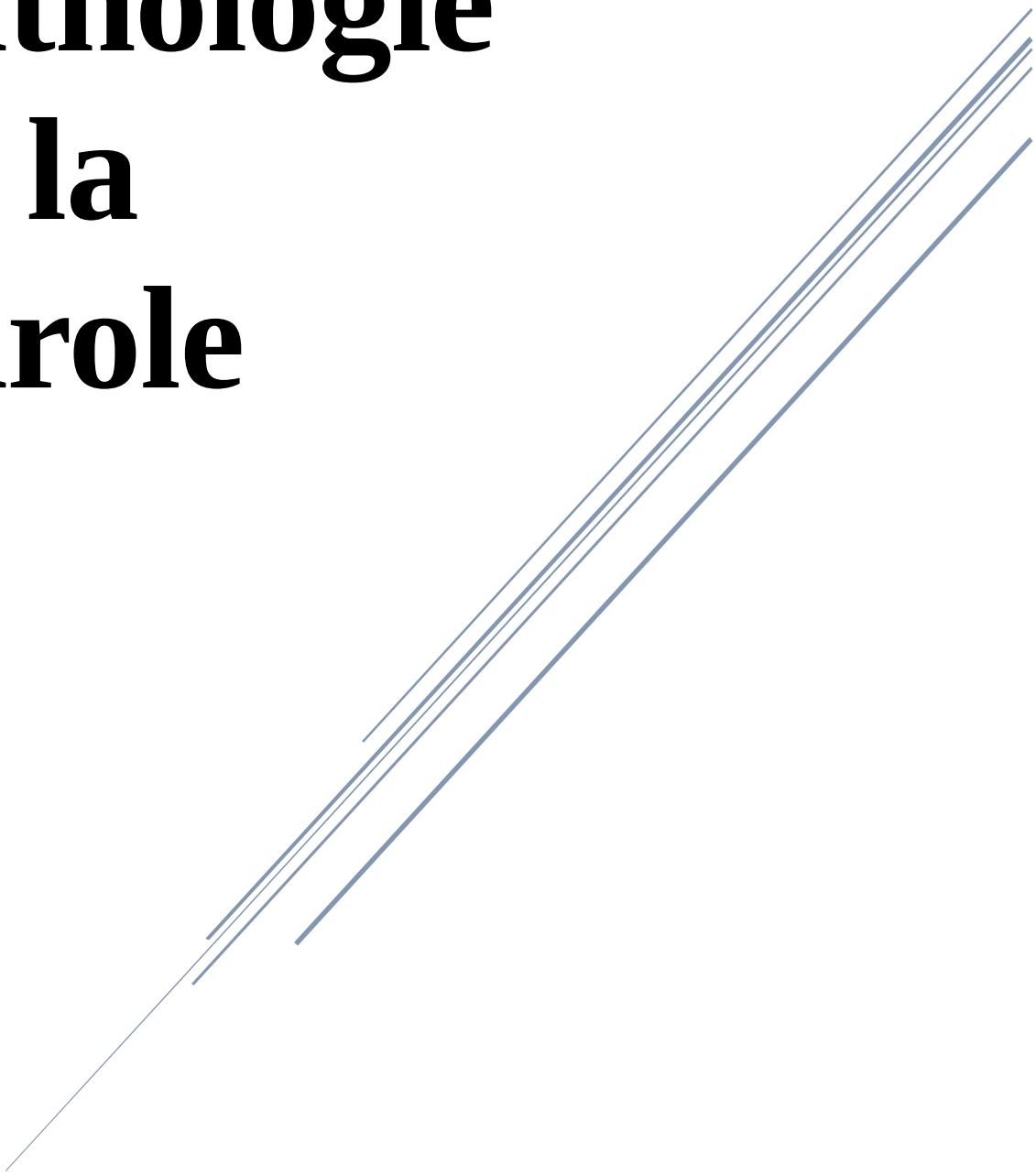
MODE D'ARTICULATION			Bilabiale	Labio-dentale	Dentale	Alvéolaire	Prépalatale	Palatale	Vélaire	Uvulaire		
MODE D'ARTICULATION	Occlusive	Médiane	p		t				k		Sourd	Orale
			b		d				g		sonore	
			m		n			ɲ				Nasale
	Constrictive			f		s	ʃ				sourd	Orale
				v		z	ʒ	j		ʁ	sonore	
		Latérale			l							
		Médiane	ɥ, w					ɥ	w			

1.7. Conclusion

Ce chapitre a permis dans sa première partie d'introduire certains concepts de base du TAP via une caractérisation du signal de parole sur le plan physiologique, acoustique et phonétique. Dans la seconde, nous avons mis en valeur les paramètres acoustiques les plus exploités et qui permettent de caractériser le signal de parole. Ce qui est un préalable indispensable à l'approche, le diagnostic, et la prise en charge des anomalies de la voix et la parole.

CHAPITRE 2 :

Pathologie de la Parole



2.1. Introduction

Dans ce chapitre, nous allons donner une brève description des diverses pathologies de la parole les plus fréquentes. Nous avons mis l'accent sur les pathologies dyslalie particulièrement la fente palatine, à savoir les paroles sur lesquelles nous avons fait une analyse acoustique et une classification automatique par rapport à la P_{Norm} (**P**arole **N**ormale) vue la sensibilité et la variabilité des paramètres.

2.2. Pathologies de la parole

Face à ces nombreuses contraintes, le système de production s'adapte par différentes stratégies, ce qui conduit à de la variabilité lors de la réalisation des entités phonétiques. Lorsque les contraintes articulatoires et acoustiques le permettent, la cible est atteinte. Lorsque ces dernières sont trop lourdes, les cibles ne sont pas atteintes de manière optimale et la variabilité apparente est intégrée au message de manière intuitive et "inconsciente" par les locuteurs auditeurs. En d'autres termes, l'articulation est compensée d'une certaine façon, afin d'atteindre la cible acoustique, à défaut de la cible articulatoire. Si aucune des deux n'est atteinte, la perception ne pourra se faire correctement.

2.2.1. Dysphonie

Une dysphonie est un trouble de la voix parlée, et on sait que la voix est un son produit par les CV (**C**ordes **V**ocales) sous l'influence de l'air expiré. Lors de la phonation, les muscles qui composent les CV se rapprochent l'une de l'autre sous contrôle moteur du nerf laryngé inférieur (nerf récurrent), puis l'air contenu dans les poumons est expulsé par la contraction des muscles abdominaux. Cet air fait vibrer de façon passive la muqueuse de recouvrement des CV. L'absence d'accolement de celles-ci est responsable d'une fuite d'air à ce niveau et ainsi d'une voix soufflée même si la muqueuse est normale (Figure 2.1).

De même, des anomalies de la muqueuse (lésion, rigidité, cicatrice, inflammation) entraînent une voix rauque en rapport avec les anomalies vibratoires de la muqueuse même si les mouvements des cordes sont normaux [12].

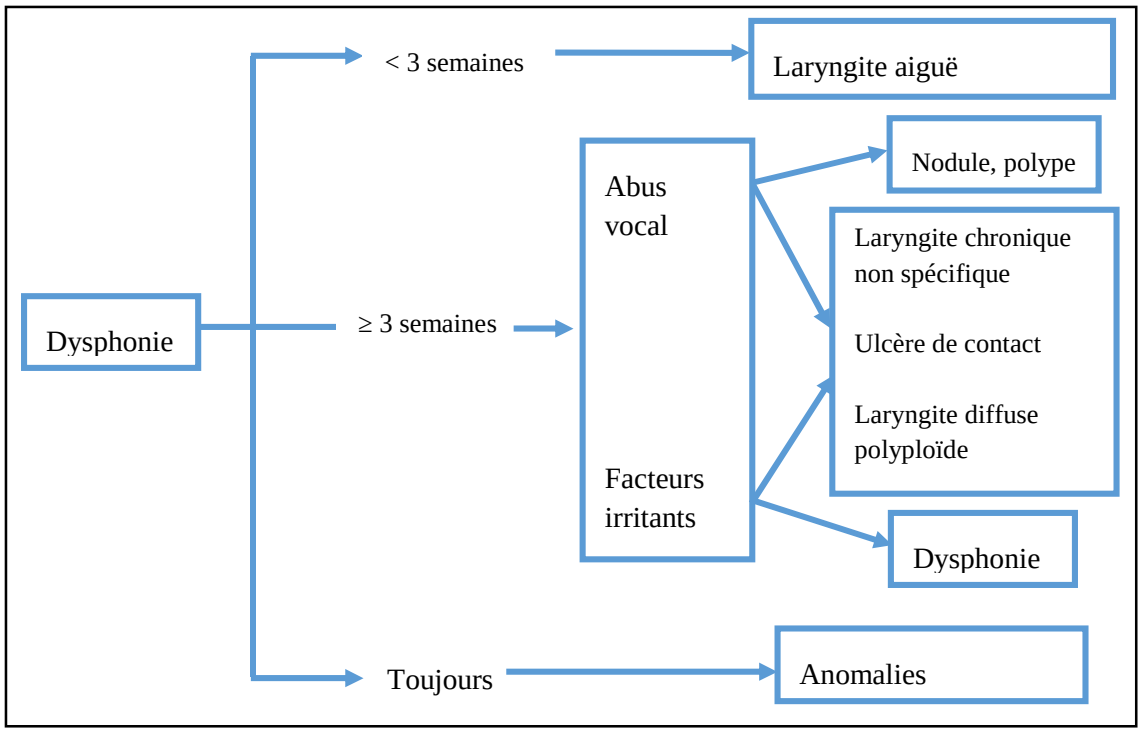


Fig. 2.1 Algorithme Diagnostic de la Dysphonie [13]

Les causes de la dysphonie sont :

Maladies chroniques

- amyloïde (une maladie rare qui se caractérise par la présence de dépôts de protéines insolubles dans les tissus) ;
- polyarthrite rhumatoïde (c'est une maladie dégénérative inflammatoire chronique, elle est caractérisée par une atteinte articulaire souvent bilatérale et symétrique, évoluant par poussées vers la déformation et la destruction des articulations atteintes) ;
- syndrome de Sjôgren (c'est une pathologie chronique auto-immune due à une hyperactivité du système immunitaire à l'encontre de substances ou de tissus qui sont normalement présents dans l'organisme) ;
- tuberculose laryngée (une maladie à déclaration obligatoire prise en charge à 100 %. Elle nécessite idéalement un isolement du patient dès le diagnostic en raison de son caractère bacillifère et jusqu'à la disparition du germe dans les crachats) ;
- maladie de Parkinson (une maladie neurologique chronique dégénérative (perte progressive des neurones) affectant le système nerveux central responsable de troubles essentiellement moteurs d'évolution progressive) ;

- sclérose en plaques (une maladie neurologique auto-immune chronique du système nerveux central. Ses manifestations cliniques sont liées à une démyélinisation des fibres nerveuses du cerveau, de la moelle épinière et du nerf optique) ;
- ataxie de Friedreich (c'est la plus fréquente des ataxies héréditaires d'origine génétique, qui se déclare généralement à l'adolescence. Les traitements actuels permettent de ralentir l'évolution. Le symptôme cardiaque (myocardiopathie) est en général bien pris en charge avec les connaissances actuelles).

Causes endocriniennes

- grossesse ;
- hypothyroïdie ;
- maladie d'Addison (une maladie endocrinienne rare caractérisée par le défaut de sécrétion des hormones produites par les glandes surrénales : glucocorticoïdes (cortisol) et minéralocorticoïdes).

Causes diverses

- Paralysie d'une corde vocale (idiopathique, post-chirurgicale, etc.) ;
- Traumatisme externe ou intubation prolongée ;
- Médicaments (par exemple, administration prolongée d'androgènes) ;
- Tumeurs bénignes.

2.2.2. Dysarthrie

Dysarthrie est un nom collectif pour un groupe de troubles de la parole neurologiques résultant d'anomalies de la force, la vitesse, la gamme, la stabilité, le ton, ou à l'exactitude des mouvements nécessaires pour le contrôle de l'appareil respiratoire, phonatoire, résonatrice, articulatoire, et les aspects prosodiques de production de la parole. Les perturbations physiopathologiques responsables sont dues à des anomalies du système nerveux central ou périphérique et reflètent le plus souvent la faiblesse ; la spasticité ; incoordination des mouvements involontaires, ou excessive, réduit, ou variable tonus musculaire. Cette définition de la dysarthrie est inspirée des travaux de Darley at al. Elle caractérise un trouble de l'exécution motrice de la parole, dont l'origine est une lésion du système nerveux central ou périphérique La dyslalie peut être d'origine fonctionnelle, tels que le bégaiement et le sigmatisme, ou organique, telles que les fentes palatines.

2.2.3. Dyslalie

La dyslalie est un trouble du langage parlé, caractérisé par des difficultés à articuler : impossibilité de prononcer certaines voyelles, permutation des consonnes, etc. Ce trouble est généralement diagnostiqué après l'âge de 4 ans, car avant cet âge, il est normal que l'enfant ait des difficultés d'élocution (en plein apprentissage du langage).

Ces problèmes d'articulation sont généralement dus à des malformations : positionnement anormal de la langue, langue trop grosse (cas typique de la trisomie 21), articulations défectueuses de la mâchoire, malformation du palais, anomalie au niveau du larynx (organe dans la gorge), etc... Un trouble cérébral peut parfois aussi être en cause : le patient ne se concentre pas suffisamment pour activer correctement les muscles de la mâchoire. Le traitement de la dyslalie varie en fonction de sa cause : opération chirurgicale, séances d'orthophonie.

2.2.3.1. Bégaiement

Le bégaiement n'est pas seulement un problème d'élocution ou de langage, c'est surtout un problème de communication. C'est le fait d'interrompre la parole ou de changer la façon naturelle de l'exprimer. C'est aussi le fait de ne pas pouvoir exprimer ce que l'on avait prévu de dire initialement. Le bégaiement peut se manifester dans la répétition d'une syllabe (façon la plus connue), en "tirant" sur les mots ou en ne pouvant pas dire la phrase, le mot ou une partie de celui-ci au moment voulu. Des tensions musculaires peuvent aussi accompagner le bégaiement : spasmes respiratoires et/ou mouvements involontaires du visage (grimaces) et du corps.

Selon les jours ou les périodes, les manifestations du bégaiement sont plus ou moins visibles. Pourtant, le bégaiement apparaîtra rarement si la personne bègue parle quand elle est seule, murmure, chante, parle avec un animal ou un bébé ou encore en prenant un accent étranger. La manifestation de ce trouble est différente selon les individus.

La personne qui bégaie peut également user de périphrases, consciente de sa difficulté voire de son impossibilité à prononcer le mot ou la syllabe voulue. Alors que certaines personnes ont des peurs diverses (ascenseur, foule, autoroute, hauteur, etc.), la personne qui bégaie peut, elle, redouter certaines lettres ou certaines syllabes [14].

2.2.3.2. Sigmatisme

Le sigmatisme est la mauvaise articulation des consonnes, en particulier les fricatives. C'est un des troubles dyslaliques les plus fréquents chez l'enfant, car ce type de consonnes nécessite une précision très importante de l'articulation. Selon son origine, nous pouvons classifier le sigmatisme en plusieurs classes :

- sigmatisme nasal, dû à un positionnement de la langue qui rend impossible le passage de l'air par la cavité buccale ;
- sigmatisme dorsal, dû à un soulèvement excessif de la langue ;
- occlusif, dû à un remplacement systématique de toute consonne fricative par la consonne occlusive dont le point d'articulation est le plus proche [10].

2.3. Fentes palatines

Nous présentons (Figure 2.4) les principales parties du palais, qui peuvent être déformées par la fente, et qui engendrent des conséquences sur la production de la parole.

2.3.1 Anatomie du palais

- la voûte palatine présente une forme concave en bas, elle mesure environ 4-5 cm de large et 7-8 cm de long, chez l'adulte, avec une flèche de 1.5 cm. Sa forme ogivale lui donne sa solidité. La voûte est constituée d'un plan osseux, recouvert d'une muqueuse avec des glandes salivaires accessoires.
- Le voile du palais est situé en arrière du palais dur et est constitué d'un élément central, l'uvule palatine. Deux replis se détachent de cette partie : un en avant vers la langue, le repli ou l'arc palato-glosse, et un en arrière vers le pharynx, le repli ou l'arc palato-pharyngien. Entre les deux arcs existe une formation lymphoïde appelée la tonsille.

Le voile du palais est constitué d'une charpente fibreuse, une muqueuse, et des muscles pairs, au nombre de 5 : le muscle palato-glosse, le muscle tenseur du voile, le muscle palato pharyngien, le muscle élévateur du voile, le muscle uvulaire (azygos de la luette). Chacun ayant un rôle très important dans la production de la parole.

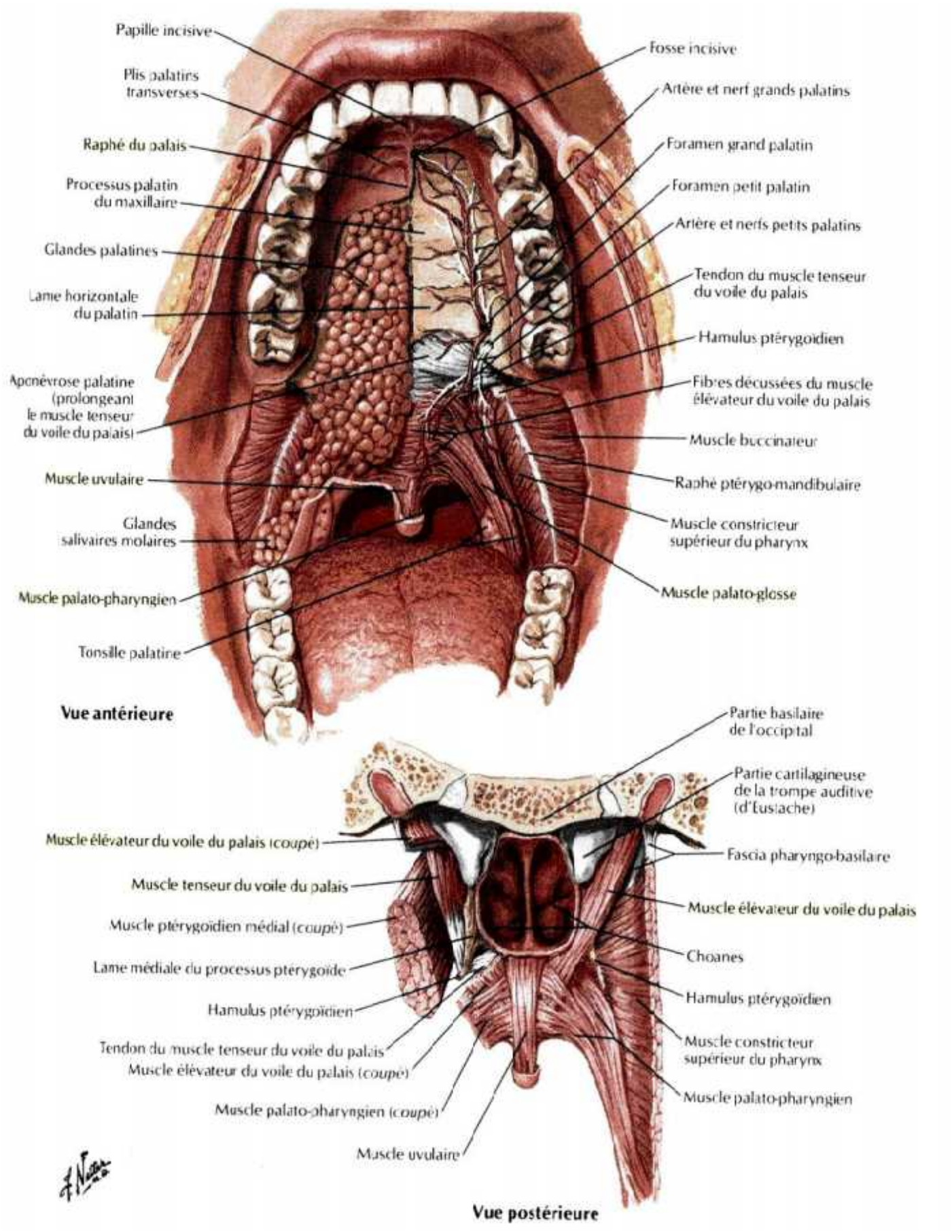


Fig. 2.2 : Vues postérieure et antérieure des principaux muscles du palais [15]

2.3.2 Fonction du palais

Le voile du palais intervient sur quatre fonctions, à savoir :

- la respiration : les muscles du voile du palais jouent un rôle important dans l'établissement d'une respiration buccale et/ou nasale ;
- la déglutition : son rôle est d'assurer la séparation des fosses nasales et de la cavité buccale, ceci évitant les reflux alimentaires par le nez ;
- la phonation : le voile du palais peut être relevé, l'articulation est alors orale. Il peut également être abaissé, dans ce cas l'articulation est nasale, et les cavités buccale et nasale sont en communication ;
- l'audition : l'ouverture de la trompe d'Eustache, qui permet l'aération de la caisse tympanique et l'équilibration des pressions de chaque côté du tympan, est permise par la contraction des muscles élévateurs du voile du palais.

Le palais dur joue quant à lui, un rôle important dans l'articulation des phonèmes postérieurs, mais il participe également à la résonance des sons puisqu'il définit la limite supérieure de la cavité buccale.

Il permet par ailleurs de créer le vide attendu lors de l'aspiration des liquides, cette étanchéité étant nécessaire à la succion.

La fonction des lèvres est aussi importante à la succion, puisqu'elles permettent la préhension et l'étanchéité de la bouche nécessaires à une tétée de qualité. Les lèvres sont également impliquées dans l'articulation des phonèmes labiaux.

Une atteinte anatomique à l'un ou à ces deux niveaux engendre donc des conséquences fonctionnelles. Nous décrirons ici les principales déformations les plus connues des formes de fentes faciales ainsi que les conséquences.

2.3.3 Classification des fentes

Devant la multitude de cas existants, de nombreuses méthodes ont été développées pour l'enregistrement de ces déformations des lèvres et du palais. Aucune de ces méthodes n'a pu être universellement acceptée à cause de leurs limites, des descriptions inadaptées aux déformations du palais et à la complexité variante liée à ces déformations.

Il reste cependant important de trouver une méthode de classification des fentes. En effet, cela permettrait une catégorisation des cas et une prise en charge clinique plus aisée. Mais la grande variabilité liée à ce type de pathologie rend cette classification très difficile, comme nous le verrons par la suite.

Kernahan a proposé la classification Y, désigné pour décrire les informations détaillées par rapport aux déformations du palais. La description anatomique des composants du palais y est notée à l'aide de 5 chiffres arabes dans l'ordre suivant :

- lèvre droite (R),
- alvéole du côté droit et palais primaire (A),
- second palais (P),
- alvéole gauche et premier palais (A),
- et lèvre gauche (L).

Selon ce chercheur, cette classification est simple et précise, donc facile à comprendre et pourrait être utilisé pour combiner les analyses des données (Figure 2.5).

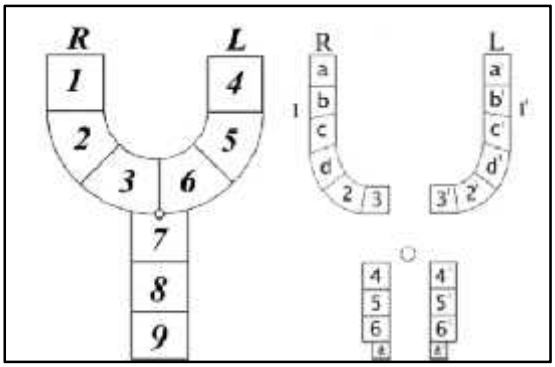


Fig. 2.3 : Malformations faciales (Classification Y) [16]

2.3.3.1 Division simple du voile

Elle intéresse le palais mou et peut être partielle ou totale. La moins importante est la luette bifide, mais une muqueuse vélaire normale peut masquer une fente sous-muqueuse qui est caractérisée par un aspect transparent sur la ligne médiane (Figure 2.6).

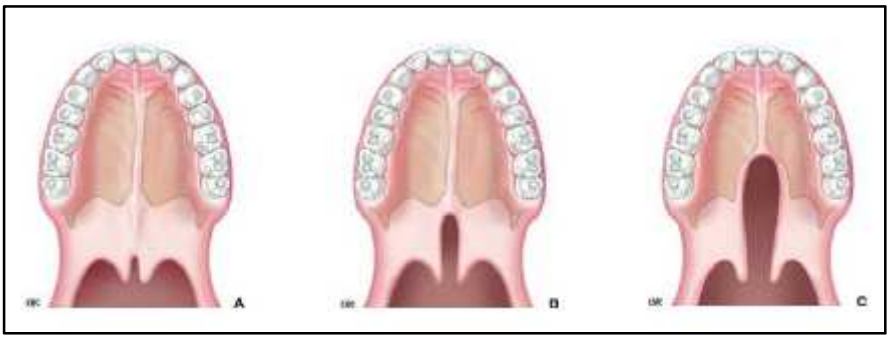


Fig. 2.4 : A- Fente vélaire partielle (luette bifide)
B- Fente vélaire partielle
C- Fente vélaire totale

2.3.3.2. Division du voile et de la voûte palatine

Elle se prolonge en avant jusqu'au foramen incisif.

.

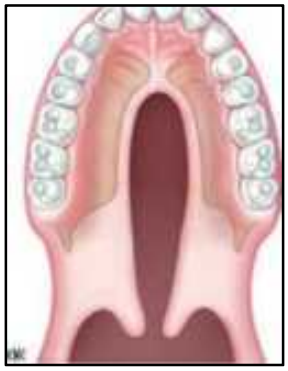


Fig. 2.5 : Fente vélo-palatine [17]

2.3.3.3. Division du voile associée à une fente labio-alvéolaire

La fente intéresse la lèvre et le procès alvéolaire qu'elle franchit dans la région distale de l'incisive latérale, zone de fusion entre le bourgeon médian et le bourgeon maxillaire. Cette dent est de ce fait fréquemment dédoublée ou absente.

Le palais est divisé en deux fragments différents : un grand fragment (côté opposé à la fente) comprenant la région incisive et la moitié du palais dur, et un petit fragment (côté adjacent à la fente). Le vomer est partiellement ou complètement fusionné au grand fragment.

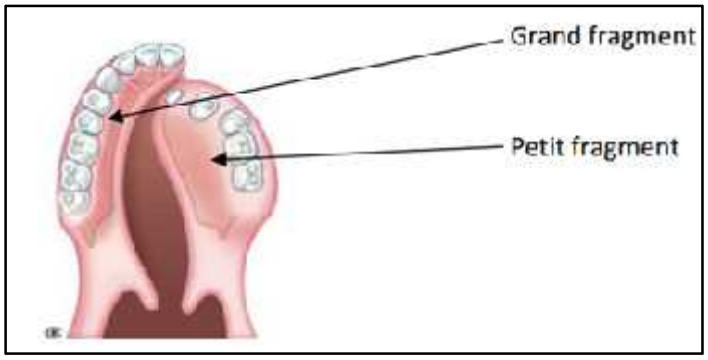


Fig. 2.6 : Fente complète unilatérale [17]

2.4 Etude de la fente palatine simple

L'incompétence vélo-pharyngée peut entraîner des troubles de l'articulation :

- des phénomènes de compensations, qui sont des mécanismes qui consistent à tenter de compenser l'insuffisance vélaire. La fuite d'air nasale est massive et empêche

toute possibilité de pression intra-buccale. Le sujet cherche alors des points d'occlusion et de constriction en amont du vélo-pharynx ;

- des coups de glotte : c'est une attaque dure qui peut remplacer les occlusives. Les cordes vocales s'accolent, l'air s'accumule dans la trachée et, à la séparation des cordes vocales, on entend une explosion glottale ;
- attaques vocaliques dures ;
- souffles rauques : les constrictives sont remplacées par une constriction produite dans la région sus-laryngée, l'air passant en sifflant entre les cordes vocales très rapprochées ;
- remplacement de phonèmes par des phonèmes voisins par leur point d'articulation ou leur mode d'articulation (par exemple [t] devient [k], [z] est désonorisé...). La postériorisation des occlusives (souvent [k]-[g] pour [t]-[d]) se rencontre particulièrement chez les fentes bilatérales totales. Elle sera rééduquée avec attention ;
- articulation atypique des phonèmes (par exemple [t] interdental, [s] et [z] sont schlintés...) ;
- sigmatisme nasal : les constrictives sont nasalisées car la langue se recule et oblige l'air à passer par le nez.

En ce qui concerne la longueur des phrases, les enfants ont tendance à utiliser des phrases plus courtes, peut-être à cause d'une moins bonne utilisation de l'air expiratoire.

L'assimilation et la durée

L'analyse des données de Abdelli-Beruh [19] révèle que les occlusives [b d g] ont des durées plus longues que leurs homologues non voisées. De plus, elles sont précédées de voyelles plus longues que les occlusives non voisées [p t k] dans trois conditions : en initial de syllabe entre voyelles, entre fricatives non voisées et entre voyelles en fin de syllabe.

En Français, lorsque deux consonnes s'opposent au niveau du trait de voisement sont en contact, l'assimilation est habituellement régressive ; c'est-à-dire que la seconde consonne va influencer le trait de voisement de la première.

Le tableau 2.1 montre qu'au niveau acoustique, les indices temporels et spectraux différencient les occlusives sonores [b d g] des non voisées [p t k]. En effet, il apparaît que les consonnes voisées ont une durée plus longue. Aussi, ces dernières influencent la durée des voyelles les précédents, qui ont elles aussi une durée plus importante.

Tab. 2.1 : Caractéristiques des occlusives en fonction de leur environnement [19]

Phonèmes	**Durée de l'occlusive**	***« phonated »* (voisée)**	**Durée V. précédente**
/ptk/	- longue	- souvent	- longue
/bdg/	+ longue	+ souvent	+ longue

2.5 Conclusion

Dans cette partie, nous avons exposé les pathologies de la parole les plus fréquentes. Nous avons décrit, en particulier, le cas de déformation des palais et nous avons pris le cas de "la fente palatine simple". Une description détaillée de ce cas de pathologie nous a permis d'avoir une meilleure compréhension des troubles phonatoires que subissent les patients.

CHAPITRE 3 :

Analyse de la Parole Pathologique

3.1. Introduction

Dans cette partie, nous allons décrire brièvement les différentes méthodes d'analyse existant en traitement de la parole qu'elles soient paramétriques ou non paramétriques, tout en insistant sur la méthode du GMM puis la modélisation.

3.2. Analyse du parole pathologique

Le signal de la parole est un signal très complexe. Il contient une quantité importante d'informations imbriquées entre elles.

L'analyse acoustique du signal de parole consiste à extraire l'information pertinente et à réduire au maximum la redondance. Généralement, on calcule un jeu de coefficients acoustiques à des intervalles de temps réguliers, sur des blocs de signal de longueur fixe. Ce jeu de coefficients constitue un vecteur acoustique.

3.2.1 Paramètres acoustiques de la Parole Pathologique

Une Parole Normale (PNorm) est analysée essentiellement par l'observation des paramètres principaux :

- la fréquence fondamentale F_0, qui permet de mesurer les vibrations des cordes vocales ;
- les formants qui permettent d'étudier les effets que subissent les sons de parole lors de leurs passages à travers les cavités vocales ;
- la durée des sons pour étudier le débit d'air et la fluidité de la parole ;
- l'intensité qui permet de distinguer un son fort d'un son faible.

Cependant, l'analyse d'une Parole Pathologique (PPath) fait appel, en addition, à d'autres paramètres aussi importants tels que le degré de perturbation de F_0 (Jitter) et le degré de perturbation de l'intensité (Shimmer), qui sont très exploités pour la caractérisation de la qualité de la PPath [20]. Ces deux paramètres sont habituellement mesurés sur les voyelles soutenues, et leurs valeurs au-dessus d'un certain seuil sont considérées comme étant liées à des PPath.

Les principales classifications des méthodes de traitement du signal vocal sont :

- les transformées usuelles comme la Transformée Discrète de Fourier qui ne se réfère pas à un modèle de production ni de perception ;

- les méthodes fondées sur la déconvolution « source - conduit vocal » cepstre et codage prédictif linéaire qui s'appuient sur le modèle de production de la parole [5].

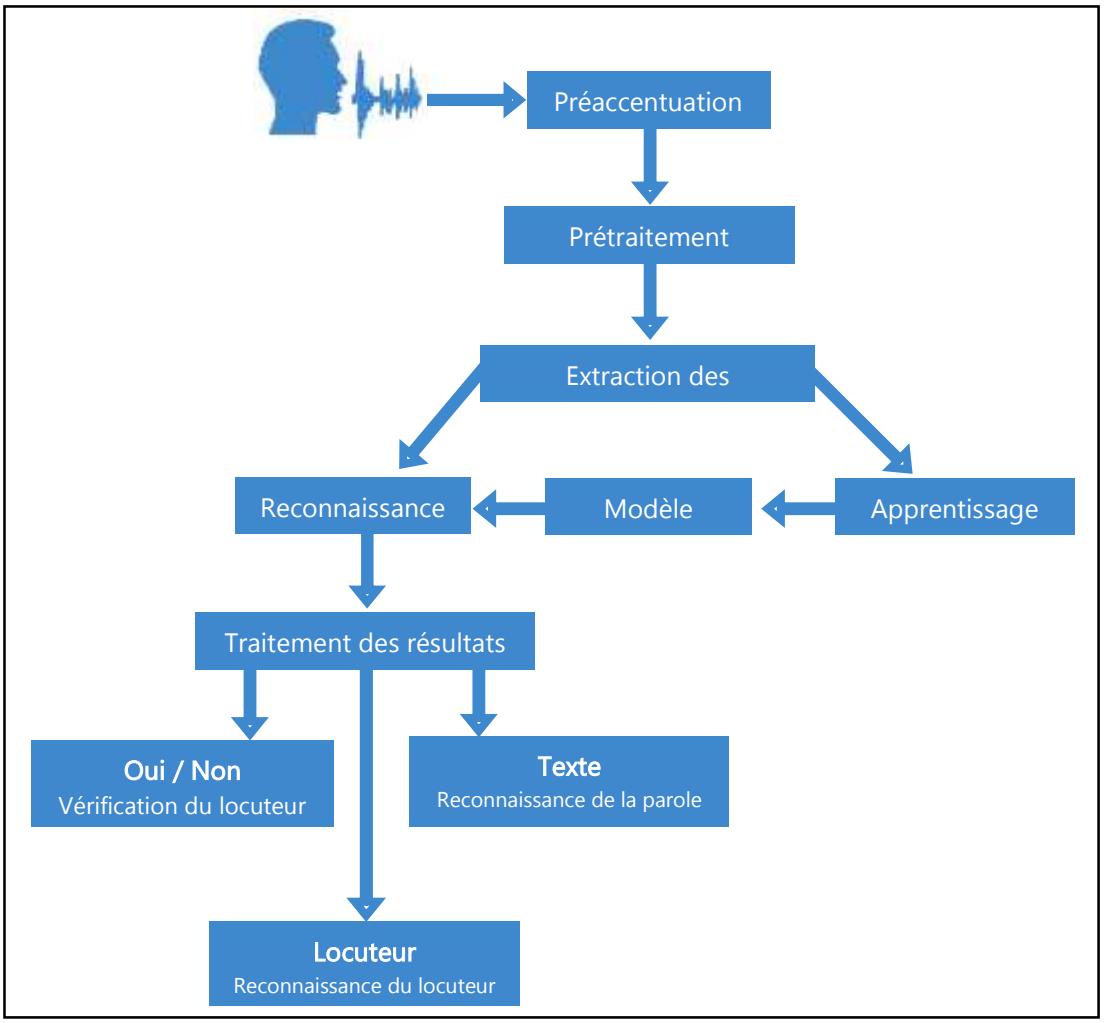


Fig. 3.1 : Organigramme de la procédure d'analyse d'un signal vocal

3.2.2. Prétraitement acoustique

Le prétraitement est utilisé généralement pour la mise en forme du signal brut avant le traitement il se compose de l'adaptation et le fenêtrage.

3.2.2.1. Préaccentuation

L'onde acoustique sortante des lèvres subit, à cause de la désadaptation entre les deux milieux intérieur et extérieur, une distorsion assimilable à une désaccentuation de 6 par octave sur tout le spectre [21]. Pour pouvoir compenser cette distorsion, et accentuer les hautes fréquences, on applique un filtre de préaccentuation passe haut de transmittance.

$$H(z) = 1 - \alpha z^{-1} \tag{3.1}$$

Avec $0.9 \leq \alpha \leq 1$.

3.2.2.2. Fenêtrage

L'étape de fenêtrage consiste à appliquer au signal vocal une fenêtre glissante de durée limitée, et ce afin de limiter le nombre d'échantillons et de réduire les effets de bords (phénomène de Gibbs).

Parmi les différentes fenêtres de pondération, les plus utilisées sont : la fenêtre rectangulaire, la fenêtre de Hamming, la fenêtre de Hanning et la fenêtre de Blackmann. En traitement de la parole, la fenêtre de Hamming est la plus utilisée.

Cette fenêtre est donnée par l'expression (3.2) :

$$w(n) = 0.54 + 0.46\cos\frac{2\pi n}{N-1} \tag{3.2}$$

Avec $0 \leq n \leq N-1$
N : Le nombre d'échantillons dans une fenêtre

3.2.3. Méthode d'analyse en traitement de parole

3.2.3.1. Coefficients de prédiction linéaire LPC

Le principe fondamental de la prédiction linéaire est qu'un échantillon donné peut être prédit à partir d'une combinaison linéaire des échantillons finis qui le précèdent. Un seul jeu de coefficients du prédicteur est déterminé en minimisant les différences entre les échantillons actuels et ceux prédits. La technique de prédiction linéaire est basée sur le modèle de la production de la parole.

La fonction de transfert du modèle de la production de la parole est décrite par :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + G \quad (n) \tag{3.3}$$

3.2.3.2 Paramètres LSP

Les paramètres LSP (Line Spectral Pair) ont été présentés la première fois par Itakura Comme représentation alternative d'information spectrale du LPC. Ils contiennent exactement la même information que Les coefficients LPC [22].

En analyse par prédiction linéaire, un segment de parole est supposé être généré comme sortie d'un filtre tous pôles H(z)= 1/A(z).Où A(z) est un polynôme en z appelé le filtre inverse dont l'expression est donnée par l'expression (3.4) :

$$A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p} \tag{3.4}$$

3.2.3.3. Coefficients cepstraux de prédiction linéaire LPCC

Les coefficients cepstraux peuvent être calculés à partir de la sortie d'un banc de filtres ou à partir des coefficients de prédiction linéaire, ainsi les coefficients **LPCC** (**L**inear **P**rediction **C**epstral **C**oefficients) sont dérivés directement des coefficients LPC.

Les coefficients cepstraux sont obtenus :

$$c_K = -a_K - \sum_{i=C}^{K-1} \left(1 - \frac{i}{K}\right) a_i c_{K-i} \quad (3.5)$$

Tel que : $k > 0$.

3.2.3.4. Coefficients MFCC (Mel Frequency Cepstral Coefficients)

Les coefficients cepstraux issus d'une analyse par Transformée de Fourier, caractérisent bien la forme du spectre et permettent de séparer l'influence de la source glottique de celle du conduit vocal.

Le cepstre du signal de parole est défini comme étant la Transformée de Fourier Inverse du logarithme de la densité spectrale de puissance. Pour ce signal, la source d'excitation glottique est convoluée avec la réponse impulsionnelle du conduit vocal [23].

$$s(t) = e(t) * h(t) \quad (3.6)$$

où *s(t)* est le signal de parole, *e(t)* est la source d'excitation glottique et *h(t)* est la réponse impulsionnelle du conduit vocal.

L'application du logarithme sur le module de la Transformée de Fourier de dans l'équation précédente donne :

$$\log|s(f)| = \log|e(f)| + \log|h(f)| \quad (3.7)$$

Par une Transformée de Fourier Inverse, on obtient :

$$s'(c\quad) = e'(c\quad) + h'(c\quad) \quad (3.8)$$

La dimension du nouveau domaine est homogène à un temps et s'appelle la quéfrence (*cef*), le nouveau domaine s'appelle donc le domaine quéfrentiel. Un filtrage dans ce domaine s'appelle liftrage.

Ce domaine est intéressant pour faire la séparation des contributions du conduit vocal et de la source d'excitation dans le signal de parole. En effet, si les contributions relevant du conduit vocal et les contributions de la source d'excitation évoluent avec des vitesses différentes dans le temps, alors il est possible de les séparer par l'application d'un simple fenêtrage dans le domaine quéfrentiel (liftrage passe-bas) pour le conduit vocal.

Les coefficients cepstraux les plus répandus sont les MFCC. Ils présentent l'avantage d'être faiblement corrélés entre eux, et qu'on peut donc approximer leur matrice de covariance par une matrice diagonale.

Pour simuler le fonctionnement du système auditif humain, les fréquences centrales du banc de filtres sont réparties uniformément sur une échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'informations utiles dans le signal de parole. Les échelles perceptives les plus utilisées sont l'échelle Mel et l'échelle Bark [23].

- Echelle Mel (Figure 3.2)

$$M \quad (f) = 2595 \log(1 + \frac{f}{7}) \tag{3.9}$$

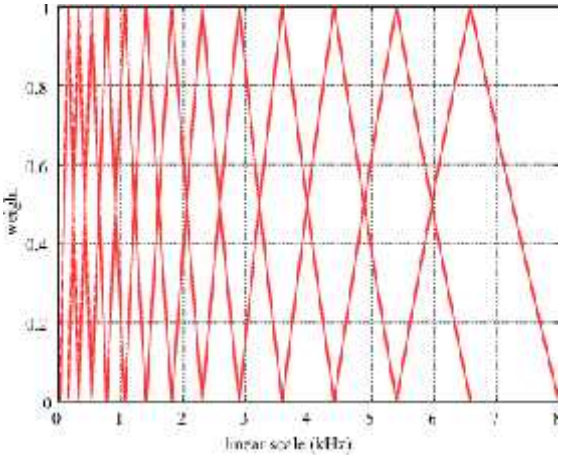


Fig. 3.2 : Banc de filtres sur l'échelle linéaire

- Echelle Bark (Figure 3.3)

$$B \quad (f) = 6\,A \quad hs\imath \quad (\frac{f}{1}) \tag{3.10}$$

f représente la fréquence [Hz].

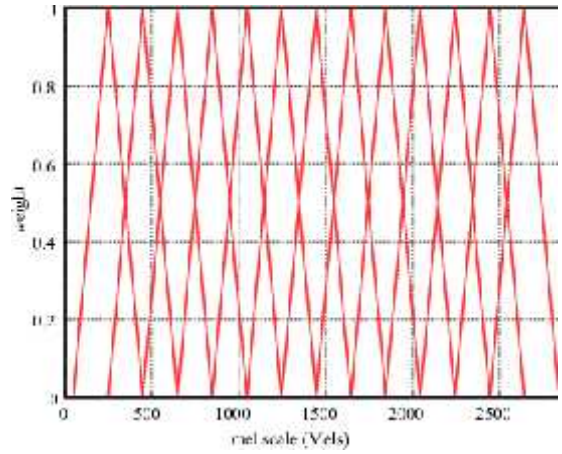


Fig. 3.3 : Le banc de filtres sur l'échelle Mel

La procédure de calcul des coefficients MFCC est illustrée dans la figure 3.4

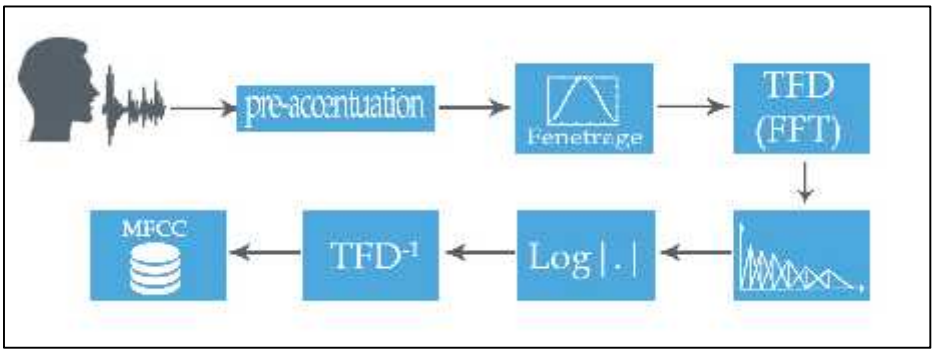


Fig. 3.4 : Schéma de calcul des MFCC

Soit un signal discret *s(n)* avec $0 \leq n \leq N-1$, *N* est le nombre d'échantillons d'une fenêtre d'analyse, F_e est la fréquence d'échantillonnage, la Transformée de Fourier Discrète à court terme *S(k)* est obtenue avec la formule :

$$S(k) = \sum_{n=0}^{N-1} s(n)\ e^{\left(\frac{-j\,2\pi\,n\,k}{N}\right)} \qquad (3.11)$$

avec $0 \leq k \leq N-1$;

Le spectre du signal est filtré par un banc de filtres triangulaires, dont les bandes passantes sont de même largeur dans le domaine des fréquences Mel. Les points de frontières des filtres en échelle de fréquence Mel sont calculés à partir de la formule :

$$B_m = B_b + m\,\frac{B_h - B_E}{M+1} \qquad (3.12)$$

$0 \leq m \leq M+1$;
M : Le nombre de filtres ;
B_h : La fréquence la plus haute du signal ;
B_b : La fréquence la plus basse du signal.
Dans le domaine fréquentiel, les points discrets correspondants sont calculés d'après :

$$f_m = B^{-1}\left(B_b + m\,\frac{B_h - B_E}{M+1}\right) \qquad (3.13)$$

Les coefficients cepstraux de fréquence en échelle Mel peuvent être obtenus par une Transformée de Fourier Inverse à partir des énergies d'un banc de filtres. Les premiers Coefficients cepstraux peuvent être calculés directement à partir du logarithme des énergies issues d'un banc M filtres par la transformée en cosinus discrète définie par :

$$c_k = \sum_{n=0}^{N-1} l\epsilon\quad E_i\,c\quad \frac{\pi}{M}\left(i - \frac{1}{2}\right), \qquad (3.14)$$

avec $1 \leq k \leq d$.

Ce qui permet d'obtenir des coefficients peu corrélés.

3.3. Modélisation des locuteurs

Le problème dans cas de la reconnaissance de la parole peut se formuler selon un problème de classification. Différentes approches ont été développées, néanmoins on peut les classer en quatre grandes familles [22], il s'agit de l'approche :

- vectorielle ;
- statistique ;
- connexionniste ;
- relative.

3.3.1. Approche vectorielle

Le signal du locuteur est modélisé par un ensemble de vecteurs de paramètres dans l'espace acoustique. Ses principales techniques sont la reconnaissance à base de DTW (**D**ynamique **T**ime **W**arping) et par la QV **(Q**uantification **V**ectorielle**)**.

3.3.1.1. Reconnaissance du locuteur à base de DTW

La reconnaissance par DTW repose sur le principe que chaque mot est représenté par une prononciation de référence (template). Compte tenu des décalages temporels entre les différentes prononciations d'un même mot, l'algorithme met en correspondance des séquences de paramètres par distorsion temporelle (Time Warping). La programmation dynamique permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage ce qui signifie que c'est une technique exclusivement utilisée en mode dépendant du texte [24].

3.3.1.2. Quantification vectorielle

Il s'agit de représenter l'espace acoustique par un nombre fini de vecteurs acoustiques. Cela consiste à faire un partitionnement de cet espace en régions, qui seront représentées par leur vecteur centroide. Pour déterminer la distance d'un vecteur acoustique à cet espace, on effectue une mesure de distance avec chacun des centroides des régions et on retient la distance minimale. Si le vecteur acoustique provient du même locuteur pour lequel on a établi le dictionnaire de quantification, la distorsion sera en général moins grande que si ce vecteur provient d'un autre locuteur. Ainsi, on va représenter un locuteur par son dictionnaire de quantification [25].

3.3.2. Approche statistique

Consiste à représenter le signal de chaque locuteur par une densité de probabilité dans l'espace des paramètres acoustiques. Elle couvre les techniques de modélisation par chaînes de Markov cachées, par les mélanges de gaussiennes et par des mesures statistiques de second ordre.

3.3.2.1 Modèles de Markov cachés

Les modèles de Markov cachés (ou **HMM** pour **H**idden **M**arkov **M**odels) ont été introduits en reconnaissance de la parole récemment. Dans cette dernière approche, il ne s'agit plus d'une mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme acoustique ait été engendrée par le modèle de référence du locuteur. Le modèle d'un locuteur est constitué de l'association d'une chaîne de Markov (Figure 3.7), une succession d'états avec des probabilités (probabilité d'observation d'un vecteur acoustique dans un état) [26].

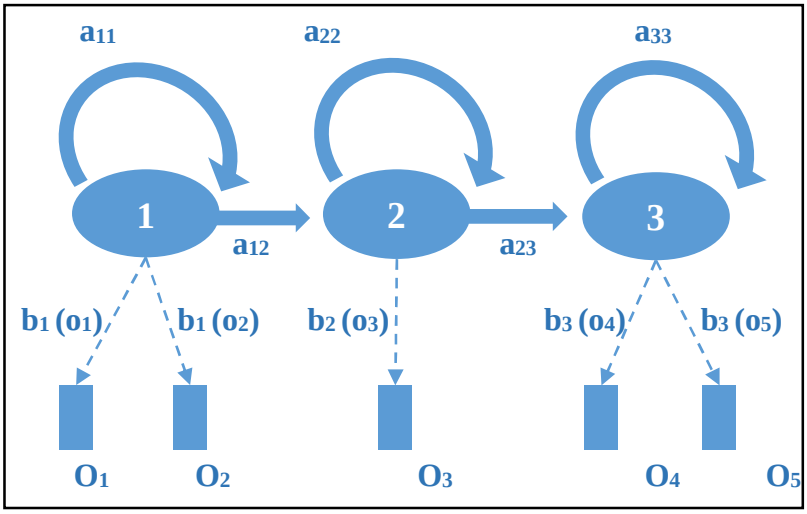


Fig. 3.7 : HMM gauche à droite a trois états.

3.3.2.2 Les Mélanges de Gaussiennes

La reconnaissance du locuteur par mélange de gaussiennes (ou **GMM** pour **G**aussien **M**ixture **M**odels) consiste à modéliser le signal d'un locuteur par une somme pondérée de composantes gaussiennes. Ainsi une large gamme de distributions peut être parfaitement représentée. Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques (Figure 3.8).

L'utilisation de ce type de modèles semble être prometteuse. Il semble bien modéliser les caractéristiques spectrales des voix des locuteurs, et il est relativement simple à mettre en œuvre. Les mélanges de gaussiennes sont considérés comme un cas particulier des HMM et une extension de la quantification vectorielle [27].

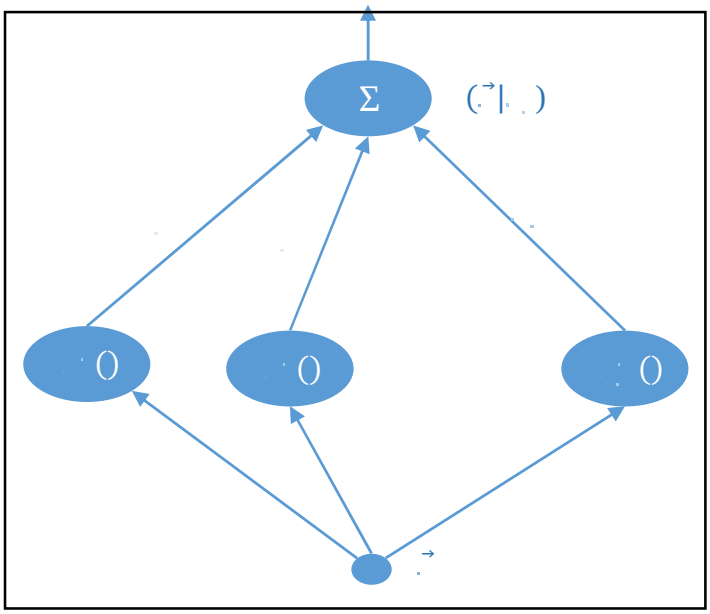


Fig. 3.8 : Modèle de GMM

3.4. Reconnaissance par le Mélange de Gaussiennes

Les GMM sont une famille de classificateurs où on suppose que la distribution des données pour chaque classe est une combinaison de plusieurs distributions Gaussiennes dans l'espace de représentation. L'apprentissage des GMM se fait généralement par l'algorithme d'**E**spérance **M**aximale (EM) qui garantit théoriquement une convergence vers une solution optimale. L'algorithme EM permet un calcul itératif des paramètres de chaque Gaussienne.

Les GMM sont utilisés pour modéliser un locuteur donné par une somme pondérée de gaussiennes. On peut assimiler un modèle GMM à un HMM à un seul état. On ne modélise donc pas les aspects temporels du signal. Cette méthode est la plus utilisée en ce qui concerne la reconnaissance du locuteur dans le cas de la parole pathologique due à sa flexibilité au type de signal et son bon compromis entre les performances du système en termes de précision et la vitesse et la complexité des algorithmes [28].

Pour obtenir une modélisation pertinente des caractéristiques d'un locuteur ce GMM est entraîné à partir des vecteurs issus du signal de parole de ce locuteur (on utilise par exemple les MFCC). S'il existe plusieurs techniques permettant de calculer les

paramètres des GMM, la plus courante consiste à maximiser la vraisemblance en utilisant l'algorithme EM (Expectation-Maximization) couplé à une entité de maximisation de la vraisemblance (ML pour Maximum Likelihood).

Les GMM consistent en la modélisation, pour chaque modèle, des données x_t sous la forme d'une somme pondérée par les coefficients k de fonctions de densité de probabilité gaussiennes $p(x_t, \mu_k, \Sigma_k)$:

$$P\left(X/\lambda\right) = \sum_{k=1}^{K} w_k \, p(x_t, \mu_k, \Sigma_k) \tag{3.15}$$

Avec K est le nombre de composantes de densité considéré pour le modèle. Chaque composante s'exprime en fonction de sa moyenne μ_k et de sa matrice de covariance $\sum$:

$$p(x_t, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} e^{\left[-\frac{1}{2}(x_t - \mu_k)'\Sigma_k^{-1}(x_t - \mu_k)\right]} \tag{3.16}$$

La matrice de covariance utilisée est diagonale, c'est -à-dire les modèles sont appris en considérant les observations associées à chacun des descripteurs de manière indépendante.

Pour chaque modèle, chacune des composantes du mélange modélise une région différente de l'espace des données appelée aussi cluster.

L'apprentissage consiste en l'estimation à partir des observations d'une même classe des paramètres des gaussiennes qui composent le modèle de cette classe.

Les paramètres à estimer sont :

- les associés poids $(w_k)_{k=1,2\ldots,K}$ à chacune des k composantes du mélange ;
- les moyennes et matrices de covariance de chacune des composantes du mélange :

$$(\mu_k, \Sigma_k)_{k=1,2\ldots,K}$$

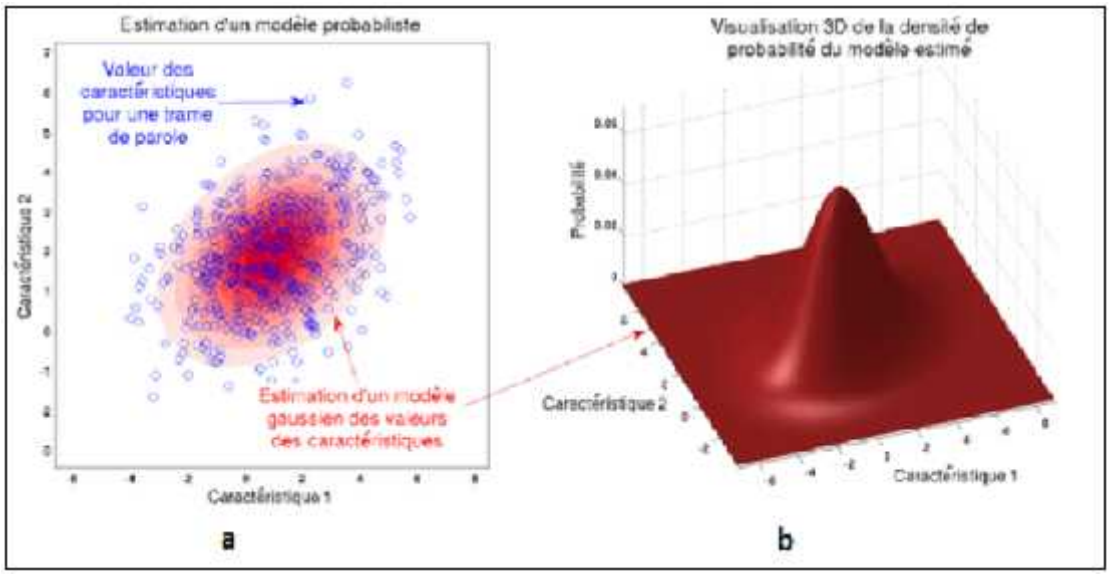


Fig. 3.9 : Les modèles probabilistes : exemples d'une distribution Gaussienne à 2 dimensions

Sur Figure 3.9 -a, on observe en bleu des valeurs de 2 caractéristiques .Chaque rond bleu représente à un instant en abscisse et en ordonnée les valeurs respectives du premier et du second coefficient Cepstral. La répartition des valeurs dans le plan peut alors être modélisée sous la forme d'une distribution Gaussienne, représentée en rouge.

Figure 3.9 -b montre la distribution Gaussienne estimée. La hauteur de la distribution représente la probabilité d'une caractéristique de prendre cette valeur. C'est pourquoi on parle de modèles probabilistes (ou statistiques).

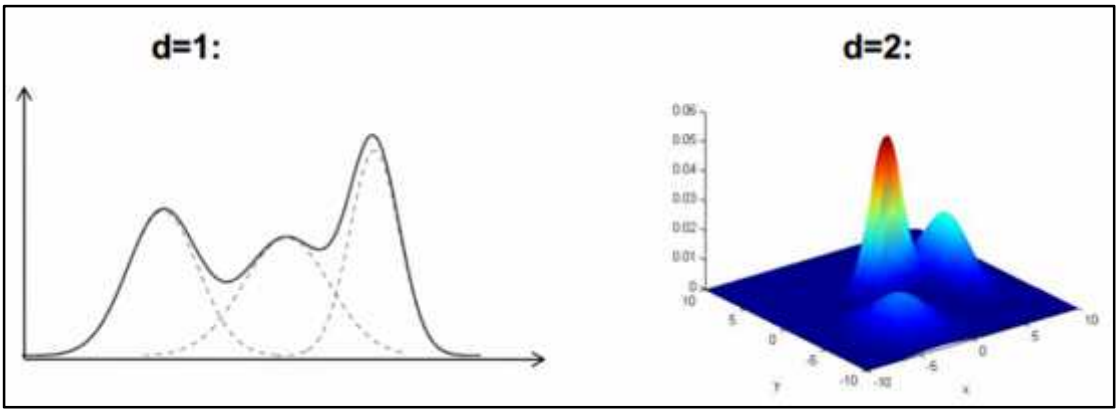


Figure 3.10 : Les modèles probabilistes : exemple d'un modèle Gaussien à 2 dimensions

- a montre un exemple de GMM à 3 gaussiennes en 1 dimension.
- b illustre la manière dont peut être adaptée une signature vocale à partir du modèle du monde.

3.4.1. Estimation des paramètres

Choisi, vis à vis des données d'apprentissage, le critère le plus utilisé pour l'apprentissage des GMM, est le critère de Maximum de Vraisemblance ML (**M**aximum **L**ikelihood). L'estimation des paramètres des GMM consiste à trouver ceux qui maximisent la fonction de vraisemblance des données d'apprentissage.

$$\vec{\lambda}_x = a \qquad {}_{\lambda}p\left(\vec{X}_t/_{\lambda}\right) \tag{3.17}$$

3.4.1.1. Algorithme des K-moyennes

L'algorithme des K-moyennes consiste à faire la répartition des vecteurs acoustiques d'une classe (locuteur) en N sous-ensemble disjoints caractérisés par un centroïde. Le résultat de cette répartition est appelé dictionnaire. L'algorithme des K-moyennes n'est que localement optimal, par conséquent, il est influencé par ses conditions initiales. L'algorithme des K-moyennes est défini comme suit :

- La première étape est l'initialisation du dictionnaire, il existe plusieurs méthodes d'initialisation comme l'initialisation aléatoire ou bien l'algorithme à seuil.
- La deuxième étape consiste à appliquer deux règles, tant qu'il y'a une amélioration importante de la distorsion moyenne donnée par la formule suivante :

$$D_m = \frac{1}{N}\sum_{K=1}^{N} d\left(\vec{X}_K, C\left(\vec{X}_K\right)\right) \tag{3.18}$$

$C\left(\vec{X}_K\right)$ est le centroïde de la région où est affecté.
$d\left(\vec{X}_K, C_i\right)$ est la distance euclidienne entre les vecteurs $\vec{X}_K$ et C_i.

Les deux règles sont définies comme suit :

- **La règle de centroïde**

Cette règle exige que tous les centroïde soient les moyennes des vecteurs acoustiques des régions représentées par ces centroide. Cela peut être formulé comme suit :

$$C_i = \frac{1}{N}\sum_{K=1}^{N_i} \vec{X}_K \tag{3.19}$$

- **La règle de plus proche voisin**

Le $\vec{X}_K$ vecteur est affecté à la région *i* si la distance euclidienne entre ce vecteur et le centroide de cette région est minimale. La formule suivante décrit cette règle.

$$r \qquad \left(\vec{X}_K\right) = i, C_i = m \quad \left(d\left(\vec{X}_K, C_i\right)\right) \tag{3.20}$$

K est le nombre de régions.
C_i est le centroïde de la $\vec{X}_K$ région.

3.4.1.2. Algorithme EM (Expectation Maximisation)

L'algorithme EM permet de régler les paramètres d'un modèle de distribution GMM pour atteindre un maximum de vraisemblance d'un ensemble d'observations. Ces dernières sont typiquement des vecteurs d'apprentissage non étiquetés. Les paramètres libres sont constitués par des poids, des vecteurs moyens et des matrices de covariance [29].

L'algorithme EM permet l'estimation de ces paramètres. Cet algorithme itératif garantit la croissance de la vraisemblance des données d'apprentissage avec les itérations. Chacune d'elle est formée de deux étapes :

- une étape E (Estimation) où la fonction vraisemblance des données complètes étant donnés les paramètres des modèles à l'itération précédente est estimée en commençant par les paramètres initiaux x_k, μ_k, Σ_k du modèle initial λ, on estime les nouveaux paramètres : $\bar{x}_t, \bar{\mu}_k, \bar{\Sigma}_k$ telle que la vraisemblance du nouveau modèle soit supérieure ou égale à la vraisemblance du modèle initial.

Dans chaque itération de l'algorithme EM, pour tous les vecteurs acoustiques, $x_t, t = 1, \dots, T$ il faut calculer la probabilité $\gamma_{n,m}$ qui indique dans quelle proportion un vecteur x_t appartient à la gaussienne *k* tel que :

$$\gamma_t = \frac{p(x_t, \mu_k, \Sigma_k)}{\sum_{k=1}^{K} p(x_t, \mu_k, \Sigma_k)}, \tag{3.21}$$

- une étape M (Maximization) où une nouvelle estimation des paramètres du modèle est obtenue en maximisant la fonction de vraisemblance précédente. Les nouveaux paramètres sont définis comme suit :

$$\bar{w}_K = \frac{\sum_{t-1}^{T} \gamma_t(k)}{\sum_{k-1}^{K} \sum_{t-1}^{T} \gamma_t(k)}, \tag{3.22}$$

$$\bar{\mu}_k = \frac{\sum_{t-1}^{T} \gamma_t(k) x_t}{\sum_{t=1}^{T} \gamma_t(k)}, \tag{3.23}$$

$$\overline{\Sigma_k^l} = \frac{\sum_{t-1}^{T_c} \gamma_t^c(K)(x_t - \mu_k)(x_t - \mu_k)'}{\sum_{t-1}^{T} \gamma_t(k)} \tag{3.24}$$

Ce processus est répété plusieurs fois jusqu'à atteindre un seuil de convergence.

La qualité des paramètres estimés de la modélisation dépend de la quantité et de la représentativité des données d'apprentissage.

3.4.2. La phase d'apprentissage

Dans cette phase, on estime les paramètres des gaussiennes qui composent un modèle GMM en se basant sur les vecteurs acoustiques déterminés dans l'étape d'extraction de paramètres. L'apprentissage se fait en deux étapes :

- la première est l'initialisation des paramètres du modèle en utilisant l'algorithme K-moyennes (K-means) ou l'algorithme LBG ;
- la deuxième est l'optimisation des paramètres obtenus dans la première étape en utilisant l'algorithme EM.

3.4.3. Rapport d'hypothèses Bayésien

En vérification du locuteur, le processus de décision est basé sur un test d'hypothèses.

Étant donné un signal de parole "S" et une identité "l_x" revendiquée par l'utilisateur, le système doit décider laquelle des deux hypothèses suivantes est la plus vraisemblable :

- H_0 : le signal "S" a été produit par ;
- H_1 : le signal "S" n'a pas été produit par.

Le rapport de vraisemblance (Likelihood Ratio - LR) entre les deux hypothèses H_0, H_1 et pour l'identité l_x est noté $L \quad (X, H_0, H_1)$ Le test bayésien est la comparaison du rapport de vraisemblance avec un seuil de décision Ω.

$$\mathcal{L} = (S, H_0, H_1) = \frac{p\left(H_0/S\right)}{p\left(H_1/S\right)} \tag{3.25}$$

3.5. Conclusion

Dans ce chapitre, nous avons présenté d'abord les méthodes d'analyse et d'extraction des paramètres puis les différentes approches de modélisation du locuteur par le mélange de gaussiennes (GMM). Après, nous avons présenté la méthode d'estimation de l'ensemble de ces paramètres et enfin, les techniques d'évaluation du système de Reconnaissance Automatique de la Parole (RAP).

CHAPITRE 4 :

Expériences et Résultats

4.1. Introduction

Dans cette dernière partie, nous avons exposé la conception et l'architecture de notre système de classification élaboré. Nous avons appliqué ensuite les mélanges de gaussiennes GMM pour reconnaitre et classifier automatiquement les paroles pathologiques par rapport à la Parole Normale. Les résultats et leurs interprétations sont discutés à la fin du chapitre.

4.2. Description de la base de données

La BD utilisée dans notre travail est constituée par des enregistrements de deux corpus différents (tableau 4.1), prononcés par quatre locuteurs en anglais : deux normaux et deux et deux pathologiques (fente palatine) (Tab 4.2).

Chaque locuteur a prononcé les différents mots des corpus, constituant la base d'apprentissage. Pour la phase test nous avons choisi de traiter des enregistrement des locuteurs, et ceci en utilisant uniquement le corpus numéro 3 où chaque locuteur doit prononcer son prénom durant cette phase.

Les enregistrements ont été effectués sous le format WAV. (Dans des cliniques), avec une fréquence d'échantillonnage 16, 44 KHz.

Tab 4.1 : Corpus en phrase et mot isolés

Corpus I	
Bobby and Bill play ball	[bɑbi ænd bɪl ple bɒl]
Bobby	[bɑbi]
Bill	[bɪl]
Play	[ple]
Ball	[bɒl]
Corpus II	
Bobby is a baby boy	[bɑbi ɪz ə bebi bɔj]
Bobby	[bɑbi]
Baby	[bebi]
Boy	[bɔj]

Tab 4.2 : Locuteurs à tester (2 Normaux et 2 Pathologiques)

Locuteurs	Age	Pathologie	Sexe
Locuteur 1	16 ans	Fente palatine postérieure	H
Locuteur 2	17 ans	Fente palatine postérieure	F
Locuteur 3	23 ans	/	H
Locuteur 4	24 ans	/	F

4.3. Description de l'application développée SOFP

Concernant l'outil MATLAB, nous avons utilisé 7 fonctions réparties en 7 fichiers.m : gaussmix.m, kmeans_gmm.m, Feature_extraction.m, inter.m, TIMITread.m, log_pdf.m, disteusq.m. Et un programme principale qui qui donne le résultat final directement principale.m, il charge la BD et fait appel aux autres fonctions.

- wavread et TIMITread charge les fichiers sonores de la BD ca dépend de type voulu ;
- Feature_extraction.m paramétrise le signal parole dans chaque fichier en calculant des MFCC ;
- calcul de la GMM est contenu dans les fichiers suivants : gaussmix.m, kmeans_gmm.m, log_pdf.m ;
- calcul de la distance euclidienne pour la mesure de ressemblance entre les paramètres du locuteur inconnu et celle de la BD, cette tâche est réalisée par fonction disteu.m.

4.4. Description de l'interface

Pour une meilleure présentation des résultats obtenus, nous avons réalisé une interface (figure 4.4) avec pour principaux composants :

- la barre de titre ;
- le menu : avec deux boutons quitter et commencer ;
- les zones de textes : (1) pour afficher le chemin du répertoire au courant et (2) pour afficher le taux d'apprentissage ;
- les cases pour choisir les corpus et le type de fichier audio.

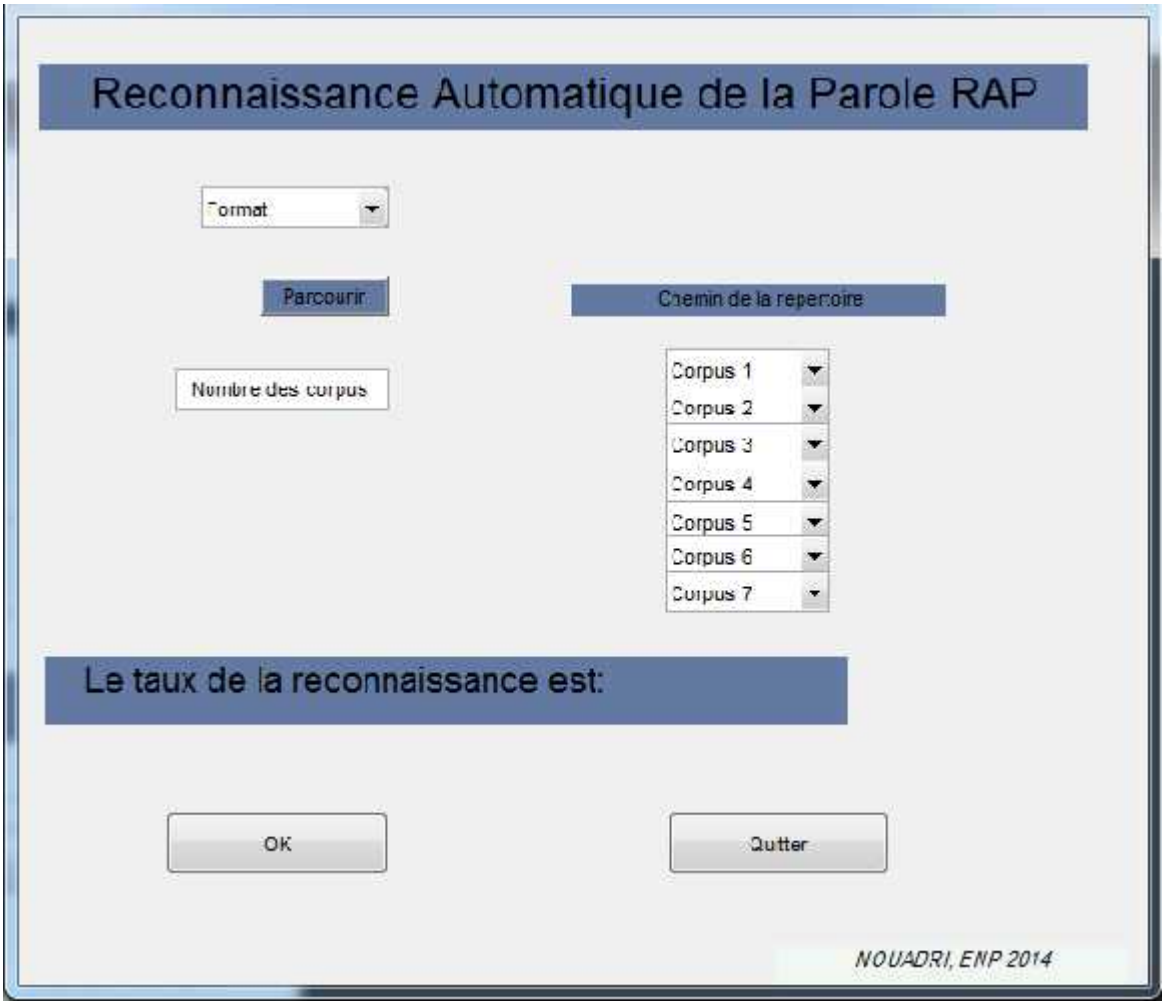


Fig. 4.1 : Interface graphique du système développé SOFP

4.5. Conception du système de classification élaboré

Ces dernières années, la reconnaissance de la parole pathologique PPath a reçu une attention particulière dans les recherches en TAP. Dans les récentes approches de classification automatique des PPath, plusieurs méthodes basées sur la reconnaissance de formes sont utilisées. Ces derniers donnent de bons résultats en reconnaissance de formes et la RAP est une technique traditionnellement connue comme un problème de reconnaissance de formes.

Une des plus importantes phases de notre système est le choix adéquat des paramètres acoustiques à exploiter comme vecteurs d'entrée. Pour mieux discriminer les PPath par rapport aux PNorm, nous avons opté pour les paramètres MFCC. Pour la mise au point des outils du prétraitement des fichiers sons (préaccentuation, détection de parole utile et fenêtrage), ainsi que l'apprentissage et la classification automatique de la PPath par rapport à la PNorm, nous avons utilisé le langage de programmation Matlab 2013.

Comme méthode d'optimisation et de calcul, nous avons utilisé les mélanges de gaussiennes GMM.

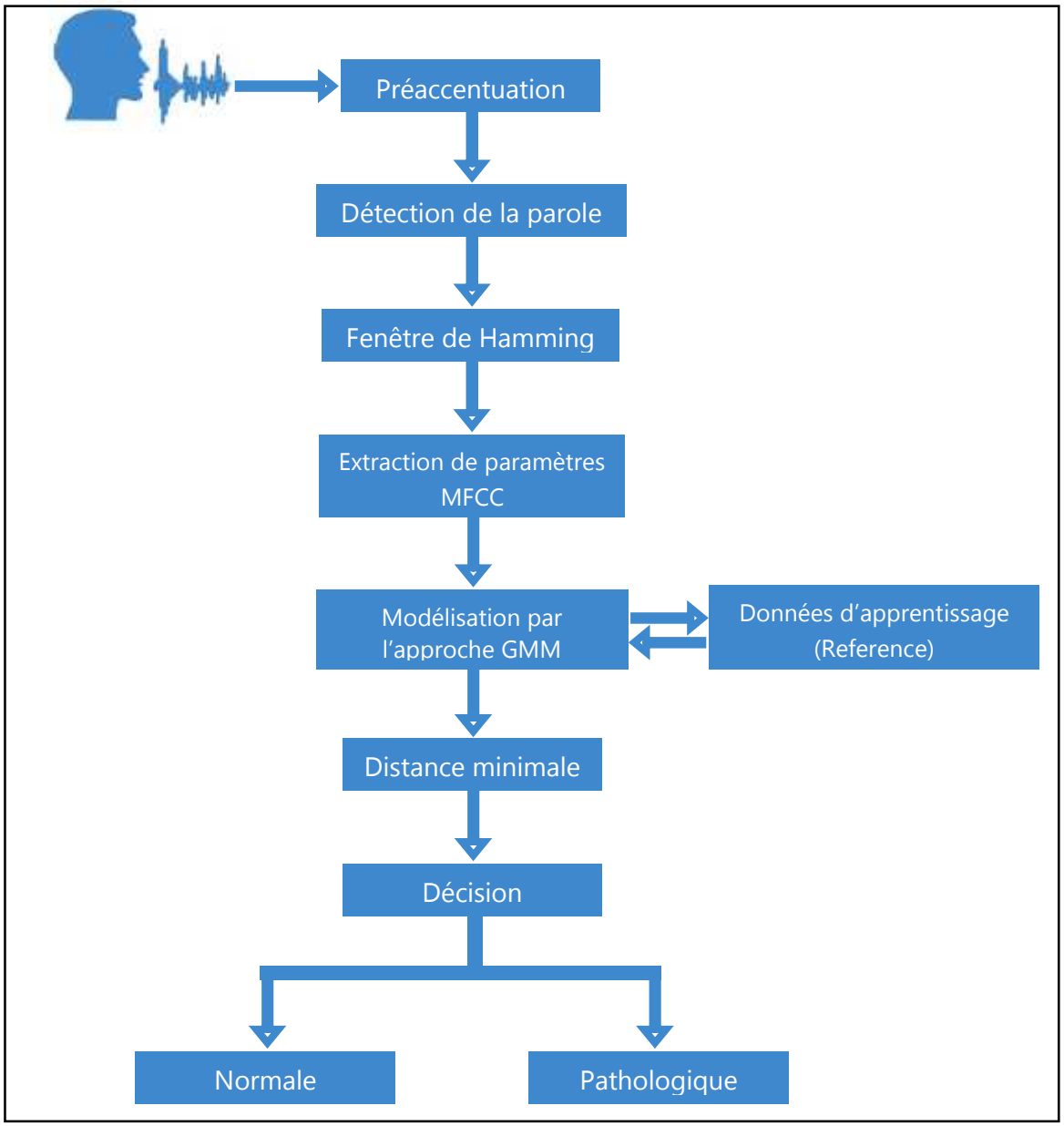


Fig. 5.2 : Organigramme de classification automatique de parole normale/pathologique

4.5.1. Enregistrements du corpus

Pour extraire les fichiers sons, nous avons exploité le même corpus utilisé pour l'analyse acoustique. Ce corpus comprend aussi bien des Paroles Pathologique (fente Palatine), que Normales (PNorm). En plus comme il été mentionné dans le chapitre 2, L'incompétence vélo-pharyngée peut entraîner des troubles de l'articulation, nous on a pris le cas des phonèmes [b] et [p].

Donc, nous avons segmenté manuellement des mots, phrases et parole continue du corpus pour obtenir des fichiers de taille moyenne de 2000 à 4000 ms et en nombre assez suffisant. Nous avons divisé notre corpus des enregistrements sonores en deux groupes : un groupe concerne 14 fichiers sonores à exploiter lors de la phase d'apprentissage (Parole normale PNorm) et un autre également de 14 autres fichiers à exploiter lors de la phase de tests de classification. En effet, une fois la GMM est entrainé, il est nécessaire de tester la fiabilité de notre système sur une autre base de données différente de celle utilisée pour l'apprentissage. Ce test permet d'apprécier les performances du système.

Avant de passer à l'extraction des paramètres acoustiques représentatifs du signal de parole à étudier, nous devrons nécessairement faire subir à ce dernier quelques prétraitements importants qui nous permettent de récupérer le signal de parole "utile".

4.5.2. Traitement du signal de parole

D'abord y a l'étape du prétraitement qui se résume en :

- une préaccentuation dont l'objectif est d'augmenter la quantité d'énergie dans les hautes fréquences et d'avoir une compensation de filtrage des effets de l'acquisition du signal (Figure 5.). Pour cela, le signal de parole enregistré est appliqué à l'entrée d'un filtre de premier ordre FIR (**F**inite **I**mpulse **R**esponse) de la forme :

$$H(z) = 1 - 0.95Z^{-1}$$

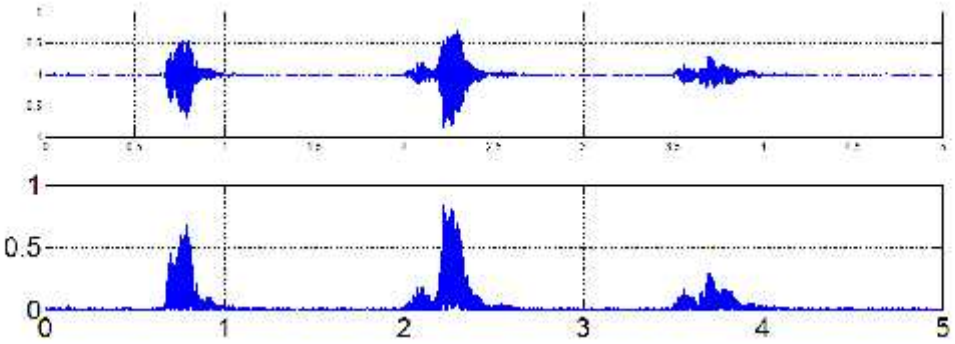


Fig. 4.3 : Signal temporel du mot [ba bo bi] après préaccentuation

- une délimitation des débuts et fins de mots et élimination de toutes les portions du signal enregistré qui ne sont pas de la parole (Figure 5.3). Le défi consiste à éliminer ces échantillons inutiles à partir du signal sans perdre ou fausser l'information pertinente véhiculée par le signal de parole. Une fonction procédure, réalisée sous Matlab 2013, utilise un seuil minimal d'énergie moyenne calculé sur la

base d'enregistrements de différents bruits d'environnement. Dès que l'énergie dépasse un seuil minimal dans une trame du signal (fenêtre de 30 ms), nous considérons que le début de parole commence à partir de cette trame et toutes les autres trames précédentes sont éliminées (figure 4.3). La même procédure est appliquée à la fin du signal de parole.

Il reste néanmoins que cette procédure donne de bons résultats pour le cas de mots isolés, mais reste très limitée dans le cas d'une parole continue car les frontières de mots sont très difficiles à distinguer (du fait des phénomènes de coarticulation), sauf si le locuteur marque explicitement une pose entre chaque mot.

4.5.3 Extraction des MFCC

Comme nous l'avons exposé dans le chapitre 3, les MFCC sont les paramètres pertinents qui différencient nos locuteurs (tableau 4.2), cette tâche a été réalisée, par l'outil HTK (Fig. 4.4).

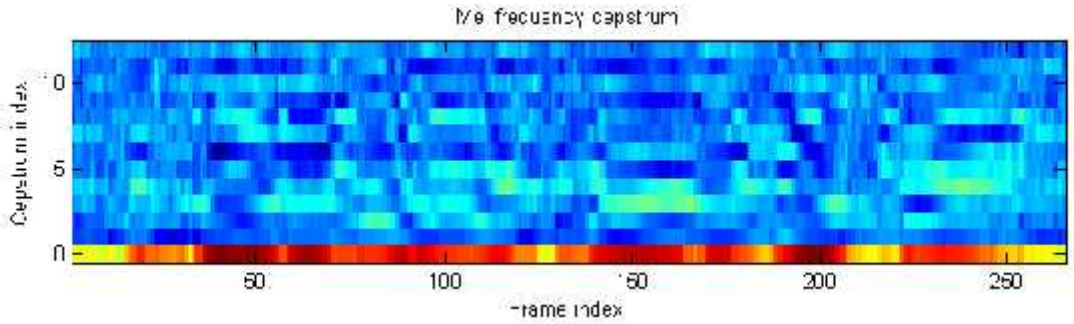


Fig. 4.4 : Représentation spectrale de la MFCC du corpus traité

Tableau 4.3 : Paramètres utilisés pour l'extraction des MFCC avec HTK

paramètres	valeurs
Tw : durée d'analyse par trame	20 (ms)
Ts : décalage par trame	10 (ms)
Alpha : coefficient de préaccentuation	0.97
M : nombre de banc de filtre sur l'échelle de Mel	20
C : nombres des MFCC	12
LF : fréquence minimale	0 (Hz)
HF : fréquence maximale	Fs/2 (Hz)

4.6. Evaluation du système développé

L'évaluation de la qualité d'un système de RAP dépend de plusieurs facteurs. Ce sont les performances en termes de taux d'erreurs qui vont en déterminer la qualité. Cependant, un système d'authentification en phase d'exploitation dépend aussi des échecs d'apprentissage, c'est-à-dire si pour des raisons de défauts matériels, ou parce que l'enregistrement a de trop mauvaise qualité pour servir à l'authentification, le système décide de rejeter le signal et de procéder à une nouvelle phase d'entraînement. Un autre critère consiste à prendre en considération le corpus sur lesquels ces mesures ont été effectuées. En effet, un corpus comportant peu de variabilité ou un trop petit nombre de locuteurs peut guider à une mauvaise interprétation des résultats [29].

4.6.1. Evaluation des performances

Les performances d'un système de RAP s'évaluent en fonction de deux taux d'erreurs. La probabilité de **F**alse **R**ejection **R**ate (FRR) ou de rejet du client à l'identité proclamée et la probabilité de **F**alse **A**cceptante **R**ate (FAR) ou d'acceptations d'impostures. Ces taux sont étroitement liés. Au point de fonctionnement, pour un certain seuil de vérification, ces deux taux sont définis. En fonction du type d'application souhaitée, le seuil de vérification peut être choisi pour minimiser le taux de FAR : application de sécurité, ou minimiser le taux de FRR pour augmenter l'ergonomie d'utilisation.

Il n'est pas possible de minimiser conjointement ces deux taux :

$$P_F = \frac{N_i \quad i_i \quad a \quad és}{n_i \quad d \ c_i \quad t_i \quad és} \tag{4.1}$$

$$P_F = \frac{N_i \quad a \ c_i \quad r \quad és}{n_i \quad d \ c_i \quad t_i \quad és} \tag{4.2}$$

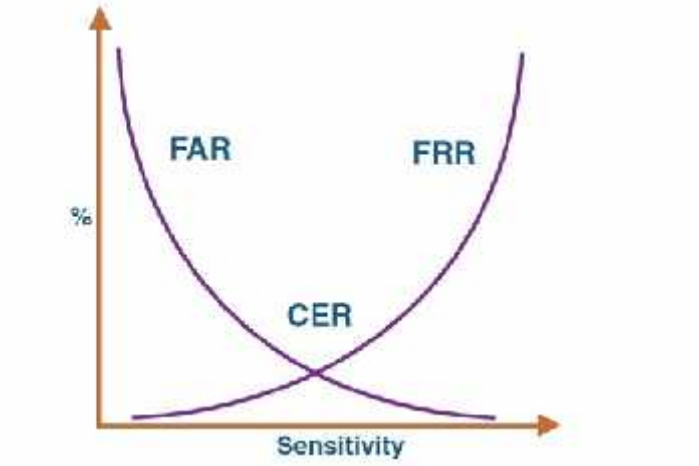


Figure 4.5 : Evaluation des taux FAR et FRR

Le taux d'erreur de décision est dépendant su seuil de décision fixé dans le module de décision et sont en générale en fonction du seuil (τ).

4.6.2 Choix du seuil de décision

La décision d'acceptation ou de rejet d'une séquence test se fait généralement par comparaison du score final à un seuil (PPath). Si l'on dispose d'un corpus de développement étiqueté, on peut fixer ce seuil pour optimiser la mesure de performance considérée pour l'application visée. Sinon, on peut procéder par validation croisée sur les données d'apprentissage. Si l'on dispose de beaucoup de données d'apprentissage pour un locuteur, on peut même envisager une approche individuelle (ou lieu d'une approche globale), c'est-à-dire choisir un seuil de décision spécifique au locuteur.

Lors des essaies effectué, nous avons obtenu un taux de reconnaissance de 92.8571%% pour les gens normaux alors que pour les malades on a obtenu un taux très réduit, 47.3482%.

Donc notre seuil est compris entre 50% et 90%.

4.7. Conclusion

Dans ce chapitre, nous avons présenté d'abord les méthodes d'analyse et d'extraction des paramètres puis les différentes approches de modélisation du locuteur par le mélange de gaussiennes (GMM). Après, nous avons présenté la méthode d'estimation de l'ensemble de ces paramètres et enfin, les techniques d'évaluation du système de Reconnaissance Automatique de la Parole (RAP).

Conclusion et Perspectives

Ce projet avait pour objectif la caractérisation des paroles pathologiques en vue de leur exploitation en réhabilitation de la parole, la conduite de diagnostics automatiques et l'établissement de systèmes experts permettant de caractériser de façon fiable les anomalies vocales.

Dans notre travail, nous avons traité le problème de la RAP pour la Parole Pathologique, cas de la fente palatine. Il s'agit d'extraire les vecteurs acoustiques, à partir des signaux de paroles enregistrés par les locuteurs de la BD, qui servent à la phase d'apprentissage des modèles représentants chaque locuteur. Pour l'analyse paramétrique nous avons utilisé les MFCC avec les GMM comme approche de modélisation.

Les résultats obtenus indiquent un Taux de Reconnaissance (TR) de 90% pour la Parole Normale et 50% pour la parole pathologique. Donc, notre seuil de décision est compris entre 50% et 90% et après plusieurs tests nous avons pris 80% comme seuil de décision pour classifier les paroles. Ce taux est lié au nombre restreint de locuteur dans la BD.

Comme perspectives à ce travail nous proposons :

- implémentation du système afin de le rendre autonome (système embarqué) ;
- utilisation d'une paramétrisation acoustique basée non seulement sur les MFCC, mais aussi sur les coefficients pathologiques comme le degré de perturbation de F_0 (Jitter) et le degré de perturbation de l'intensité (Shimmer) ;
- mise au point d'une BD plus riche à enregistrer en milieu ambiant à l'échelle nationale, car il est important de prendre, des échantillons de locuteurs et paroles pathologiques plus importants pour avoir des résultats plus représentatifs et plus fiables ;
- hybridation des techniques pour améliorer la performance du système, exemple, combinaison de la technique "pitch" avec la GMM, "pitch-GMM".

Références Bibliographiques

[1] R. Benslimane, Transformation de voix en temps réel, Département de Traitement du Signal, Université de la Marne la Vallée, France, juillet 2000.

[2] http://www.ikonet.com/fr/ledictionnairevisuel/etre-humain/anatomie/appareil-respiratoire/appareil-respiratoire.php

[3] C. Jacquier, Étude d'indices acoustiques dans le traitement temporel de la parole chez des adultes normo-lecteurs et des adultes dyslexiques, Thèse de Doctorat Neurosciences et Cognition, Université de Lyon, France, 2008.

[4] L. Buniet, Traitement automatique de la parole en milieu bruité : étude des modèles connexionnistes statiques et dynamiques, Thèse de Doctorat de l'université Henri Poincaré, Nancy 1, France, Février 1997.

[5] Calliope. La parole et son traitement automatique. Collection technique et scientifique des télécommunications, CNET - ENST, Masson, 718 pages, 1989.

[6] O. Godin, Chapitre 5-Analyse de la parole IMN317, Université de Sherbrooke/Canada, 2011.

[7] M. Guerti, Contribution à la synthèse de la parole en Arabe standard, synthèse par diphones et technique de prédiction linéaire. Thèse de Magister, ILP, Alger, Algérie, 1983.

[8] C. Abry, D. & J.V. ABRY, La phonétique : audition, prononciation, correction Clé International, Paris, 2007.

[9] P. Munot, F.Xavier, Nève, Une introduction à la phonétique : manuel à l'intention des linguistes, orthophonistes et logopèdes, Editions du CEFAL, 2002.

[10] K. Ferrat, Classification de la Parole Pathologique par Réseau de Neurones Artificiels. Thèse de Doctorat, ENP, Alger, Algérie, 2014.

[11] http://cfcc.ie-eg.com/formatio/phonetique%20cours.htm

[12] Item 337 : Trouble aigu de la parole. Dysphonie, Collège Français d'ORL, Université Médicale Virtuelle Francophone, 2010.

[13] J. Germain, F. Parent, Les troubles de la voix : du diagnostic au traitement, Le Médecin du Québec, 1994.

[14] http://www.vaincre-le-begaiement.fr/index.php?p=1

[15] D. H. McFarland, F.H. Netter, "L'anatomie en orthophonie, Parole, déglutition et audition", 2ème édition, Masson, 2009.

[16] P. Montoya et H. Baylon, 1996 et Thibault, 2007.

[17] E. Noirrit-Esclassan, P. Pomar, R. Esclassan, B. Terrier, P. Galinier, V. Noisard, Plaques palatines chez le nourisson porteur de fente labiomaxillaire. Encycl. Med. Chir (Elsevier SAS). Stomatologie. 22-066-B-55, 2005.

[18] J.C. Murray, Gene/environment causes of cleft lip and/or palate. Clinical Genetics. April 2002. 61(4): 248-56.

[19] Abdelli-Beruh, The stop voicing contrast in French sentences: contextual sensitivity of vowel duration, closure duration, VOT, stop release and closure voicing. Phonetica, 61 (4), 201-219, 2004.

[20] J. Kreiman and B.R. Gerratt, Perception of aperiodicity in pathological voice, Journal of the Acoustical Society of America, 117, pp.2201-2211, 2005.

[21] M. Bouchamekh, Identification du locuteur indépendante du texte, Mémoire de magister à l'ENP, 2006.

[22] F. Itakura, Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals, J. Acoust. Soc. Am, 57, 535(a), s35 (A), 1975.

[23] JOSEPH P. CAMPBELL, Speaker Recognition: A Tutorial, Proc. IEEE, Vol. 85, NO. 9, September 1997.

[24] Booth, M. Barlow, B. Watson, Enhancement to DTW and VQ decision algorithms for speaker recognition, and I Speech Communication, 7-433 1993.

[25] K. Yu, J. Mason and J. C. Oglesby, Speaker Recognition using Hidden Markov Models, Dynamic time Warping and Vector Quantization. Vision, Image and Signal Processing, 142(5) p 313-316, 1995.

[26] M. Savic, and S. K. Gupta, Variable parameter Speaker verification System based on Hidden Markov Modeling, ICASSP, Volume 1, p281-284, 1990.

[27] F. Bimbot, I. Magrin-Chagnalleau and L. Mathan Second-order statistical measures for text-independent speaker identification Speech Communication, 17: 177-192, 1995.

[28] Ke. Chen, Towards better making a decision in Speaker Verification, Pattern Recognition, N° 36, pp. 329 – 346, 2003.

[29] N. Scheffer, Structuration de l'Espace Acoustique par le Modèle Générique pour la Vérification du Locuteur, Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse, France, 2006.