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

Université Dr Yahia Farès de Médéa
Faculté des Sciences et de la Technologie
Département du Génie Électrique et de l'Informatique



Mémoire

Présenté en vue d'obtention du diplôme de Magister

Spécialité : Électronique

Option : Instrumentation et Microélectronique

Présenté par :
KERRACHE Soumia

Thème

Conception d'une méthode robuste de reconnaissance de la parole pour un système embarqué

Devant le jury :

Pr. Salam ABUDURA	Professeur, U. Médéa	Président
Dr. Noureddine BATEL	MCA, U. Médéa	Examinateur
Pr. Abderrezak GUESSOUM	Professeur, U. Blida	Examinateur
Dr. Mohamed OULD ZMIRLI	MCA, U. Médéa	Encadreur

Résumé

Les systèmes actuels de reconnaissance automatique de la parole RAP permettent d'atteindre des performances relativement élevées. Ces systèmes nécessitent souvent de grande capacité de mémoire et de calcul du fait de la complexité des algorithmes auxquels ils font appel. De nos jours on assiste de plus en plus à l'intégration des systèmes RAP dans divers domaines, ce qui a pour conséquence la nécessite de limité fortement la consommation de ressources matérielles, et ceci dans le but de réduire le coût du produit ou du service dans des marchésde plus en plus compétitifs.

L'objectif de ce mémoire est de concevoir et implémenté sous MATLAB une méthode de reconnaissance de la parole destinée à reconnaitre des chiffres isolés basée sur la programmation dynamique (DTW) pouvant fonctionner aisément sur système embarqué. Les performances de reconnaissance ont été testées en effectuant des expériences sur un corpus de plusieurs locuteurs extraits de la base de données TIDIGIT de Texas Instrument.

ABSTRACT

The Current systems of automatic speech recognition ASR achieve relatively high performance. These systems often require large memory and computation due to the complexity of the algorithms that they use. Today we see more and more integration of ASR systems in various fields, which results in the need to severely limited the consumption of material resources and this in order to reduce the cost of the product or service in markets increasingly competitive.

The objective of this thesis is to design and implemented in MATLAB a method for speech recognition to recognize isolated digits based on dynamic time warping (DTW) can operate easily on embedded system. Recognition performances were tested by conducting experiments on a corpus of several speakers extracted from the data base TIDIGIT Texas Instruments.

ملخص

الأنظمة الحالية للتعرف الأوتوماتيكي على الكلام تحقق أداء عال نسبيا، هذه الأنظمة غالبا ما تتطلب ذاكرة كبيرة و قدرة عالية على الحساب نظرا للتعقيد الذي تعرفه الخوارزميات التي يستخدمونها. اليوم نرى المزيد من الدمج لأنظمة التعرف الأوتوماتيكي على الكلام في مختلف المجالات ، ما يؤدى إلى ضرورة الحد بشدة من استهلاك الموارد المالية ، و هذا من أجل خفض تكلفة المنتج أو الخدمة في أسواق تنافسية على نحو متزايد .

الهدف من هذه الرسالة هو تصميم و تنفيذ في Matlab طريقة للتعرف على الكلام،هذه الطريقة مخصصة للتعرف على أرقام معزولة على أساس البرمجة الديناميكية (DTW)بحيث يمكن أن تعمل بسهولة على نظام مضمن. تمّ اختبار أداء طريقة التعرف هذه بإجراء التجارب على العديد من المتحدثين المستخرجة من قاعدة بياناتTIDIGIT.

Remerciement

Avant tout nous tenons nos remerciements à notre dieu de nos avoir donné la force et le courage.

Aussi , Nous tenons à remercier vivement Dr. OULD ZMIRLI Mohamed mon promoteur qui a fourni des efforts énormes, par ses informations ses conseils et ses encouragements.

Nous nos vifs remerciement à messieurs les membres de jury pour l'honneur qu'ils nous ont fait en acceptant de siéger à notre soutenance, tout particulièrement : Pr.ABUDURA Salam pour nous avoir fait l'honneur de présider le jury de ce mémoire.

Nos vifs remerciements aussi à : Pr .GUESSOUM AbderrezakDr.BATEL Noureddine, et Tous deux maîtres assistants, pour avoir accepté d'être examinateurs de ce travail.

Et tous les professeurs de département de Génie électrique Et de l'informatique

A tous ce qui furent à un moment ou à tout instant, partie prenante de ce travail.

Nos plus chaleureux remerciements pour tous ceux qui de prés ou de loin ont contribué à la réalisation de ce mémoire.

Dédicace

Je dédie ce mémoire à . . . ✍

A ma très chère mère *:*

Affable, honorable, aimable : Tu représentes pour moi le symbole de la bonté par excellence, la source de tendresse et L'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi.

Ta prière et ta bénédiction m'ont été d'un grand secours pour mener à bien mes études.

A mon très cher papa :

Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour vous. Rien au monde ne vaut les efforts fournis jour et nuit pour mon éducation et mon bien être.

Ce travail est le fruit de tes sacrifices que tu as consentis pour mon éducation et ma formation.

A mon très cher mari :

Tes sacrifices, ton soutien moral et matériel, ta gentillesse sans égal, ton profond attachement m'ont permis de réussir mes études.
Sans ton aide, tes conseils et tes encouragements ce travail n'aurait vu le jour.

Que dieu réunisse nos chemins pour un long commun serein et que ce travail soit témoignage de ma reconnaissance et de ma sincère fidélité.

A mes très chers frères :

Les mots ne suffisent guère pour exprimer l'attachement, l'amour et l'affection que je porte pour vous. Mes anges gardien et mes fidèles compagnons dans les moments les plus délicats de cette vie mystérieuse.

Je vous dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite.

A mes très chères sœurs :

Malgré la distance, vous êtes toujours dans mon cœur. Je vous remercie pour votre hospitalité sans égal et votre affection si sincère.

Je vous dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite

A mes chers beaux-parents :

Vous m'avez accueilli à bras ouverts dans votre famille. En témoignage de l'attachement, de l'amour et de l'affection que je porte pour vous.

Je vous dédie ce travail avec tous mes vœux de bonheur, de santé et de réussite.

A mes beaux-frères :

Que dieu vous assistes.

A mes belles sœurs :

Je t'exprime à travers ce travail mes sentiments de fraternité etd'amour.

A tous les membres de ma famille, petits et grands

Veuillez trouver dans ce modeste travail l'expression de mon affection

Liste d’abréviation

AR : Auto Regressise.
DAV : Détection d’Activité Vocale.
DTW : Dynamique Time Working.
DCT : Discret Cosine Transform.
HMM : Hidden Markov Model.
FFT : Fast Fourier Transformer.
TFD : Transformer de Fourier Discrète.
LPC : Linear Predictive Coding.
LPCC : Linear Prediction Cepstral Coefficients.
MFCC : Mel Frequency Cepstral Coefficients.
RAP : Reconnaissance Automatique de la Parole.
QV : Quantification Vectoriel.
ZC : Zero Crossing.

Sommaire

Introduction générale 1
Chapitre I : LE PROCESSUS DE PRODUCTION ET DE PERCEPTION DE LA PAROLE CHEZ LES HUMAINS 3
I.1 Introduction 3
I.2 Production de la parole 3
I.2.1 Description de l'appareil phonatoire 3
I.2.2 Les sons en parole 6
I.2.2.1 Définitions : le phonème 6
I.2.2.2 Classifications des phonèmes de la langue française 6
I.2.3 Représentation schématique de production de la parole 8
I.2.4 Les types de sons produits 9
I.2.4.1 Sons voisés 9
I.2.4.2 Sons non voisés 9
I.3. Audition et perception de la parole 10
I.3.1 Physiologie du système auditif 10
I.3.1.1 L'oreille externe 11
I.3.1.2 L'oreille moyenne 11
I.3.1.3 L'oreille interne 12
I.3.2 Acoustique de l'audition 13
I.3.2.1 Les sensations auditives 13
I.3.2.2 Masquage et bande critique 14
I.3.2.2.a Le Masquage 14
I.3.2.2.b Les Bandes Critiques 15
I.4 Conclusion 15
Chapitre II : TRAITEMENT NUMERIQUE DU SIGNAL VOCAL 16
II.1 Introduction 16
II.2 la chaine de traitement du signal vocal 16
II.2.1 Prétraitement du signal vocal 17
II.2.1.1 La pré-accentuation 18
II.2.1.2 Le fenêtrage 18
II.3 Les outils du traitement du signal 20
II.3.1 Énergie à court-terme 20
II.3.2 le taux de passage par zéro pour un signal vocal 21

II.3.3. Etude spectrale 22
II.3.4 l'autocorrélation à court terme 23
II.4 La paramétrisation du signal 23
II.4.1 La méthode par prédiction linéaire 23
II.4.1.1 Modèle auto régressif 24
II.4.1.2 L'estimation des coefficients de prédiction 25
II.4.2 L'analyse homomorphique 27
II.4.2.1 Les coefficients cepstraux 28
II.4.2.2. Les coefficients MFCCs 29
II.4.2.3 Les coefficients LPCCs 30
II.5 Conclusion 33
Chapitre III : LA RECONNAISSANCE VOCALE 34
III.1 Introduction 34
III.1.1 Niveaux de complexité 34
III.1.2 Classification des méthodes de reconnaissance vocale 35
III.2 Reconnaissance de type acoustico-phonétique 36
III.2.1 extraction des paramètres pertinents 36
III.2.2 Segmentation et étiquetage des sons de parole 37
III.2.3 Décision et stratégie de contrôle 37
III.3 les distances dans l'espace acoustique 37
III.3.1La mesure de distorsion 37
III.3.2 La distance Euclidienne 38
III.3.3 La distance d'Itakura 38
III.3.4 La distance cepstrale 39
III.3.5 La distance cepstrale pondérée 39
III.3.5.1 La pondération par la fréquence k 39
III.3.5.2 La pondération par variance 40
III.3.5.3 La pondération par filtrage passe-bande (bandpass liftering) 40
III.4 Reconnaissance basée sur l'alignement temporel des formes 41
III.4.1 Présentation du problème 41
III.4.2 La programmation dynamique 41
III.4.3 Les modèles MMCs 44
III.4.3.1 concept général 44
III.4.3.2 présentation des modèles de Markov cachés 45

III.4.3.3 Apprentissage des modèles 47
III.4.4 La quantification vectorielle 47
III.4.5 Algorithme de classification 48
III.5 Conclusion 51
Chapitre IV : La réalisation du système de reconnaissance 53
IV. Introduction 53
IV.2 Description du système de reconnaissance 53
IV.2.1 La détection d'activité vocale 54
IV.2.1.1 introduction 54
IV.2.1.2 Description de l'algorithme utilisé 55
IV.2.1.3 L'utilité de la VAD 57
IV.2.2 Extraction des paramètres 58
IV.2.3 Le dictionnaire de référence (phase de l'apprentissage) 59
IV.2.4 Reconnaissance des chiffres 59
IV.3 Corpus utilisé 59
IV.4. Contexte expérimental et résultats de simulations 59
IV.4.1 Cas d'un système multi-locuteurs dépendants 59
IV.4.2 Cas d'un système multi-locuteurs indépendants 61
IV.4.3 Cas d'un système multi-locuteurs indépendants bruité 63
IV.4.4 Cas d'un système multi-locuteurs indépendants bruité avec plusieurs techniques de Paramétrisation spectrale 70
IV.5 Conclusion 74
Bibliographie

Introduction générale

La reconnaissance de la parole est un domaine en pleine expansion depuis plusieurs années, grâce au développement de systèmes électroniques de plus en plus élaborés, offrant des gains significatifs de puissance et d'espace mémoire, et permettant l'utilisation, pour les applications destinées au grand public, de méthodes jusque là principalement cantonnées aux séances d'essais en laboratoires. Néanmoins, ces avancées technologiques ont surtout permis aux outils tels que les logiciels de dictée vocale de se démocratiser, tandis que de nombreuses autres applications sont embarquées sur des systèmes aux ressources encore trop limitées pour permettre une reconnaissance de qualité dans des conditions adaptées à leur utilisation.

Parmi ces systèmes embarqués, nous pouvons citer les processeurs de traitement du signal présents notamment dans les téléphones mobiles. Apparue au milieu des années 90.Ces processeurs sont parfaitement adaptés aux systèmes mobiles en raison de leur faible consommation et de leur propension à effectuer avec une grande efficacité les opérations de multiplication-addition, très présentes dans le traitement de la parole.

La reconnaissance de mots, prononcés isolément ou de façon enchaînée, constitue la quasi-totalité des applications existantes, dans la mesure où il s'agit d'une simplification du problème général de la reconnaissance de la parole naturelle prononcée continûment. La plupart des systèmes actuels sont monolocuteurs et nécessitent un apprentissage préalable pour tout nouveau locuteur. Néanmoins, de bonnes performances sont maintenant possibles en multi-locuteurs et en adaptation automatique au locuteur. Les systèmes de reconnaissance de mots sont pour l'essentiel fondés sur des modèles formels qui ont été progressivement optimisés: programmation dynamique qui utilise par ailleurs un ensemble de connaissances phonétiques, modèles stochastiques de Markov, réseaux neuronaux, plus récents mais prometteurs.

L'objectif de ce mémoire est de concevoir une méthode de reconnaissance de la parole destinée à reconnaître des chiffres isolés basée sur la programmation dynamique(DTW) pouvant fonctionner aisément sur un système embarqué, c'est-à-dire répondant aux contraintes d'espace mémoire et de temps d'exécution limités. L'intérêt ici sera de proposer un algorithme adapté au vocabulaire considéré, à savoir les chiffres allant de 0 à 9, prononcés en anglais, afin d'optimiser au maximum son fonctionnement selon les caractéristiques du système, tout en obtenant un taux de réussite convenable, pour un grand nombre de locuteurs différents.

Le système de reconnaissance vocale proposé est composé principalement de trois module celui de la phase de paramétrisation, celui de l'apprentissage et en dernier celui de la phase de reconnaissance ou de décision. Dans la phase de paramétrisation, chaque trame de parole est convertie en un ensemble de paramètres LPCs /LPCCs/ MFCCs. Ce type de paramètres permet avec une pondération adéquate, une parfaite représentation des traits acoustiques du signal vocale, ce qui est d'une très grande utilité pour la reconnaissance. La seconde phase celle de l'apprentissage a pour rôle de fournir le dictionnaire des mots de référence à partir des chiffres isolés. Enfin la phase de reconnaissance.

Ce mémoire est organisé de la manière suivante, dans le premier chapitre nous avons exposé les notions théoriques de la production et de la perception de la parole, le deuxième chapitre traite les outils de base pour le traitement et la paramétrisation spectrale du signal vocale, le troisième chapitre est un aperçu sur les principales méthodes de la reconnaissance de la parole, et quelque algorithmes de classifications y sont aussi présentés, dans le quatrième chapitre nous décrirons notre système ainsi que les résultats expérimentaux obtenus .

Chapitre I

LE PROCESSUS DE PRODUCTION ET DE PERCEPTION DE LA PAROLE CHEZ LES HUMAINS

I.1 Introduction

La parole est la faculté de communiquer la pensée par un système de sons articulés émis par les organes de phonation ; c'est le moyen de communication privilégié entre les humains qui sont les seuls êtres vivants à utiliser un tel système structuré. Toutefois l'avantage principal du message oral qui réside dans l'instantanéité de son émission et de sa transmission, a pour contrepartie son immatérialité, laquelle a longtemps constitué un obstacle rédhibitoire pour pénétrer les arcanes de la parole. Ne laissant derrière elle aucune trace de son éphémère passage dans le temps et dans l'espace, la parole a été longtemps synonyme d'abstraction et de virtualité.

I.2 Production de la parole

Production de la parole est une action volontaire et coordonnée d'un certain nombre de muscles du système articulatoire. En guise d'introduction, nous sommes consacrés à l'appareil que l'être humain utilise pour produire des sons durant la phonation.

I.2.1 Description de l'appareil phonatoire

Voici tout d'abord en coupe, l'appareil phonatoire humain.

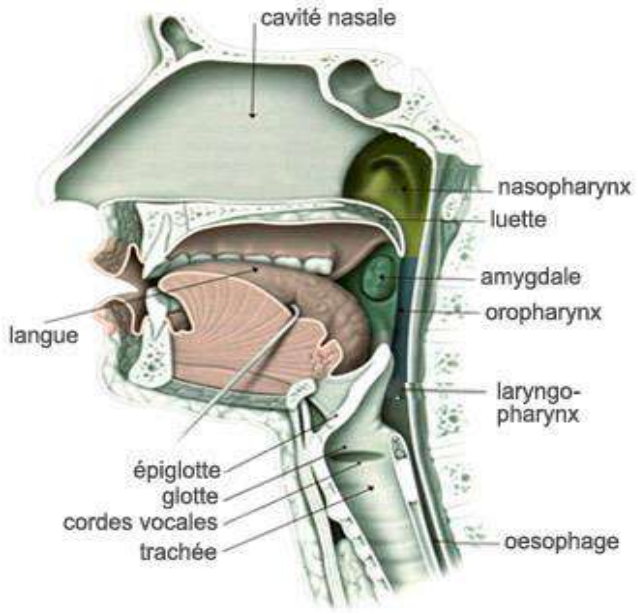


Figure I.1. L'appareil phonatoire [1].

L'appareil phonatoire est composé principalement de trois éléments qui contribuent l'ensemble à la production de la parole. Ces éléments dont le contrôle et la coordination sont assurés par le système nerveux central, sont :

- L'ensemble poumons+bronches+trachée (l'appareil respiratoire)

Les poumons permettent en premier lieu d'alimenter les tissus du corps en oxygène par une alternance de cycles d'inspiration et d'expiration. La phonation se produit dans la très grande majorité des cas au cours de l'expiration. Le flux d'air expulsé des poumons se déplace jusqu'à arriver au niveau du larynx où plusieurs cas se présentent.

- Le larynx

Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée. La figure I.2 nous présente une description du larynx, vu sous deux coupes différentes.

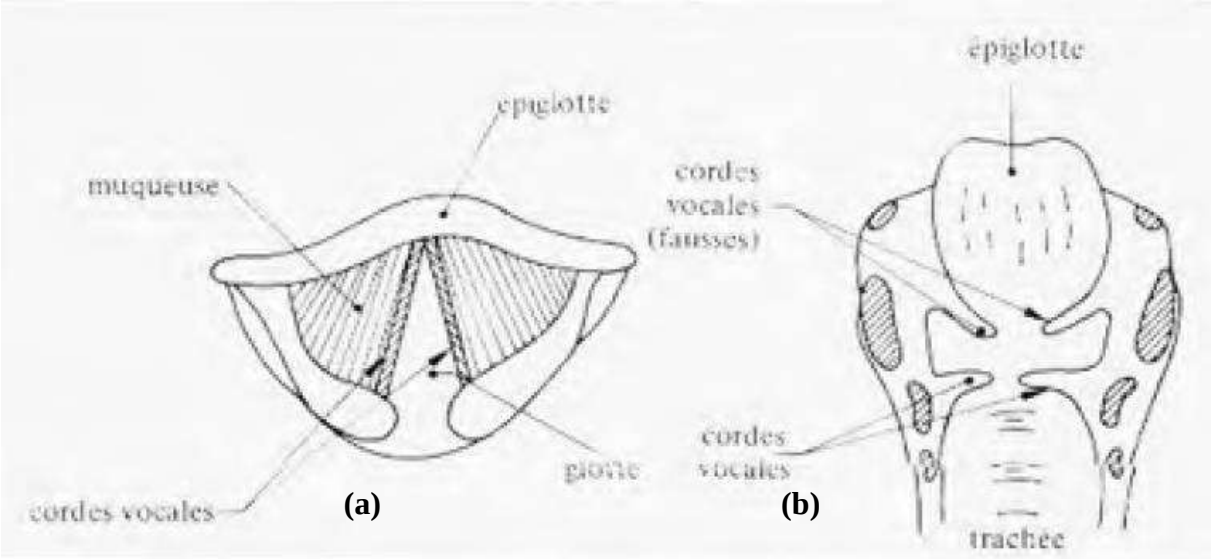
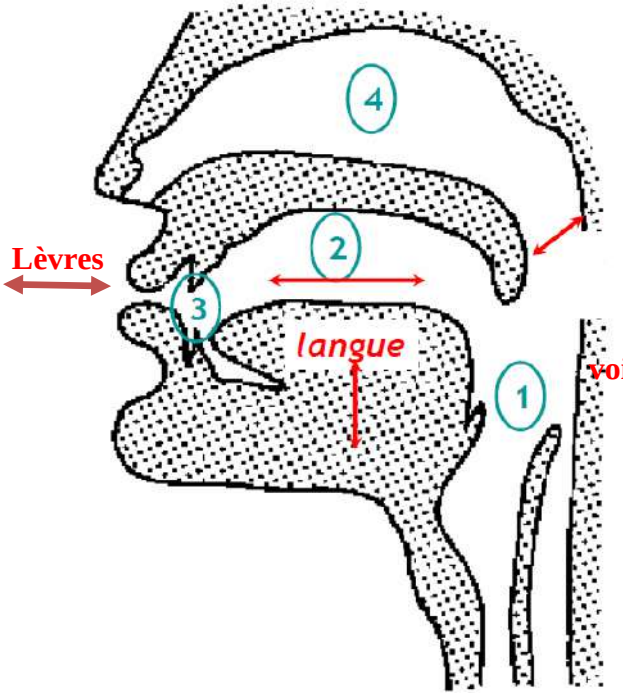


Figure I.2. Vues de larynx : (a) vue de haut ; (b) coupe verticale [2]

Les cordes vocales sont en fait deux lèvres symétriques placées en travers du larynx ; ces lèvres peuvent fermer complètement le larynx et, en s'écartant, déterminer une ouverture triangulaire appelée glotte. [2]

- Le conduit vocal

Le conduit vocal est le terme qui désigne la cavité remplie d'air qui s'étant de la glotte (larynx) aux lèvres. Il peut être considéré comme une succession de tubes ou cavités acoustiques de sections diverses.



1. pharynx
2. bouche
3. cavité labiale
4. Fosses nasales

Figure I.3. Les résonateurs [3]

Le conduit vocal est composé du pharynx, de la cavité orale et de la cavité nasale. Toutes ces Cavités jouent un rôle très important dans la forme finale du son produit. Elles sont en quelque sorte une caisse de résonance de l'appareil de production de la parole. Le pharynx, débutant au croisement de la trachée et de l'œsophage et s'étendant jusqu'au voile du palais, se contente juste de faire circuler le son en provenance de la glotte, mais sa forme peut être modifiée en bougeant d'autres éléments du conduit vocal. La cavité orale est constituée d'éléments extrêmement mobiles tels que la langue, la mâchoire, le palais et les lèvres qui, en modifiant la forme de la caisse de résonance, modifient les sons qui en sortent. La cavité nasale est immobile, mais peut être connectée ou déconnectée du reste du conduit vocal en modifiant la position du voile du palais.[4]

Maintenant que nous connaissons les principaux acteurs du processus de production de la parole, nous allons décrire brièvement leur fonctionnement ainsi que la façon dont ils influent sur la nature du son crée.

Lors de la production, un flot d'air émanant des poumons est contraint à passer dans le larynx, qui par le biais des cordes vocales va générer une vibration, les sons étant des ondes (donc des vibrations). La pression est en quelque sorte transformée en une suite de pulsation quasi-périodique ou aléatoire.

Les sons de parole sont produits soit par les vibrations des cordes vocales, dans ce cas on parle de **sons voisés**, soit par l'écoulement turbulent de l'air dans le conduit vocal, soit lors de relâchement d'une occlusion de ce conduit, alors on parle de son non- voisés.

L'onde ainsi obtenue est ensuite modulée en fréquence en passant au travers du conduit vocal.

Sous l'effet résonateur de différentes cavités, et en fonction des positions des organes tels que les dents, la langue, les lèvres, la bouche ou le voile du palais, différents types de sons peuvent alors être produits pour un même type de pulsation. Une très petite variation peut entrainer de grosses modifications dans la signification des mots.

I.2.2 Les sons en parole

Dans le processus de communication parlée, pour une langue donnée, les sons permettent de distinguer les différentes unités de signification du langage [5]. L'unité élémentaire d'un son permettant la distinction des différents mots est le phonème.

I.2.2.1 Définitions : le phonème [2]

Est la plus petite unité présente dans la parole et susceptible par sa présence de changer la signification d'un mot. Par exemple riz et rat, pari et mari,…

Le nombre de phonème est toujours très limité, normalement inférieur à 50.

Un système de phonème est un ensemble d'images acoustiques emmagasinées dans le cerveau du locuteur dans la mesure où celui-ci maitrise la langue.

I.2.2.2 Classifications des phonèmes de la langue française :

On considère généralement que la langue française comprend 36 phonèmes répartis en classes et en sous-classes ; chaque classe est liée à un mode articulatoire de l'appareil des phonations [6].

La figure I.4 illustre la classification des phonèmes pour la langue française.

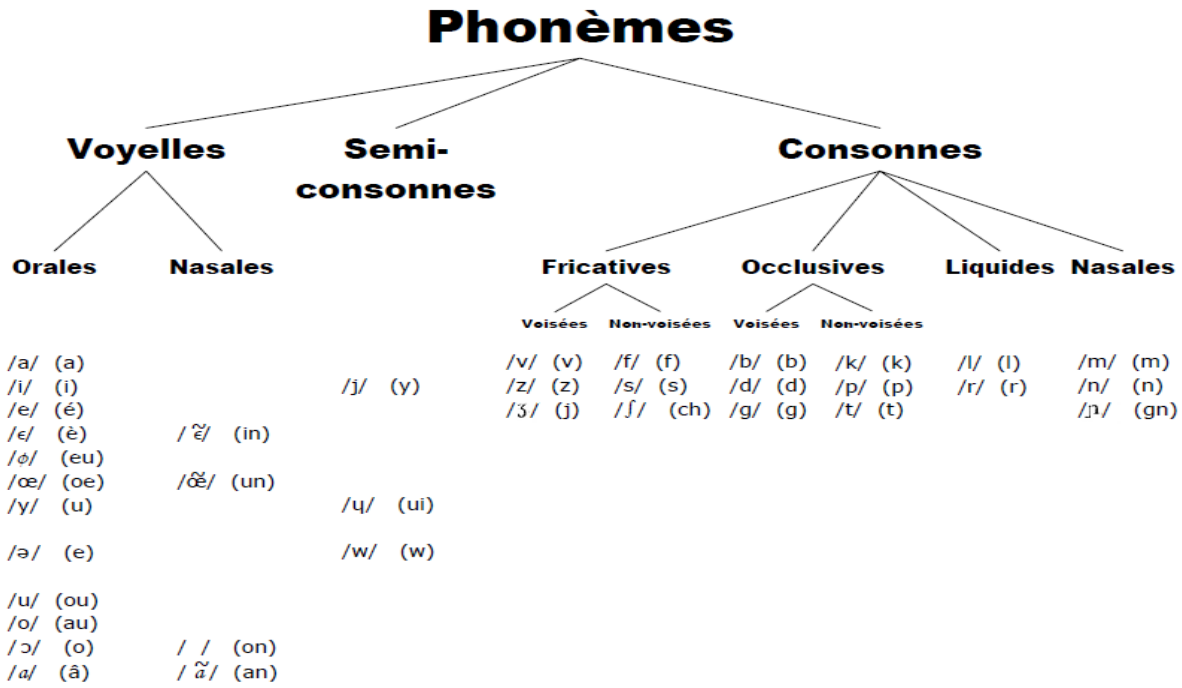


Figure I.4. Phonèmes de la langue française.[7]

Les phonèmes peuvent êtres rangés en catégories selon des "traits distinctifs" qui indiquent une similitude au niveau articulatoire, acoustique ou perceptif. Les voyelles peuvent être rangées selon [6] :

- La nasalité ;
- L'ouverture du conduit vocal ;
- La position de la constriction du conduit vocal ;
- L'arrondissement des lèvres.

On classe principalement les consonnes en fonction de leur :

- Mode d'articulation (occlusif, nasal, fricatif) ;
- Nasalisation.
- Le lieu d'articulation (labiale, dentale, palatale).

Les traits acoustiques du signal de parole sont évidement liés à sa production. L'intensité du son est liée à la pression de l'air en amont du larynx. Sa fréquence, qui n'est rien d'autre que la fréquence du cycle d'ouverture/fermeture des cordes vocales, est déterminée par la tension de muscles qui les contrôlent.[8]

I.2.3 Représentation schématique de production de la parole

De ce que nous avons vu précédemment, nous pouvons dire que chaque son de parole est fruit des actions successives des poumons, du larynx et du conduit vocal, tous agissant indépendamment les uns des autres pour permettre à l'être humain de pouvoir produire une très grand quantité de sons différents. Nous pouvons ainsi résumer le cheminement du son, depuis son origine jusqu'à sa sortie du corps humain dans la figure I.5

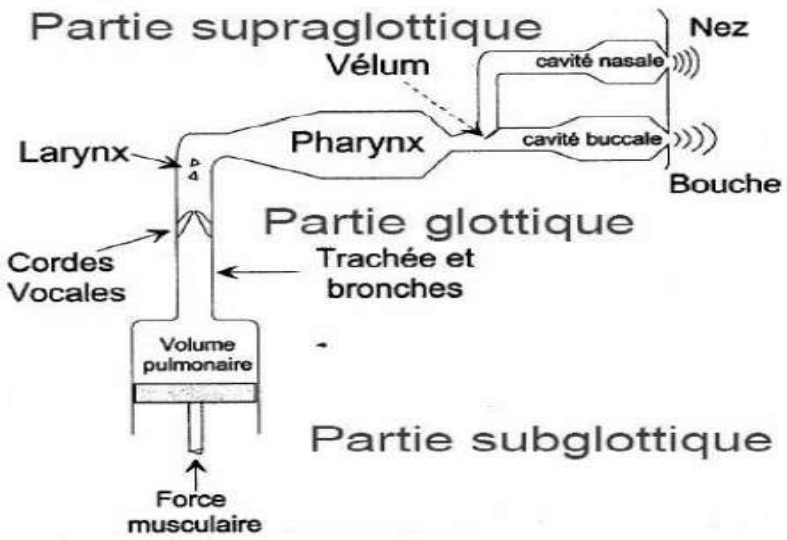


Figure I.5. L'appareil phonatoire humain schématisé.

Une représentation fonctionnelle du modèle de production, séparant sources, conduit et rayonnement aux lèvres, est donnée par la figure I.6.

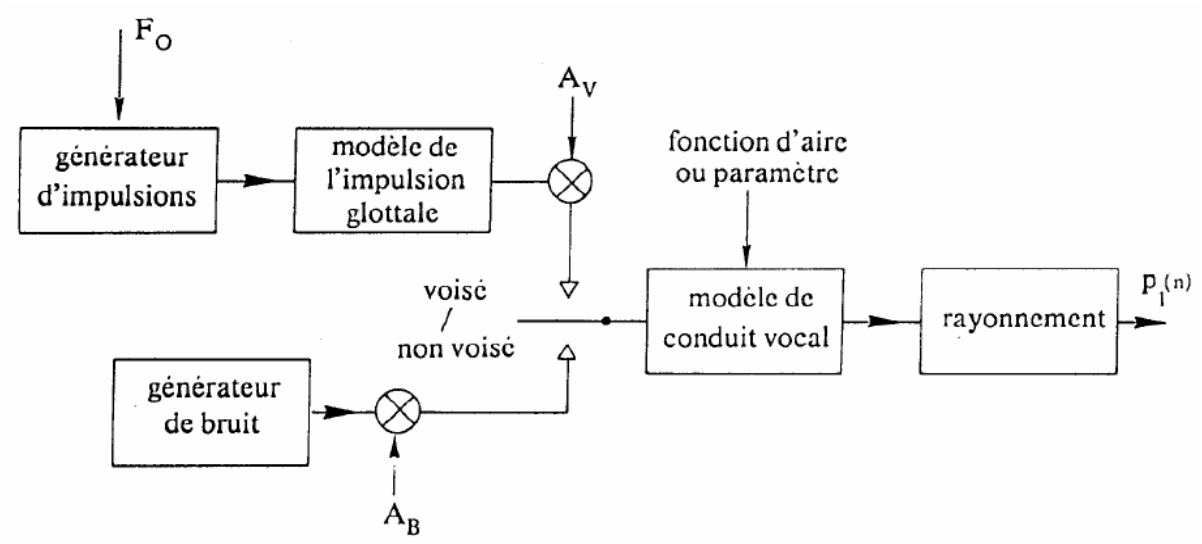


Figure I.6. Modèle fonctionnel de production de la parole.

I.2.4 Les types de sons produits

La parole est une succession d'évènements sonores faisant alternativement apparaître deux types de sons, selon que les cordes vocales vibrent (on dit alors que le son est **voisé**) ou pas (le son est alors **non voisé**).

I.2.4.1 Sons voisés

Dans le premier cas, les vibrations des cordes vocales modulent le flux d'air issu des poumons. Cela résulte en une variation de la pression de l'air passant la glotte. Les oscillations des cordes vocales étant sinusoïdales, l'onde sonore produite au dessus de la glotte est également périodique sinusoïdale, en première approximation. Cette onde se propage alors dans le conduit vocal jusqu'aux lèvres. La fréquence fondamentale du signal, également appelée pitch, est alors déterminée par la fréquence de vibration des cordes vocales, alors que son intensité est liée à la pression de l'air en amont du larynx.

La fréquence du fondamental peut varier : [8]

- de 80 à 200 Hz pour une voix masculine,
- de 150 à 450 Hz pour une voix féminine,
- de 200 à 600 Hz pour une voix d'enfant.

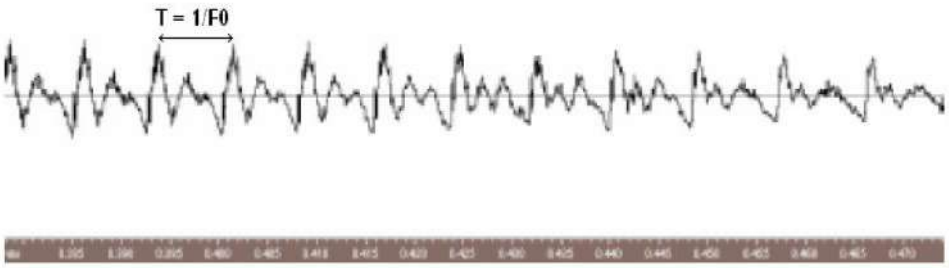


Figure I.7. Exemple de son voisé.

I.2.4.2 Sons non voisés

Lorsqu'on produit des sons non voisés, les cordes vocales ne vibrent pas. Le son est alors produit par le frottement de l'air dans le conduit vocal entre la glotte et les lèvres. Si l'air ne rencontre aucun obstacle sur son chemin, le son produit est très faible (son de la respiration). Par contre, si l'air rencontre une **constriction**, c'est-à-dire un rétrécissement du conduit vocal, sur son passage (par exemple entre les dents et les lèvres pour un 'f'), le son produit est un bruit (son apériodique continu) plus fort. Cela est dû à des propriétés mécaniques de

l'écoulement de l'air à travers un tube : dans certaines conditions, le flux d'air qui s'écoule librement et régulièrement avant la constriction (on dit qu'il s'agit d'un écoulement **laminaire)** est transformé en écoulement turbulent plus du tout régulier au passage d'une constriction. C'est cet écoulement **turbulent**, généré par le frottement de l'air dans la constriction qui provoque le bruit des sons non voisés [9].

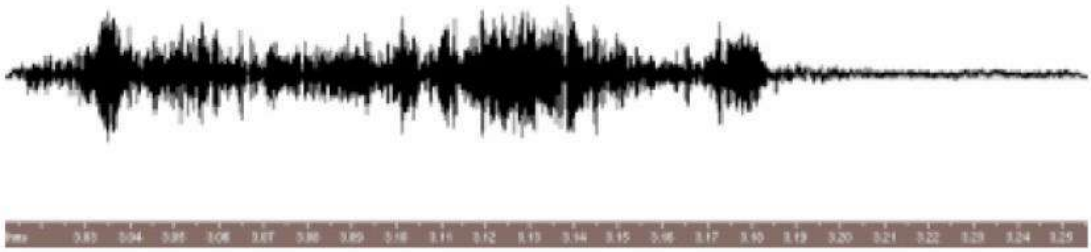


Figure I.8. Exemple de son non voisé.

I.3. Audition et perception de la parole

Le processus de compréhension par l'auditeur d'un message oral peut être décomposé en deux niveaux [5] :

- Une transformation de l'information contenue dans le signal acoustique par l'oreille, qui transmet le résultat au cerveau.
- La reconstitution dans le cerveau du message linguistique, séparé en une étape d'interprétation des indices fournis par l'oreille et une étape d'accès à la sémantique du message.

I.3.1 Physiologie du système auditif

Les ondes sonores sont recueillies par l'appareil auditif, ce qui provoque les sensations auditives. Ces ondes de pression sont analysées dans l'oreille interne qui envoie au cerveau l'influx nerveux qui en résulte; le phénomène physique induit ainsi un phénomène psychique grâce à un mécanisme physiologique complexe.

L'appareil auditif comprend l'oreille externe, l'oreille moyenne, et l'oreille interne, comme présenté sur la figure I.9 [10].

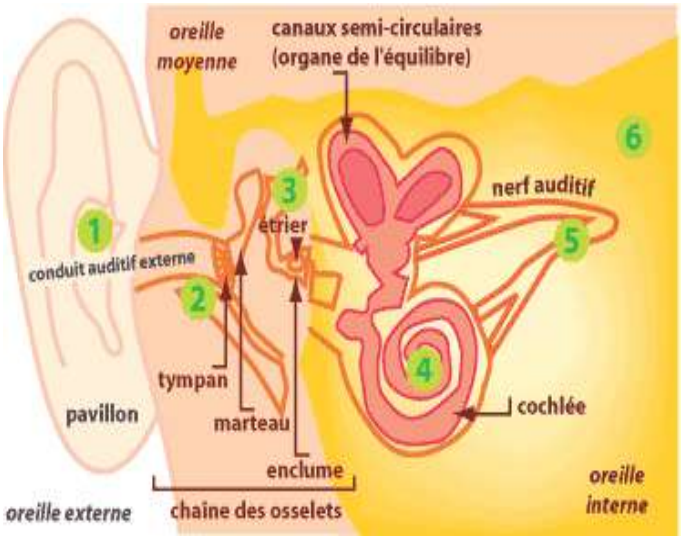


Figure I.9. Appareil auditif humain.

Où :

1. Le son arrive dans le conduit vocal.
2. Le son fait vibrer le tympan.
3. Le marteau et l'enclume transmettent les vibrations.
4. L'oreille interne transforme le son et transmet ces signaux au nerf auditif.
5. Le nerf auditif véhicule l'influx nerveux vers le cerveau.
6. Le cerveau interprète les sons.

I.3.1.1 L'oreille externe

L'oreille externe est constituée du pavillon, qui dirige le flux sonore vers le conduit auditif externe, un tube acoustique de section uniforme fermé à une extrémité (tympan), son premier mode de résonnance est situé vers 3000 Hz, ce qui accroit la sensibilité du système auditif dans cette gamme de fréquence.

I.3.1.2 L'oreille moyenne

L'oreille moyenne est une cavité appelée caisse du tympan, son mécanisme permet une adaptation d'impédance entre l'air et le milieu liquide de l'oreille interne, ainsi une protection contre les sons très élevés qui risquent de l'endommager.

L'oreille moyenne comprend : [6]

- Le tympan, qui est une membrane sensible aux variations de pression acoustique.

L'équilibrage de la pression des deux cotés du tympan est assuré par la trompe d'Eustache, reliée à l'appareil respiratoire.

- Les osselets : le marteau, l'enclume et l'étrier. Ils sont reliés à la fois au tympan (pour le marteau) et à la fenêtre ovale (pour l'étrier). Ils jouent le rôle de levier, amplifiant les vibrations reçues par le tympan. Les muscles du marteau et de l'étrier peuvent modifier les caractéristiques de la chaîne, modifiant ainsi la perception des sons.

I.3.1.3 L'oreille interne

L'oreille interne est un milieu liquide qui comprend :

- Les trois canaux semi-circulaires qui sont le départ du nerf vestibulaire.
- La cochlée, ou limaçon, qui est un canal membraneux enroulé, elle contient la membrane basilaire qui transforme les vibrations mécaniques en impulsions nerveuses. La membrane s'élargit et s'épaissit au fur et à mesure que l'on se rapproche de l'apex de la cochlée ; elle est le support de **l'organe de corti** qui est constitué par environ 25000 **cellules ciliées** raccordées au nerf auditif.la réponse en fréquence du conduit à l'endroit d'une cellule est esquissée à la figure I.10 [11]. La fréquence de résonance $f_{0,i}$ dépend de la position occupée par la cellule sur la membrane.

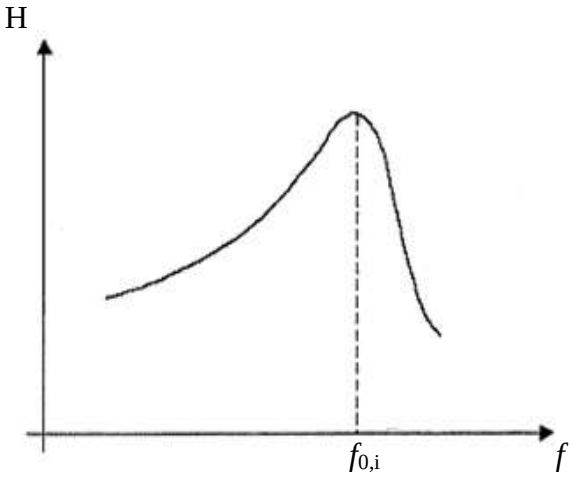


Figure I.10. Réponse en fréquence d'une cellule ciliée.

I.3.2 Acoustique de l'audition

Le système auditif humain est surtout sensible dans une gamme de fréquences situées entre 800Hz à 8000 Hz ; les limites extrêmes sont respectivement 20 et 20 000 Hz [2].la figure I.11 présente le champ auditif humain, délimité par la courbe du seuil de l'audition et celle du seuil de la douleur.[14]

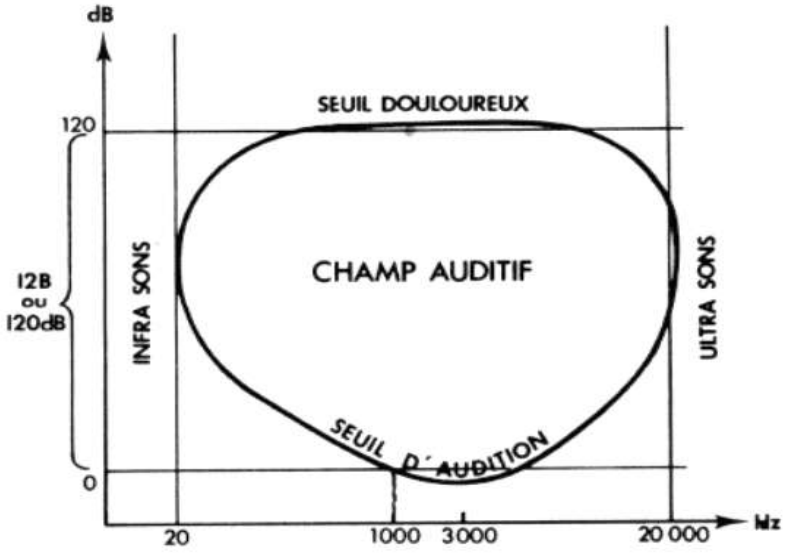


Figure I.11. Limites du champ auditif humain.

Il reste très difficile de nos jours de dire comment l'information auditive est traitée par le cerveau. On a pu par contre étudier comment elle était finalement perçue, dans le cadre d'une science spécifique appelée psychoacoustique, une branche de la psychophysique, a pour objet l'étude expérimentale des relations quantitatives entre les stimulus acoustiques mesurables physiquement et les réponses de l'ensemble du système auditif de l'être humain: sensations et perceptions auditives. [12]

I.3.2.1 Les sensations auditives

Ces sensations sont : le phone, la hauteur perçue ou tonie, le timbre et la durée. [5]

La hauteur (ou tonie) d'un son pur est liée à la fréquence de l'onde sonore. Mais, au-delà de 1000 Hz, il faut plus que doubler la fréquence pour percevoir un doublement de la hauteur (il faut un signal de 3120 Hz pour avoir une sensation de 2000 Hz). L'échelle de tonie est graduée en mels. Un écart constant en mels est perçu comme un écart constant en hauteur. Dans le domaine d'audition, l'oreille est capable de discerner 1400 hauteurs distinctes.

L'unité permettant de mesurer le niveau de l'intensité subjective est le phone. Les courbes d'isosoniques (figure I.12) montrent que la courbe de réponse de l'oreille dépend de la fréquence et de la pression sonore. Elles révèlent un maximum de sensibilité dans la plage [500Hz-10kHz], en dehors de laquelle les sons doivent être plus intenses pour être perçus. L'oreille peut distinguer 280 niveaux différents d'intensité.

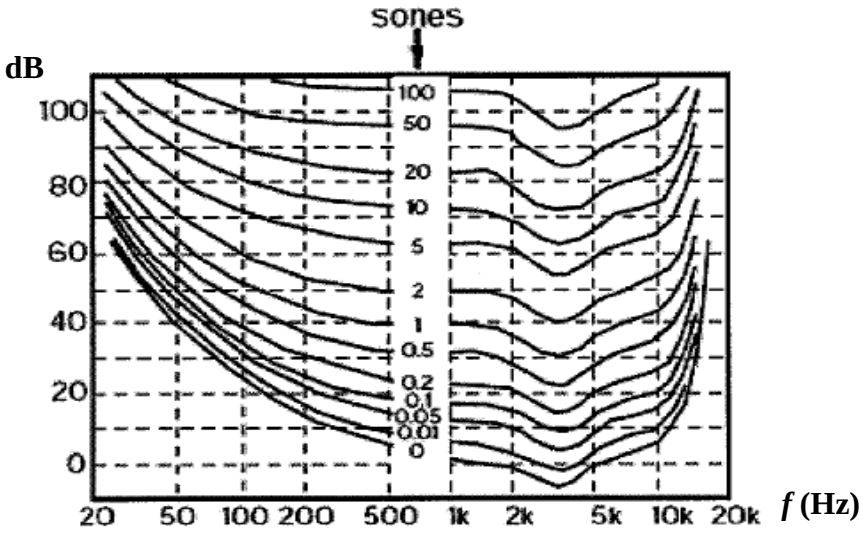


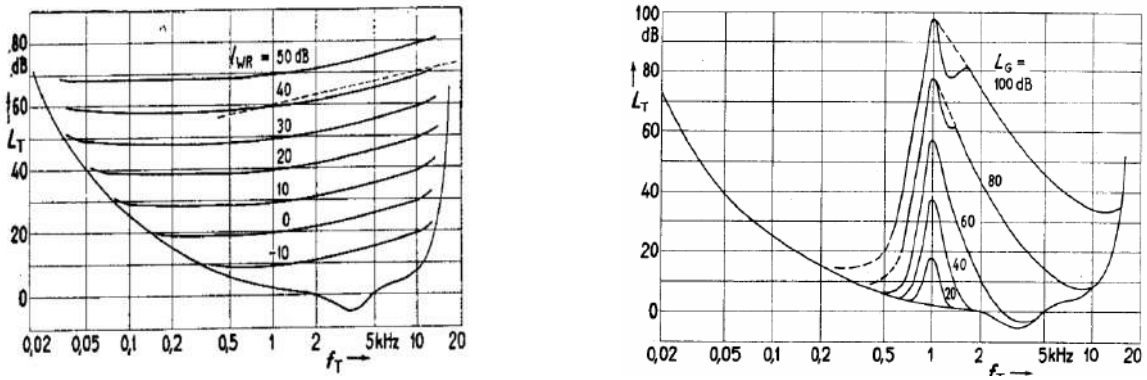
Figure I.12. Courbes isosoniques en champ ouvert [8]

I.3.2.2 Masquage et bande critique

I.3.2.2.a Le Masquage

La notion de masquage est étroitement liée à la perception de l'intensité autant qu'à celle des hauteurs. Le principe de ce phénomène de masquage est qu'un son fort en masque (empêche de percevoir) un autre plus faible, surtout s'ils sont dans des fréquences proches. [13]

Les courbes obtenues quand on enregistre les pressions acoustiques nécessaires pour qu'un son test devienne audible ou inaudible en présence d'un son parasite ou masquant sont tout aussi importantes pour caractériser l'ouïe.



(a) (b)

Figure I.13. (a) Courbes d'effet de masque d'un bruit blanc ;(b) Courbes d'effet de masque de bruits à bande étroite. [12]

I.3.2.2.b Les Bandes Critiques

Fletcher a suggéré que le système auditif se comporte comme un banc de filtres qui se chevauchent et dont les fréquences centrales s'échelonnent continûment. Il a modélisé cette fréquence par une série de filtres rectangulaires dont la largeur a été appelée bande critique, et un son dont la fréquence est à l'intérieur de cette bande peut influencer la perception des autres sons dans la même bande, mais pas en dehors de cette bande critique. [5]

L'existence des bandes critique a permis de montrer que l'oreille découpe son spectre en bandes critiques, qu'elle examine ensuite le niveau de chaque bande et le compare au seuil d'audition de la bande correspondante. Si le niveau est supérieur au seuil, la bande est audible et un certain stimulus est transmis au cerveau. Dans le cas contraire, la bande est inaudible et est dite masquée. [8]

I.4 Conclusion :

Ce premier chapitre permis de définir le comportement global du signal vocal, nous avons passé en revue le principe de son audition ainsi que les caractéristiques générales du signal vocal.

Principalement, l'information d'un message parlé réside dans les fluctuations de la pression de l'air, engendrées, puis émises, par l'appareil phonatoire. Ces fluctuations constituent le signal vocal ; elles sont détectées par l'oreille. Les résultats sont transmis au cerveau, qui interprète.

Le prochain chapitre est consacré au traitement numérique d'un signal vocal, nous présenterons les différents outils de traitements de la parole, et quelques méthodes de paramétrisation du signal vocal.

Chapitre II

TRAITEMENT NUMERIQUE DU SIGNAL VOCAL

II.1 Introduction

Le traitement du signal vocal s'inscrit dans une succession de procédures, que ce soit pour la reconnaissance automatique ou pour la synthèse de la parole. Analyse et synthèse sont deux activités duales, l'analyse fournissant une description du signal acoustique que la synthèse utilise pour le reproduire.

Le traitement est aussi utilisé pour réduire la redondance du signal vocal, ou en extraire des paramètres pertinents pour la reconnaissance. Cette réduction par traitement permet aussi de compresser l'onde avant le stockage ou la transmission.

L'objectif de ce chapitre est de connaitre les caractéristiques particulières du signal de parole afin de déterminer la méthode d'extraction des paramètres des informations exploitables contenues dans l'onde sonore. Pour ce faire, nous décrivons les différents outils nécessaires au traitement de la parole, nous commencerons par un bref aperçu sur le traitement à court-terme, ensuite nous exposerons les concepts de base de la paramétrisation du signal vocal avec la présentation de l'analyse par prédiction pour les coefficients LPC(Linear Predictive Coding), ainsi que les coefficients LPCCs et MFCCs.

II.2 la chaine de traitement du signal vocal

Le signal de parole est considéré comme une forme d'onde transportant simultanément des informations linguistiques et émotionnelles.

Le $n^{ème}$ échantillon est représenté par le terme x(n) à l'aide d'un convertisseur analogique numérique (CAN), qui sélectionne la valeur du signal analogique d'entrée à chaque instant $T_n=T_{n-1}+T_e$, où T_e est la période d'échantillonnage donnée par la formule suivante:

$$T_e=1/f_e \qquad (2.1)$$

Il est possible de représenter la chaîne de traitement du signal vocal au moyen de la figure (II.1) [15], qui donne un bon aperçu global des différentes tâches à effectuer avant d'entrer dans le processus de reconnaissance.

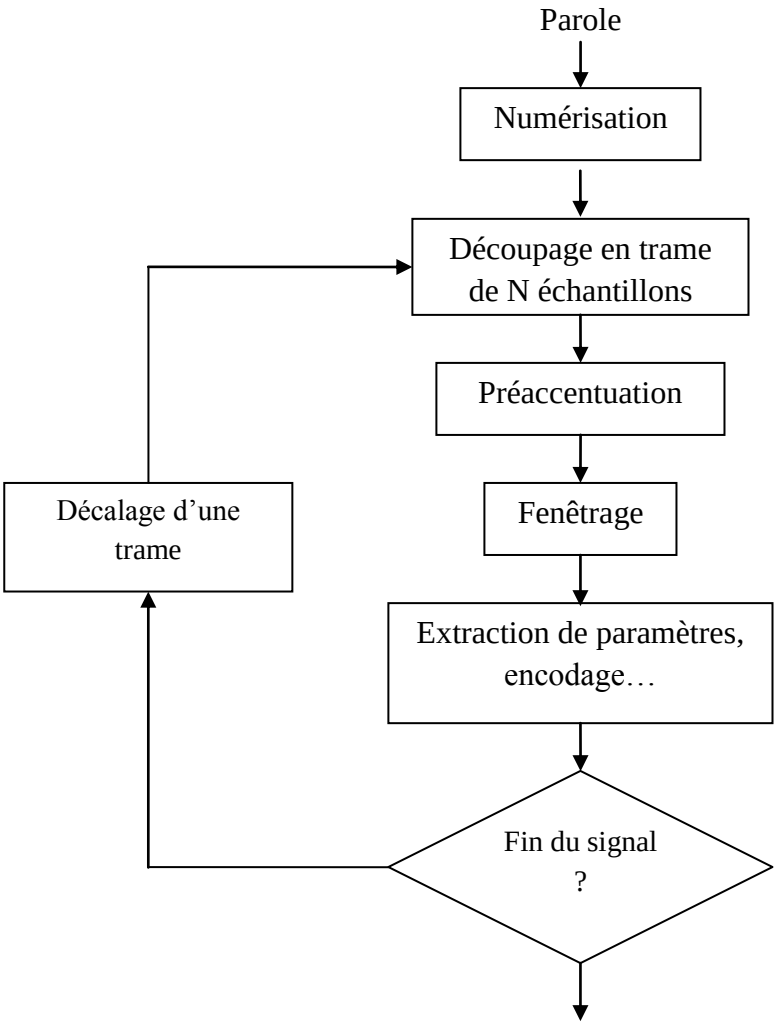


Figure II.1: La chaine de traitement de signal [15].

II.2.1 Prétraitement du signal vocal

Pour obtenir une représentation précise de l'état du signal à un instant donné, il est important de travailler sur des tranches temporelles très courtes, bien que suffisamment longues pour contenir assez de renseignements sur l'information contenue dans le signal. Généralement, on préfère donc travailler sur des périodes allant de 10 à 30 ms, durant lesquelles le signal est considéré comme étant quasi stationnaire [2].

Une fois le signal numérisé, la première action d'un système de traitement de la parole est de regrouper les échantillons entrants en un bloc de taille finie, appelé trame. Par la suite, nous noterons N la taille de la trame, en nombre d'échantillons, et toutes les opérations effectuées, dites alors « court terme », ne concerneront que cette trame. Le reste du signal sera traité en décalant la trame de façon répétée, d'un nombre d'échantillons L, L était inférieur à N, jusqu'à la fin du signal. C'est ensuite sur ces trames de parole que peuvent s'appliquer les opérations destinées à préparer le signal d'entrée pour la phase d'extraction des informations.

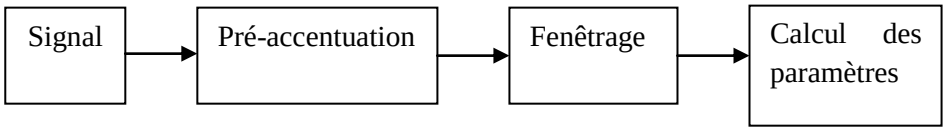


Figure II.2. Pré-traitement et extraction des paramètres.

II.2.1.1 La pré-accentuation [16]

Le filtre de pré accentuation (c'est un filtre numérique), souvent un filtre à réponse impulsionnelle finie(R.I.F) non-récursif du 1er ordre (donc à un seul coefficient) dont la fonction de transfert est donnée par:

$$H_1(z)=1-a_1z^{-1} \quad (2.2)$$

Les valeurs typiques de a_1 sont comprises entre [0.4 et 1.0], une valeur proche de 1 est souvent utilisée pour la reconnaissance (de l'ordre de 0.96).

Dans le domaine temporel, cette opération se traduit par une différentielle:

$$x'(n)=x(n)-a_1x\ (n-1) \quad (2.3)$$

Les avantages procurés par l'utilisation d'un tel filtre de pré accentuation sont:

- Les segments voisés du signal possèdent une pente spectrale négative approximativement de 20dB par décade (donc une atténuation de 20 dB par décade),ceci est dû aux caractéristiques physiologique de l'appareil phonatoire. La pré-accentuation sert à annuler cette atténuation avant analyse spectrale par exemple.
- L'appareil phonatoire est plus sensible à la région se trouvant au dessus de 1KHz, la pré accentuation amplifie cette région.
- On note aussi qu'un tel filtre de pré-accentuation relève les fréquences au-dessus de 5KHz, une région pour laquelle le système d'audition devient de moins en moins sensible.

II.2.1.2 Le fenêtrage

Généralement le découpage du signal dans le domaine temporel équivaut à multiplier le signal par une fonction rectangulaire, ce qui équivaut à une convolution dans le domaine fréquentiel entre le spectre analysé et celui de la fenêtre. Dans la majorité des cas la fenêtre rectangulaire s'avère trop brutal. En effet il a été démontré [17] que toute variation rapide dans le domaine temporel correspond à des hautes fréquences dans le domaine fréquentiel qui se traduit par des ondulations sur le spectre. Alors on lui préfère d'autres fenêtres plus douces.

La figure II.3 montre les fenêtres les plus utilisés, on trouve la fenêtre de Bartlett; de type triangulaire, de forme convexe comme celle de Kaiser, ou bien sinusoïdale comme celle de

Hamming, Von Hann et Blachman, l'objectif ici n'est pas de toutes les décrire, mais de présenter celle qui semble la plus adaptée à nos besoins.

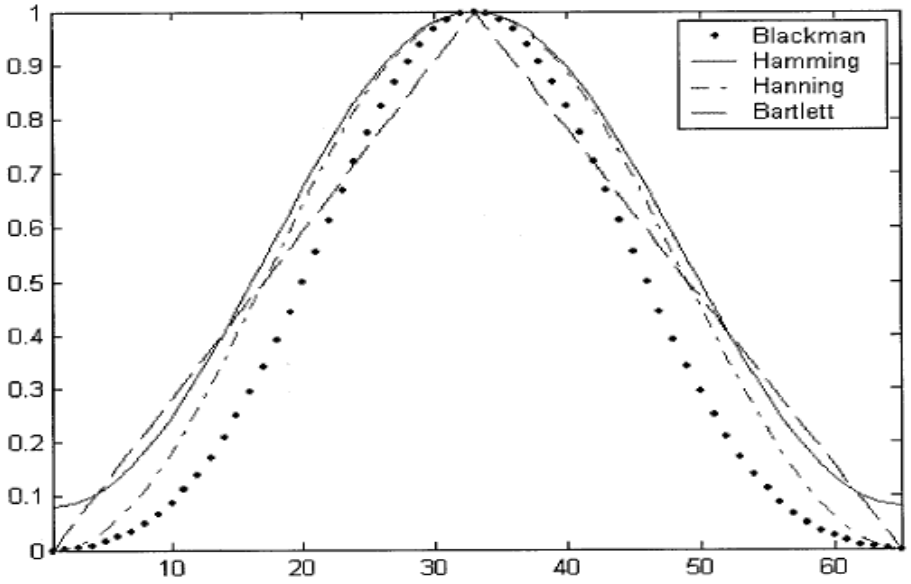


Figure II.3. Exemples de fenêtres de pondération [18]

En pratique, la fenêtre de Hamming, est la plus utilisée dans le domaine de reconnaissance vocale, car elle n'annule pas le signal aux extrémités, permettant ainsi une meilleure transition entre les trames successives de signal, que l'on peut réaliser à l'aide de la formule (2.4)

$$W(I) = \begin{cases} k_1 - k_2 \cos\left(\frac{2\pi l}{L-1}\right) + k_3 \cos\left(\frac{4\pi l}{L-1}\right) & 0 \le l \le L-1 \\ 0 & \text{autrement} \end{cases} \quad (2.4)$$

Où k_1= 0.54, k_2=0.46 et k_3 =0.

La figure II.4 est une illustration du spectre concernant la fenêtre de Hamming et celui de la fenêtre rectangulaire. Dans cette figure on voit clairement que la fenêtre de Hamming permet une grande atténuation en dehors de la bande passante comparativement à la fenêtre rectangulaire d'où son avantage.

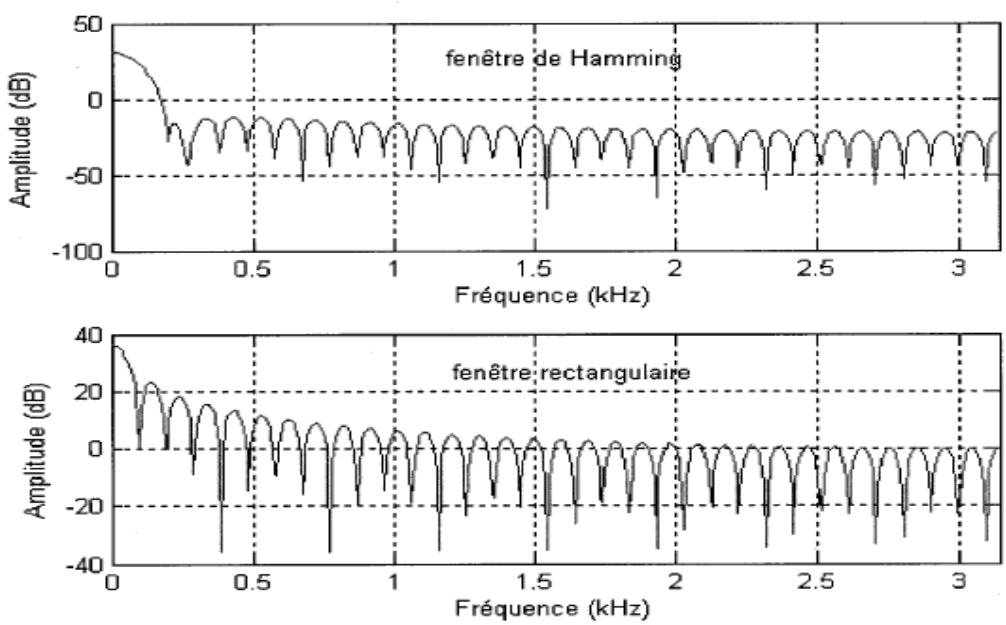


Figure II.4. Spectre des fenêtres de Hamming et rectangulaire [19]

Cette première phase de traitement du signal permet donc de préparer idéalement le signal en vue de la phase d'extraction de paramètres décrivant l'information qu'il contient, phase dont nous décrivons plusieurs méthodes fréquemment utilisées, au cours de la partie suivante.

II.3 Les outils du traitement du signal

Le signal de parole est un processus aléatoire non stationnaire, or les outils de traitement du signal conventionnels sous-entendent la stationnarité du signal, alors on va exploiter le fait que le signal parole soit quasi stationnaire sur des courts segments de parole appelés "frame" en anglais.

Au cours de cette partie, nous détaillerons principalement les outils nécessaires à la conception de notre méthode de reconnaissance, ou tout du moins certaines notions qui nous seront fort utiles par la suite.

II.3.1 Énergie à court-terme

Le premier paramètre intuitif pour caractériser le signal de parole l'énergie, c'est l'un des outils qui permettent de fournir une représentation fidèle des variations de l'amplitude du signal vocal x(n) dans le temps voir figure(II.5). Elle est définie par la formule [20] :

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \qquad (2.5)$$

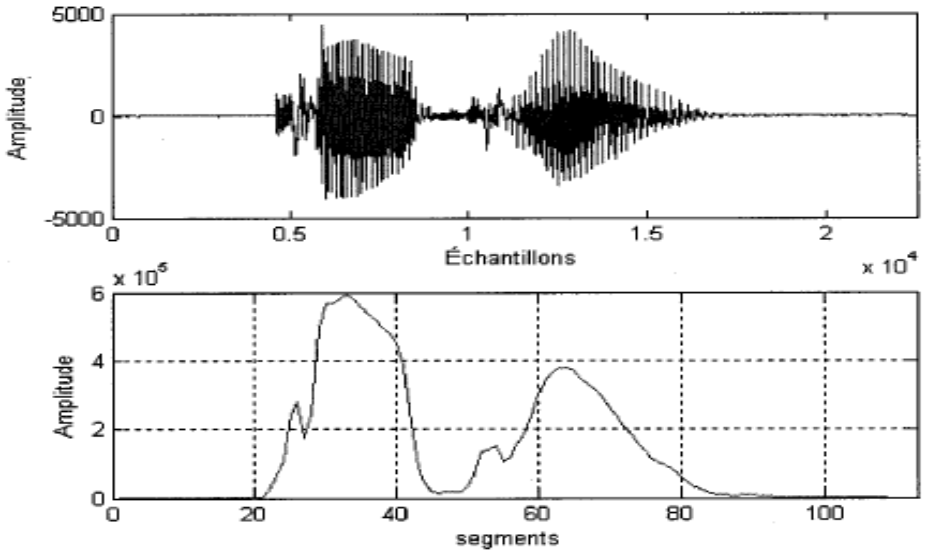


Figure II.5. L'énergie à court terme d'un signal vocal

II.3.2 le taux de passage par zéro pour un signal vocal

Le taux de passage par zéro (Zero Crossing Rate, en anglais) représente le nombre de fois que le signal, dans sa représentation amplitude/temps, passe par la valeur centrale de l'amplitude (généralement zéro).

Pour un signal échantillonné x(n), on dit qu'il ya passage par zéro si deux échantillons successifs sont de signe opposés:

$$\begin{cases} x(n).x\,(n-1) < 0 \\ 0 < n < N-1 \end{cases} \quad (2.6)$$

Avec

N: taille du signal

Le taux de passage zéro court terme peut être estimé par la formule:[2]

$$Z_n = \frac{1}{2}\sum_m |\mathrm{sgn}[x(m)] - \mathrm{sgn}[x(m-1)]|w(n-m) \quad (2.7)$$

Où

$$w(n) = \begin{cases} \frac{1}{N} & ,0 \le n \le N \\ 0 & ,\text{ailleurs.} \end{cases} \quad (2.8)$$

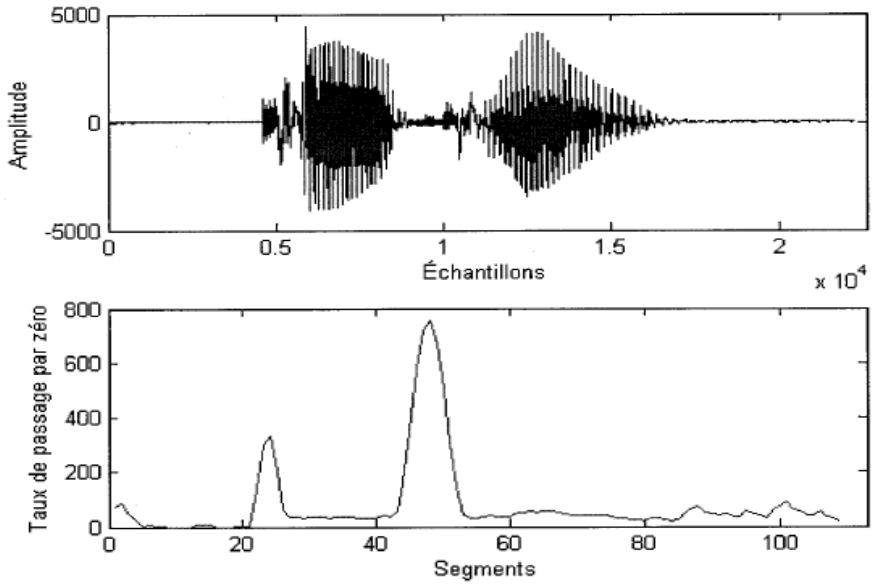


Figure II.6. Le taux de passage par zéro d'un signal vocal.

Remarque:

- Le taux de passage par zéro est élevé pour le son non voisé et faible pour le son voisé.
- C'est un outil important pour la détection du début et fin de la parole dans un signal vocal.

II.3.3. Etude spectrale

L'outil le plus utilisé pour étudier l'intérieur du signal (observer ses composants spectrales) est la transformé de Fourier qui donne le contenu harmonique d'un signal, c'est-à-dire la répartition de ses composantes fréquentielles ainsi que leurs amplitudes. Pour un signal discret, la définition de la transformée de Fourier continue est donnée par la formule 2.9 [21].

$$X\left(e^{j\omega}\right) = \sum_{n=0}^{\infty} x(n) \times e^{-j\omega n} \quad (2.9)$$

La transformée de Fourier d'un signal de taille finie (à court terme) est donnée par la formule 2.10 [22].

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi \frac{nk}{N}} \quad (2.10)$$

L'inversion de la TFD est donnée par la formule 2.11 [22]

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{j2\pi \frac{nk}{N}} \quad (2.11)$$

II.3.4 l'autocorrélation à court terme

L'utilisation de l'autocorrélation pour la détection de la fréquence fondamentale est très classique en traitement de la parole. Pour un signal x_n, la fonction d'autocorrélation $\hat{r}(k)$ est définie par: [23]

$$\hat{r}(k) = \frac{1}{N-k}\sum_{n=0}^{N-k-1} x_n x_{n+k} \quad (2.12)$$

Où le paramètre N contrôle l'horizon sur lequel est calculée la fonction. On définit également La fonction d'autocorrélation normalisée r(k) par:

$$\hat{r}(k) = \frac{\sum_{n=0}^{N-k-1} x_n x_{n+k}}{\left(\sum_{n=0}^{N-k-1} x_n^2\right)^{1/2}\left(\sum_{n=0}^{N-k-1} x_{n+k}^2\right)^{1/2}} \quad (2.13)$$

L'idée de l'utilisation de la fonction d'autocorrélation est de déterminer à quel point deux échantillons successifs d'un signal se ressemblent [17]. Parmi ses autres applications on peut citer son utilisation pour l'estimation de la fréquence du fondamental (ou pitch). En effet ce dernier se manifeste par un pic dans la fonction de l'autocorrélation qu'il faut isoler adéquatement pour pouvoir l'évaluer.

II.4 La paramétrisation du signal

Il existe de nombreuses façons de paramétrer le signal, la plupart d'entre elles se basant sur certaines propriétés spécifiques au signal vocal pour en extraire un maximum d'informations, en un minimum de données. C'est notamment le cas du codage par prédiction linéaire (souvent appelé LPC, pour Linear Prédiction Coding*)*, qui est une conséquence de la modélisation autorégressive de la parole, et qui peut ensuite donner lieu à d'autres méthodes de paramétrisation dérivées de ses coefficients.

II.4.1 La méthode par prédiction linéaire

L'objectif de cette méthode est la représentation directe du signal vocal sous la forme d'un nombre limité de paramètre. Sa puissance est fondée sur un modèle simple de production de la parole qui s'approche du système phonatoire.

II.4.1.1 Modèle auto régressif

La production d'un signal vocal peut être assimilée à l'excitation d'un filtre par une source pseudo-périodique pour les signaux voisés, par contre pour les signaux non voisés, l'excitation est un bruit blanc de moyenne nulle et de variance unité.

Nous pouvons considérer le signal vocal comme étant la convolution de la source u(n) et de la réponse impulsionnelle du filtre représentant le conduit vocal, h(n), comme présenté la figure II.7

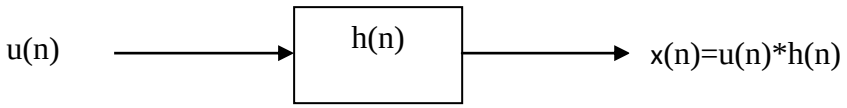


Figure II.7. Modèle simplifié de la production de la parole.

Où u(n) est donnée par la formule :

$$u(n) = \sum_k \delta(n-k) \tag{2.14}$$

Le signal x(n) au domaine discret peut s'écrire:

$$X(z)=U(z)\,H(z) \tag{2.15}$$

Où H(z), qui représente la transmittance du conduit vocal, est un filtre *tous-pôles* d'ordre p, dont la formule est donnée par l'équation (2.16):

$$H(z) = \frac{\sigma}{1+\sum_{i=1}^{P} a(i).z^{-i}} \tag{2.16}$$

En rajoutant un terme correspondant au gain de la source d'excitation, nous pouvons donc dire que la production de la parole peut être modélisée par l'équation (2.17):

$$X(z) = \frac{\sigma.U(z)}{1+\sum_{n=1}^{P} a(i).z^{-i}} \tag{2.17}$$

Si l'on repasse la formule précédente dans le domaine temporel, nous obtenons l'expression (2.18):

$$x(n) + \sum_{i=1}^{P} a(i)\, x(n-i) = \sigma.u(n) \tag{2.18}$$

Ainsi on peut prévoir l'échantillon x(n) à partir d'une combinaison linéaire des P échantillons qui le précédent, en plus de l'excitation u(n). Ce modèle (figure II.8) est appelé modèle auto régressif (AR). Les coefficients a(i) représentent les caractéristiques du conduit vocal. Ces coefficients sont d'ailleurs appelés « coefficients de prédiction », car en cas d'excitation nulle, chaque échantillon x(n) pourrait être prédit à partir des p échantillons précédents, associés aux coefficients de prédiction.

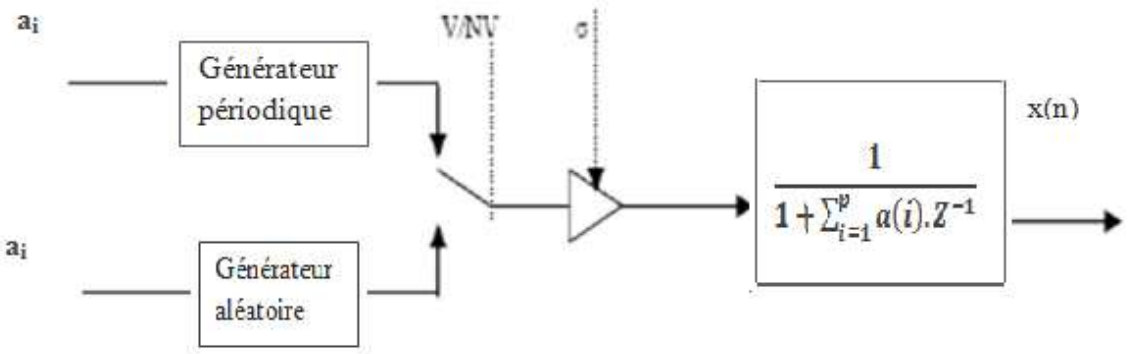


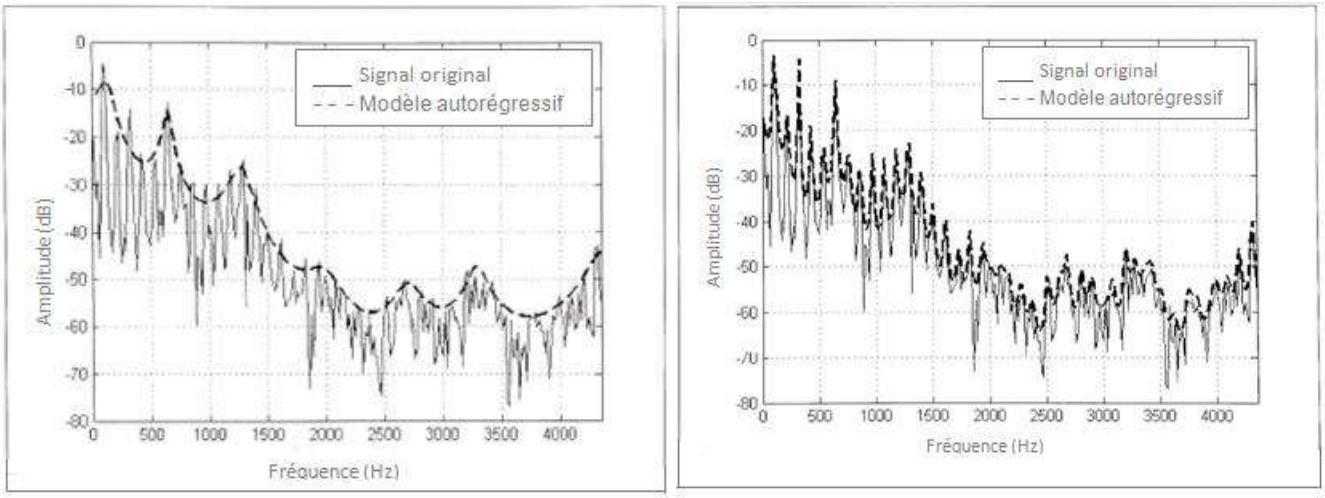
Figure II.8. Modèle auto régressif de production de parole [23]

La densité spectrale du signal synthétisé est donnée par la formule:

$$S_x(e^{j\omega}) = S_U(e^{j\omega}).\left|\frac{\sigma}{1+\sum_{k=1}^{P} a(k).(e^{j\omega})^{-k}}\right|^2 \qquad (2.19)$$

Où $S_u(e^{jw})$ est la densité spectrale de la source.

Il est utile de visualiser l'importance de l'ordre de prédiction dans la précision de modélisation d'un signal. La figure II.9 présente une comparaison entre la densité spectrale autorégressive d'ordre 20 et d'ordre 100 du son "a".



(a) (b)

Figure II.9.densité spectrale d'un modèle autorégressif. (a): d'ordre 20

(b): d'ordre 100

II.4.1.2 L'estimation des coefficients de prédiction:

Le problème de l'estimation d'un modèle LPC, souvent appelée analyse LPC revient à déterminer les coefficients du filtre connaissant le signal de sortie, mais pas l'entrée. Il est par conséquent nécessaire d'adopter un critère, afin de faire un choix parmi l'infinité de solutions

possibles. Le critère classiquement utilisé est celui de la minimisation de l'énergie de l'erreur de prédiction. La mise en équation de ce problème conduit aux équations dites de Yule-Walker.

Nous avons vu que le signal vocal obéissait au modèle autorégressif d'ordre p défini par la formule 2.19, faisant de lui un signal autorégressif Cette formule peut également s'écrire d'une autre façon, présentée par l'équation 2.20.

$$\bar{x}(n) = -\sum_{k=1}^{P} a_k x(n-k) \tag{2.20}$$

Où x(n) est une estimation du signal original. La différence entre cette estimation et le signal réel est appelée erreur de prédiction linéaire, notée e(n), définie par la formule 2.21

$$e(n) = x(n) - \tilde{x}(n) = x(n) + \sum_{k=1}^{P} a_k x(n-k) \tag{2.21}$$

D'autre part on définit l'énergie résiduelle de prédiction par [2] :

$$E_n = \sum_m e^2(m) = \sum_m (x_n(m) - \tilde{x}_n(m))^2 \tag{2.22}$$

La minimisation de cette erreur est à la base de la détermination des coefficients de prédiction a_i. Ainsi les coefficients a_i optimaux seront tirés de l'équation suivante:

$$\frac{\partial E_n}{\partial a_k} = 0 \, , k = 1,2, \dots \dots . , p \tag{2.23}$$

Il reste ainsi à résoudre un système de p équations à p inconnues, dites équations de Yule-Walker, définies par la relation (2.24)

$$-\sum_{n=0}^{N+p-1} x(n-i)\, x(n) = \sum_{k=1}^{p} a_k \sum_{n=0}^{N+p-1} x(n-i).x(n-k), \qquad i = 1, \dots \dots . , p \tag{2.24}$$

Pour simplifier ces équations, on peut faire intervenir la formule de l'autocovariance, donnée par l'équation 2.25, permettant ainsi d'exprimer les équations de Yule-Walker sous la forme donnée par la formule 2.26.

$$\emptyset(i,k) = \sum_n x(n-i).x(n-k) \tag{2.25}$$

$$\emptyset(i,0) = \sum_{k=1}^{p} a_k . \emptyset_n(i,k) \tag{2.26}$$

Pour résoudre ce système, les deux méthodes les plus couramment utilisées sont la méthode de l'autocorrélation, introduite pour la première fois par Itakura et Saito en 1968 [24], et la méthode de la covariance, établie par Atal et Schroeder en 1968. Nous allons présenter la première méthode, Son hypothèse de départ est la réduction de la fonction d'autocovariance, présentée en 2.26, à la fonction d'autocorrélation, comme présenté par l'équation 2.27.

$$\emptyset(\mathrm{i},\mathrm{k}) = \mathrm{r}(\mathrm{i}-\mathrm{k}) = \sum_{\mathrm{n}=0}^{\mathrm{N}-1-(\mathrm{i}-\mathrm{k})} \mathrm{x}(\mathrm{n}).\mathrm{x}(\mathrm{n}+\mathrm{i}-\mathrm{k}) \qquad \mathrm{i} = 1,\ldots..,\mathrm{P} \tag{2.27}$$

P étant l'ordre de la prédiction, et r_n la fonction d'autocorrélation.

La forme matricielle de cette équation donne:

$$\begin{bmatrix} \mathrm{r_n}(0) & \mathrm{r_n}(1) & \mathrm{r_n}(2) & \cdots & \mathrm{r_n}(\mathrm{p}-1) \\ \mathrm{r_n}(1) & \mathrm{r_n}(0) & \mathrm{r_n}(1) & \cdots & \mathrm{r_n}(\mathrm{p}-2) \\ \mathrm{r_n}(2) & \mathrm{r_n}(1) & \mathrm{r_n}(0) & \cdots & \mathrm{r_n}(\mathrm{p}-3) \\ \vdots & \vdots & \vdots & & \vdots \\ \mathrm{r_n}(\mathrm{p}-1) & \mathrm{r_n}(\mathrm{p}-2) & \mathrm{r_n}(\mathrm{p}-3) & \ldots & \mathrm{r_n}(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} r_n(1) \\ r_n(2) \\ r_n(3) \\ \vdots \\ r_n(p) \end{bmatrix} \tag{2.28}$$

On peut remarquer que nous sommes en présence d'une matrice symétrie, dont tous les éléments de la diagonale sont égaux, il s'agit alors d'une matrice dite matrice de Toeplitz, une caractéristique très important qui va faciliter la résolution de ce système d'équation. En effet il existe une méthode simple pour le résoudre, c'est l'algorithme de Levinson-Durbin[25], dont l'idée de base est l'utilisation d'une récursion pour arriver à la (i+1)^{ème} solution, à partir de la i^{ème} solution, et ainsi de suite jusqu'à atteindre l'ordre P, et à chaque i^{ème} niveau on obtient le coefficient a_i du filtre qui permet de minimiser l'énergie résiduelle, ou aussi l'erreur de prédiction.

II.4.2 L'analyse homomorphique

L'analyse cepstrale a pour objectif de séparer les contributions de la source et du conduit vocal, apportant ainsi certains avantages par rapport à la transformée de Fourier : d'une part l'observation d'un spectre lissé correspondant aux caractéristiques du conduit vocal, et d'autre part une détermination aisée de la fréquence fondamentale.[5]

Le point de départ de l'analyse cepstrale est le traitement homomorphique, son principe consiste à calculer le logarithme de la transformé en z du signal pour séparer les deux composants de la convolution. Le résultat de cette méthode, des coefficients plus fiables pour la reconnaissance que les classique LPC.

Le principe en est illustré sur la figure II.10:

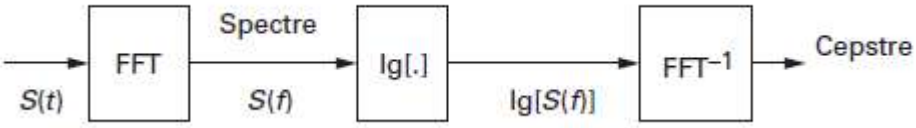


Figure II.10. Principe de l'analyse homomorphique

La figure II.11 est un exemple d'analyse homomorphique, on remarque la présence de fréquence du fondamental sous forme de pic périodique, notons aussi que seulement les premiers coefficients cepstraux sont nécessaire pour la reconnaissance.

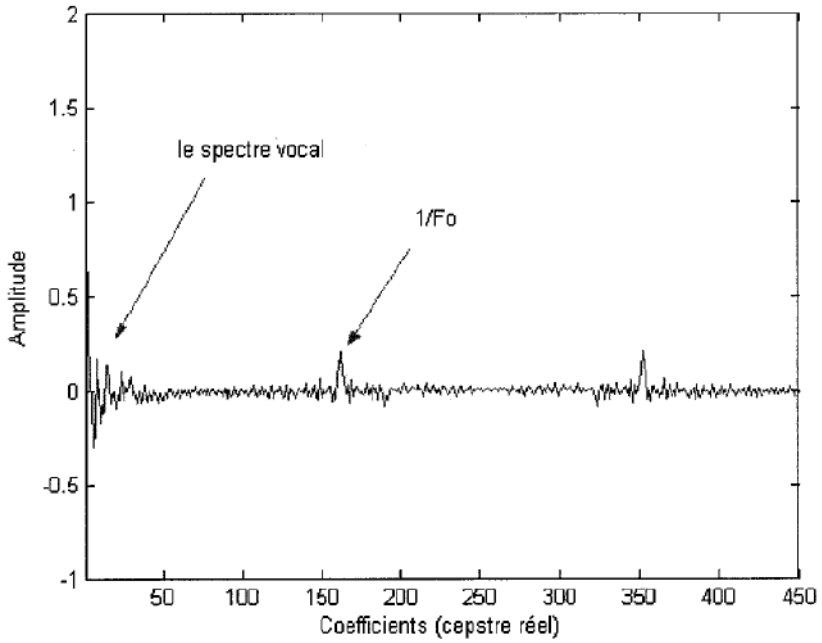


Figure II.11. Exemple d'analyse homomorphique

II.4.2.1 Les coefficients cepstraux:

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. Les techniques de paramétrisation acoustique sont nombreuses néanmoins on peut les regrouper en trois grandes familles:

- Analyse par bancs de filtres.
- Analyse par transformée de Fourier.
- Analyse par prédiction linéaire.

Les coefficients cepstraux issus d'une analyse par transformée de Fourier caractérisant bien la forme du spectre et permettent de séparer l'influence de la source de celle du conduit vocal. Ils peuvent aussi être calculés à partir d'une analyse de prédiction linéaire.

Le cepstre du signal de parole est défini comme la transformée de Fourier inverse du logarithme du module du spectre. Pour ce signal, la source d'excitation glottique est convoluée avec la réponse impulsionnelle du conduit vocal considéré comme un filtre linéaire

$$s(t) = e(t) * h(t) \tag{2.29}$$

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

L'application à l'équation (2.29) du logarithme du module de la transformée de Fourier donne:

$$log|S(f)| = \log|E(f)| + \log|H(f)| \tag{2.30}$$

Par une transformée de Fourier inverse, on obtient :

$$s'(cef) = e'(cef) + h'(cef) \tag{2.31}$$

La dimension du niveau domaine est homogène à un temps et s'appelle la quéfrence. Le nouveau domaine s'appelle donc le quéfrentiel. Un filtrage dans ce domaine s'appelle liftrage.[26]

Principalement on distingue deux types de coefficients cepstraux [19]:

- Les coefficients LPCC.
- Les coefficients MFCC.

On opte pour l'un ou l'autre des deux types selon des considérations pratiques dictées par les exigences de l'application ou selon les performances obtenues lors des expériences. Nous utiliserons coefficients MFCCs dans notre projet.

II.4.2.2. Les coefficients MFCCs

La MFCC (Mel Frequency Cepstral Coefficients) est une extraction de caractéristique du signal développée autour de la FFT et de la DCT, ceci sur une échelle de Mel.

La figure (II.12) montre les phases d'obtention de coefficients MFCC à partir d'un signal.

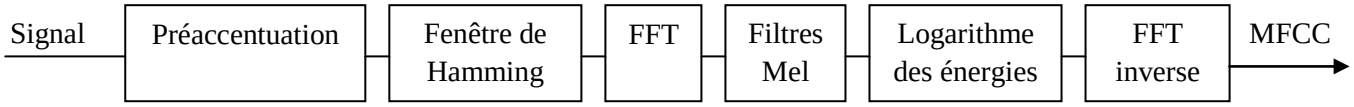


Figure II.12. Principe de calcul des coefficients MFCCs

La MFCCs se décompose en phases:

Phase1: découper le signal préaccentué en plusieurs fenêtres qui se recoupent entre elles. On applique la MFCCs à chaque fenêtre.

Phase 2: appliquer en suite la FFT à la fenêtre pour en ressortir le spectre, on obtient donc le spectre. La phase de la FFT du signal de parole ne contient pas d'informations suffisamment pertinents pour la reconnaissance de la parole, alors il est judicieux de garder juste la partie réelle, c'est ce qu'on va faire on prenant seulement le module au carré de la FFT.

Phase 3 : On passe à l'échelle de Mel. En effet, après des études sur l'oreille humaine, il a été montré que l'homme se base sur une échelle fréquentielle spécifique.

Pour simuler l'oreille humaine, il faut passer par un Banc Filtre, un filtre pour chaque fréquence que l'on cherche. Ces filtres ont une réponse de bande passante triangulaire. Pour connaitre l'intervalle entre chaque filtre, on utilise une constante: Mel-Frequency interval.

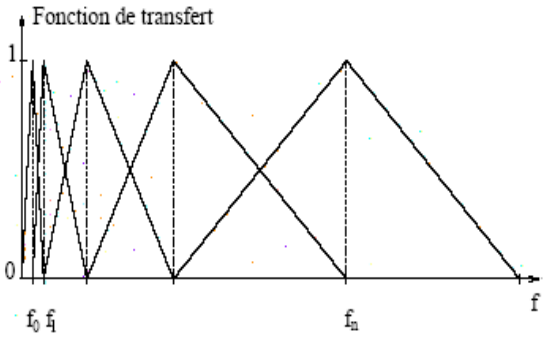


Figure II.13.Banc de filtres sur l'échelle. Mel.

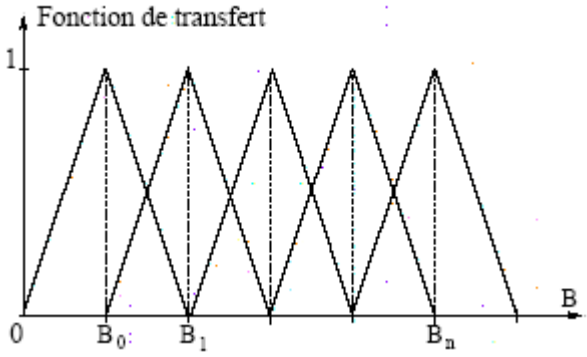


Figure II.14 Banc de filtres sur l'échelle linéaire

Phase 4: le logarithme est appliqué pour transformer la multiplication en addition.

Phase 5: Pour finir, on travaille avec le Cepstre, on convertis le spectre logarithmique de Mel en temps au moyen de la DCT (Discret Cosinus Transform) La formule de cette transformation est simple [28]:

$$\tilde{c} = \sum_{k=1}^{K}(\log \tilde{E}_k)\cos[n(k-\frac{1}{2})\frac{\pi}{K}],\ n=1,2,\ldots,L. \qquad (2.32)$$

Avec :

K: le nombre de filtres

L: le nombre de coefficients qu'on veut avoir.

E_k : l'énergie à la sortie des filtres, k=1,2,…,K

II.4.2.3 Les coefficients LPCCs

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 (*Linear Prediction Cepstral Coefficients*) sont dérivés directement des coefficients LPC.

Le processus de l'extraction des paramètres LPCCs est illustré par la figure II.15.

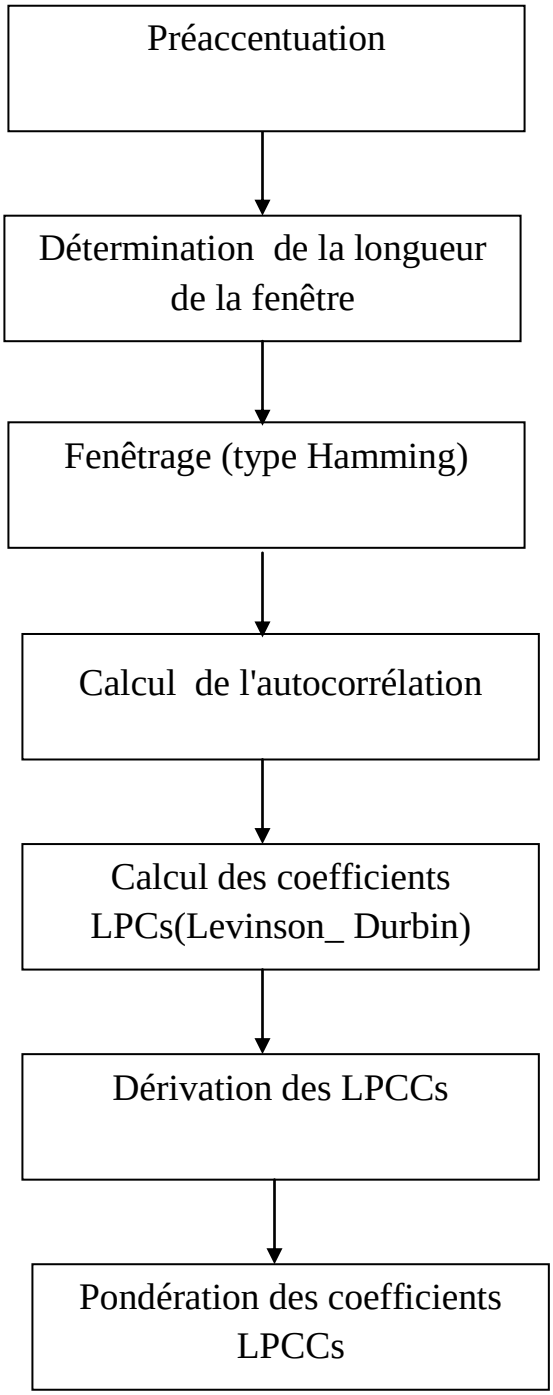


Figure II.15 : processus de l'extraction des paramètres LPCCs pondérés

- **Préaccentuation**

La préaccentuation du signal permet d'amplifier les hautes fréquences, afin de leurs donner une amplitude supérieure au bruit de fond.

Elle consiste en un passage dans un filtre de transmittance $1-\mu z^{-1}$, avec μ compris entre 0.9 et 1 (0.93 pour notre cas).

- **Calcul des coefficients LPC**

La méthode utilisée pour déterminer les paramètres LPC à partir des coefficients de l'autocorrélation est l'algorithme de Levinson-Durbin , son principe est le suivant [29] :

$$E^{(0)} = r(o)$$

$$k_i = \left(r(i) - \sum_{j=1}^{l=1} \alpha^{(i-1)}\ r(|i-j|) \right) ,1 \leq i \leq p$$

$$\alpha_i^{(i)} = k_i \tag{2.29}$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i\ \alpha_{i-j}^{i-1}$$

$$E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)}$$

Où : $E^{(0)}$ est le terme d'erreur initial

K_i : le coefficient de réflexion

$\alpha_i^{(i)}$: le coefficient prédicteur

Cet ensemble d'équations est résolu d'une manière récursive. La solution finale donne les coefficients LPCs α_m :

$$\alpha_m = \alpha_m^{(p)} , 1 \leq m \leq p \tag{2.30}$$

- **Dérivation des coefficients cepstraux**

L'obtention des coefficients LPCCs se fait par dérivation à partir des coefficients LPCs en utilisant les formules de conversion suivantes [19]:

$$C_n = \begin{cases} a_{\ n} , n = 1 \\ a_n \sum_{k=1}^{n} \left(\frac{n-k}{n}\right) c_{n-k} a_k , 2 \leq n \leq p \\ \sum_{k=1}^{p} \left(\frac{n-k}{n}\right) c_{n-k} a_k , p \leq n \leq q \end{cases} \tag{2.31}$$

Pour le nombre de coefficients cepstraux, nous avons choisi q =12

On ce qui concerne le nombre de coefficients cepstraux, il a été démontré qu'un nombre égale à trois fois l'ordre de prédiction P, est suffisant pour fournir une bonne représentation du spectre vocal.

II.5 Conclusion

Ce chapitre nous aura donc permis de présenter tous les éléments indispensables à la réalisation d'un système de reconnaissance de la parole, en ce qui concerne la phase d'extraction de l'information contenue dans l'onde sonore. Diverses méthodes de traitement ont été étudiées, allant du simple calcul de l'énergie jusqu'aux procédés se basant sur la structure particulière du processus de traitement de la parole, et une méthode de paramétrisation adaptée à nos besoins a été retenue pour représenter le signal. En l'état,

Nous allons pouvoir présenter, au cours du prochain chapitre, les différents procédés généralement utilisés pour transcrire l'information extraite du signal en un jeu de caractères représentant la signification visuelle du son émis par le locuteur.

Chapitre III

LA RECONNAISSANCE VOCALE

III.1 Introduction

Le problème de la reconnaissance automatique de la parole consiste à extraire, à l'aide d'un ordinateur, l'information lexicale contenue dans un signal de parole. Depuis les années 70, des recherches intensives dans ce domaine ont été accomplies par de nombreux laboratoires internationaux. Des progrès importants ont été accomplis grâce au développement d'algorithmes puissant ainsi qu'aux avancées en traitement du signal.

Différents systèmes de reconnaissance de la parole ont été développés, couvrant des domaines aussi vastes que la reconnaissance de quelques mots clés sur lignes téléphoniques, les systèmes à dicter vocaux, les systèmes de commande et contrôle sur PC, et allant jusqu'aux systèmes de compréhension du langage naturel (pour applications limitées).

Dans ce chapitre, nous discutons des principes qui sont à la base de la plupart de ces systèmes.

Malgré que nous ayons appris beaucoup concernant la reconnaissance de la parole et la mise en œuvre de systèmes pratiques et utiles, il reste encore beaucoup de questions fondamentales concernant la technologie pour lesquelles nous n'avons pas de réponses. Il est clair que le signal de parole est un des signaux les plus complexes. En plus de la complexité physiologique inhérente au système phonatoire et des problèmes de coarticulation qui en résultent, le conduit vocal varie également très fort d'une personne à l'autre.

Enfin, la mesure de ce signal de parole est fortement influencée par la fonction de transfert (comprenant les appareils d'acquisition et de transmission, ainsi que l'influence des milieux ambiants).

III.1.1 Niveaux de complexité

Pour bien appréhender le problème de la reconnaissance automatique de la parole, il est bon d'en comprendre les différents niveaux de complexités et les différents facteurs qui en font un problème difficile.

- La redondance: ce qui nous oblige à extraire du signal les informations pertinentes pour réduire le traitement.

- La variabilité: on parle de la variabilité interlocuteur, liée aux caractéristiques automatiques et aux habitudes linguistiques, et de la variabilité intra locuteur tel que le débit d'élocution, l'état émotionnel…etc.
- La continuité: ainsi lorsqu'on écoute une personne parler, on perçoit une suite de mots alors que l'analyse du signal vocal ne permet de déceler aucun séparateur.
- La coarticulation: elle est provoquée par le fait que lors de la prononciation d'un phonème, l'appareil articulatoire se prépare pour la production du phonème suivant.

III.1.2 Classification des méthodes de reconnaissance vocale

Si les recherches sur la reconnaissance vocale ont emprunté de nombreuses voies, c'est en partie à cause des différents types de problèmes à résoudre. Tout d'abord au niveau de la façon d'aborder la tâche à accomplir, et ensuite au niveau des conditions associées à la réalisation de cette tâche déjà mentionné auparavant.

Pour le premier point, il existe effectivement différents types de tâches, selon la nature du système que l'on souhaite créer [5] :

✓ Reconnaissance ou compréhension : à savoir si l'on souhaite simplement reconnaître les mots, ou bien faire en sorte que la machine comprenne également leur signification, comme dans le cadre d'un dialogue;

✓ Système « auto-organisateur » ou « fondé sur des connaissances », entre lesquels la différence vient principalement de la façon de traiter le signal : simplement extraire les informations à l'aide d'outils mathématiques, ou bien se fonder directement sur les propriétés linguistiques pour effectuer un choix;

✓ Reconnaissance globale [2]: dans cette approche l'unité de base, est le mot considéré comme entité globale non décomposable. Elle a pour avantage d'éviter les effets de coarticulation, dans ce type de méthode on compare globalement le mot ou la phrase, aux différentes références stockées dans un dictionnaire. Dans cette catégories on trouve principalement la méthode DTW (Dynamic Time Warping), et la méthode HMM (Hidden Markov Models) et la méthode de la quantification vectorielle (VQ). Cette méthode est utilisée pour les mots isolés et les mots enchaînés.

✓ La méthode analytique: dans cette approche est exploitée la structure linguistique du mot. En effet dans ce cas les unités de base sont les phonèmes, les syllabes, et les demi-syllabes, la reconnaissance dans cette méthode passe par la segmentation du signal de la parole en unité de décision, puis par l'identification de ces unités en utilisant des méthodes de

reconnaissance de formes. Cette méthode est beaucoup mieux adaptée pour les systèmes à grands vocabulaires et pour la parole continue.

III.2 Reconnaissance de type acoustico-phonétique

Cette approche se base sur l'hypothèse que le langage humain est constitué d'une quantité finie d'unités phonétiques, et que ces unités peuvent être identifiées au travers de certaines propriétés visibles du signal ou de son spectre [28]

On peut visualiser le déroulement global du décodage acoustico-phonétique de la parole au moyen de la figure III.1.

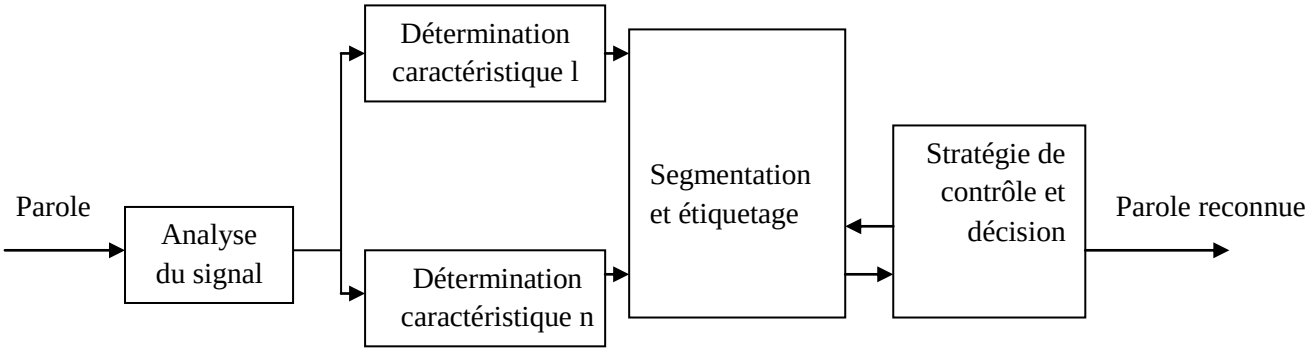


Figure III.1. Structure d'un système de reconnaissance acoustico-phonétique

Un système de reconnaissance de la parole par décodage acoustico-phonétique se compose de trois étapes après la phase de la représentation des caractéristiques spectrales du signal, que nous allons décrire par la suite.

III.2.1 extraction des paramètres pertinents

L'objectif de cette étape est de convertir les caractéristiques spectrales du signal en un jeu d'indices acoustiques décrivant les propriétés physiques des sons. Ces indices peuvent être binaires, et donc notifiés par leur présence ou leur absence, comme :

- le voisement;
- la nasalité;
- la frication.

Il reste alors à combiner judicieusement ces indices pour distinguer les unités phonétiques du mot. C'est le rôle de la phase de segmentation et d'étiquetage.

III.2.2 Segmentation et étiquetage des sons de parole

La segmentation consiste à découper le signal de parole en unités clairement définies, en se basant sur l'évolution temporelle des indices déterminés à l'étape précédente. L'objectif est donc de trouver des régions stables, à l'intérieur desquelles les valeurs des indices ne varient pas, ou peu.

L'étiquetage phonétique d'un segment se fait ensuite en combinant les valeurs des indices déterminés auparavant, afin d'identifier le son parmi l'ensemble des phonèmes ou unités acoustiques du vocabulaire considéré. Si la grande diversité des phonèmes peut rendre cette tâche très ardue [5].

III.2.3 Décision et stratégie de contrôle

La stratégie de contrôle consiste à exploiter les contraintes lexicales pour revenir éventuellement sur la phase de segmentation et réajuster le découpage.

III.3 les distances dans l'espace acoustique

La reconnaissance de la parole est effectuée normalement au niveau acoustique. Le spectre du mot à reconnaitre est comparé ç ceux d'un ensemble de mots appelés mots de référence. Il est pertinent de se demander comment mesurer le degré de similarité entre une occurrence et une autre lors d'un processus de décision. En d'autres termes il faut établir une distance ou une mesure de dissemblance entre ces deux occurrences. Cependant il faut s'assurer de réduire au minimum la sensibilité de cette distance aux fluctuations des débits d'élocution.

III.3.1La mesure de distorsion

Considérons un ensemble quelconque E de points. Nous dirons que E est un espace métrique réel d'il existe une fonction appelée distance, notée :

$\mathrm{d} : E \times E \rightarrow \mathrm{R}$,

Vérifiant les quartes propriétés suivantes :

- Séparabilité : $\forall (a,b) \in \mathrm{E}^2, \mathrm{a} \neq \mathrm{b} \Rightarrow \mathrm{d(a,b)} > 0$,
- Réflexivité : $\forall\, a \in E, d(a,a) = 0$,
- Symétrie : $\forall (a,b) \in \mathrm{E}^2, d(a,b) = d(b,a)$,
- Inégalité triangulaire : $\forall (a,b,c) \in \mathrm{E}^2, \mathrm{d(a,b)} \leq \mathrm{d(a,c)} + \mathrm{d(c,b)}$.

En parole ces conditions ne sont pas toutes satisfaites (c'est le cas par exemple de la symétrie), on parle plutôt de la mesure de dissemblance ou de mesure de distorsion.

Les distances utilisées pour comparer deux occurrences sont étroitement liées aux types de paramètres utilisés. Une définition particulière de la distance entre deux spectres doit être [2] :

- Significative sur le plan acoustique.
- Formalisable d'une façon efficiente sur le plan mathématique.
- Définie dans un espace de paramètres judicieusement choisi.

III.3.2 La distance Euclidienne

Pour l'analyse spectrale ou cepstrale, le choix se porte généralement sur les distances associées à la norme dite de Holder [2], pour des vecteurs à K composantes, cette norme est :

$$d_p(a,b) = [\sum_{k=1}^{k} |a_k - b_k|^p]^{\frac{1}{p}}, p = 1 \text{ à } \infty \quad (3.1)$$

Où *a* est un vecteur de paramètres du mot de référence, et *b* un vecteur de paramètres du mot à reconnaitre.

Pour p=2, cette distance est connue sous le nom de la distance euclidienne, qu'on peut utiliser dans le domaine spectral ou cepstral.

III.3.3 La distance d'Itakura

La distance dItakura est utilisée pour comparer deux vecteurs a et b de (p+1) coefficients de prédiction linéaire chacun, p étant l'ordre de la prédiction. Cette distance est définie par [2] :

$$d_t = \ln\left[\frac{aRa^T}{bRb^T}\right] \quad (3.2)$$

Où *a* est vecteur de référence et R est la matrice$(p+1) \times (p+1)$ des coefficients d'autocorrelation évalués sur le segment de signal correspondant à b, et a^T est le vecteur colonne transposé de a. pour le numérateur il s'agit de l'énergie résiduelle on peut l'évaluer par :

$$aRa^T = r(0)r_a(0) + 2\sum_{n=1}^{p} r(k)r_a(k) \quad (3.3)$$

Où les r(k) sont les coefficients d'autocorrélation sur le segment du signal correspondant à b, et $r_a(k)$ sont les coefficients d'autocorrélation correspondant au vecteur a.

Une autre variante de la distance d'ITAKURA est la distance appelée rapport de vraisemblance (Likelihod-Ratio) dont la forme est :

$$d_{LR} = \frac{aRa^T}{bRb^T} - 1 \quad (3.4)$$

III.3.4 La distance cepstrale

Soit deux vecteurs C_t et C_r qui contiennent respectivement les coefficients cepstraux d'un segment du mot de référence et d'un segment du mot à reconnaitre. La distance cepstrale, d_{CEP} est la distance euclidienne entre ces deux vecteurs, elle est définie par :

$$d_{CEP} = [c_t(0) - c_r(0)]^2 + 2\sum_{k=1}^{\infty}[c_t(k) - c_r(k)]^2 \quad (3.5)$$

En pratique, on ne prend pas en considération le premier terme de la distance.

La distance cepstrale est généralement tronquée, elle est évaluée le long d'un nombre fini de coefficients typiquement 10à30 [28]. Cependant ce nombre ne doit pas être inférieur à l'ordre de prédiction p si les spectres sont issus d'un modèle tous pôles d'ordre p.

La distance cepstrale sera alors :

$$d_{CEP} = \sum_{k=1}^{L}[c_t(k) - c_r(k)]^2 \quad (3.6)$$

Avec L est le nombre de coefficients le long desquelles la distance est calculée.

Remarque :

Les coefficients cepstraux dans l'équation (3.6) peuvent être des coefficients $MFCC_s$ ou des coefficients $LPCC_s$.

III.3.5 La distance cepstrale pondérée

Une autre variété de la distance cepstrale est la distance cepstrale pondérée, il a été démontré [21] que cette pondération permet d'améliorer considérablement le taux de reconnaissance, l'idée est d'ajouter une pondération w(k) dans la formule (3.6), la forme générale de la distance sera alors :

$$d_{WCEP} = \sum_{k=1}^{L}(w(k)(c_t(k) - c_r(k)))^2 \quad (3.7)$$

Plusieurs méthodes de pondération ont été introduites par des chercheurs, dans ce qui suit nous présenterons quelques unes.

III.3.5.1 La pondération par la fréquence k

Cette pondération a été utilisée par Hanson et Wakita [19], elle permet de normaliser la contribution de chaque coefficient cepstral, elle est définie par :

$$W(k)=k$$

$$d_{WCEP} = \sum_{k=1}^{L} k^2[c_t(k) - c_r(k)]^2 \quad (3.8)$$

III.3.5.2 La pondération par variance

Cette méthode de pondération à été utilisée par Tokhura [21] dans "la reconnaissance multi-locuteurs des chiffres isolés en utilisant les distances pondérées ", son origine est la distance suivante :

$$d_{MCEP} = (c_t - c_r)V^{-1}(c_t - c_r)^T \quad (3.9)$$

Cette distance (3.9) connue sous le nom de distance de Mahanalobis [19], s'applique difficilement en reconnaissance de la parole, entre autre à cause de multiplication de la matrice de la covariance V, alors on l'a remplacée par la distance (3.10) ou juste les éléments de la diagonale de la matrice de la covariance sont utilisés au lieu de la matrice V en entier :

$$w(k) = \frac{1}{V_{kk}}$$

$$d_{WCEP} = \sum_{k=1}^{L} \left(\frac{1}{V_{kk}}\right)^2 [c_t(k) - c_r(k)]^2 \quad (3.10)$$

Ou V_{kk} est le $k^{ème}$ élément de la diagonale de la matrice de la covariance.

III.3.5.3 La pondération par filtrage passe-bande (bandpass liftering)

Cette pondération a été introduite par Juang et al [29], son rôle est de réduire la variabilité sepctrale en raison de la sensibilité du banc de filtres à tout changement de la fréquence du fondamental dans le cas des coefficients MFCCs ou les effets indésirables sur les spectres causés par la représentation par un modèle tous pôles, dans le cas des coefficients LPCCs [29], trois types de fenêtres ont été avancées comme solution :

Type 1

$$w_1(k) = \begin{cases} 1, & k = 1, \ 2, \dots, L \\ 0, & ailleurs. \end{cases} \quad (3.11)$$

Type 2

$$w_2(k) = \begin{cases} 1 + h \cdot (k-1)/(L-1), k = 1, \ 2, \dots, L \\ 0, ailleurs. \end{cases} \quad (3.12)$$

Type 3

$$w_3(k) = \begin{cases} 1 + h.\sin(k\pi/L), & k = 1, \ 2, \dots, L \\ 0, & ailleurs. \end{cases} \quad (3.13)$$

Juang et al ont retenu la dernière forme (3.13) comme meilleure choix pour les applications de reconnaissance et ont suggéré comme valeur pour la variable *h* la valeur L/2, L étant le nombre des coefficients cepstraux.

III.4 Reconnaissance basée sur l'alignement temporel des formes

III.4.1 Présentation du problème

La tâche de reconnaissance consiste ici à identifier un mot prononcé par un locuteur en déterminant sa distance vis-à-vis de l'ensemble des mots du vocabulaire de référence. Étant donné qu'un mot est représenté par une succession de vecteurs acoustiques court-terme, l'opération consiste donc à mesurer les distances locales entre chacun des vecteurs spectraux, x, et y, des deux formes à comparer, X et Y, de longueur T, la distance totale entre ces deux mots étant alors obtenue par accumulation de toutes ces distances locales, comme présenté par la formule 3.14.

$$D(X,Y) = \sum_{i=1}^{\tau} d(x_i, y_i) \qquad (3.14)$$

Toutefois, et c'est bien là l'une des complexités majeures de la reconnaissance de la parole, un vocabulaire est généralement constitué de mots de tailles différentes, et un même mot peut être prononcé à des vitesses d'élocution variables, engendrant ainsi une transformation non linéaire de l'échelle temporelle. Afin de rendre la mesure de distance indépendante des fluctuations du rythme de prononciation et de la vitesse d'élocution, les chercheurs ont mis en œuvre des algorithmes de comparaison dynamique.

III.4.2 La programmation dynamique

L'orsqu'un locuteur, même entraîné, répète plusieurs fois une phrase ou un mot, il ne peut éviter les variations du rythme de prononciation ou de la vitesse d'élocution, ces variations entraînent des transformations non linéaires dans le temps du signal acoustique, ce qui fait qu'on ne pourra comparer directement point à point (matching) deux formes acoustiques sans correction temporelle au préalable.

Pour établir une meilleure correspondance entre les axes temporels des deux mots, en même temps que leurs comparaisons, on utilise une technique appelée technique d'alignement temporel dynamique ou DTW [2] pour *Dynamic Time Warping*. La notion de programmation dynamique a été introduite par Bellman en 1954, mais utilisée pour la première fois en reconnaissance vocale par Vintsyuk en 1968. L'algorithme consiste alors à calculer de façon récursive la distance minimale accumulée pour chaque point (i,j) en tenant compte de certaines contraintes locales sur la façon avec laquelle le chemin optimum atteint ce point. Pour plus de clarté, la figure 3.8 présente la comparaison de deux mots X et Y, de tailles respectives Tx et Ty (en terme de nombre de trames). Chaque point visible sur cette figure correspond à une distance locale $d(x_i,y_j)$.

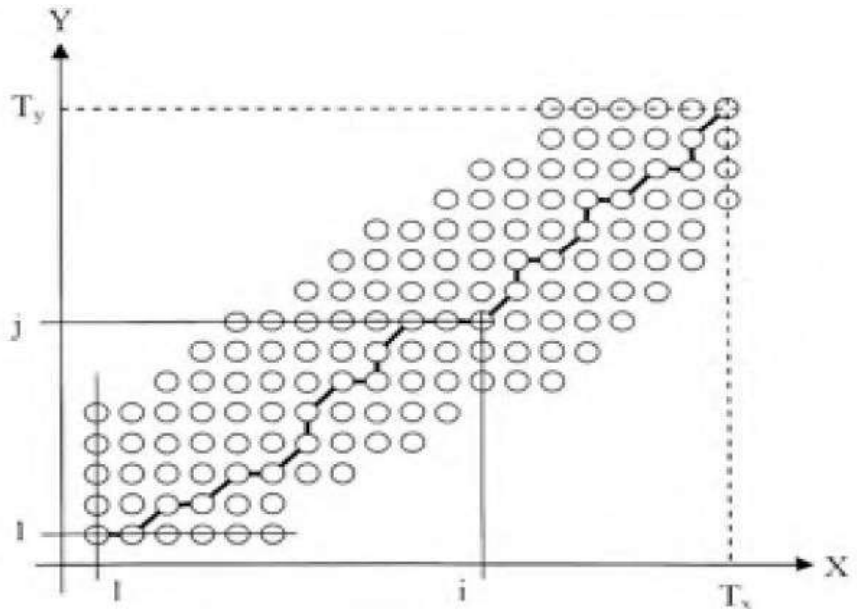


Figure III.2. Alignement temporel entre les allocutions X et Y

La comparaison est donc basée sur la recherche du chemin optimum w de longueur K, défini par la formule 3.12, et dont la distance associée est donnée par la formule 3.15 [21].

$$\begin{cases} w: [i(k), j(k)]; & k = 1,2, \dots, K \\ i(1) = 1 & j(1) = 1 \\ i(K) = T_x, & j(K) = T_y \end{cases} \tag{3.15}$$

$$D_w(X,Y) = \frac{\sum_{k=1}^{K} d(x_{i(k)}, y_{j(k)}) \times g(k)}{N(g)} \tag{3.16}$$

Où g(k) est un coefficient de pondération, et N(g) un facteur de normalisation généralement fixé selon la formule 3.17.

$$N(g) = T_X + T_y \tag{3.17}$$

La recherche du chemin optimum, c'est-à-dire de distance accumulée minimale, se fait donc au moyen de contraintes locales destinées à réduire le nombre de calculs et à prendre en compte le caractère temporel. la distance accumulée entre le point (1 , 1) et le point (i , j), notée D(i,j), sera déterminée soit par:

- Contrainte de type (a):

$$\widetilde{D}(i,j) = d(i,j) + min\begin{cases} \widetilde{D}(i-1,j) \\ \widetilde{D}(i-1,j-1) + d(i,j) \\ \widetilde{D}(i,j-1) \end{cases} \quad (3.18)$$

- Contrainte de type (b):

$$\widetilde{D}(i,j) = d(i,j) + \min[\widetilde{D}(i-1,j), \widetilde{D}(i-1,j-1), \widetilde{D}(i-1,j-2) \quad (3.19)$$

- Contrainte de type (c):

$$\widetilde{D}(i,j) = d(i,j) + min\begin{cases} \widetilde{D}(i-2,j-1) + 2 \cdot d(i-1,j) \\ \widetilde{D}(i-1,j-1) + d(i,j) \\ \widetilde{D}(i-1,j-2) + 2 \cdot d(i,j-1) \end{cases} \quad (3.20)$$

Où d(i, j) la distance locale, et D la distance globale, ou l'accumulation des distances locales.

Ces contraintes peuvent être présentées par la figure 3.3:

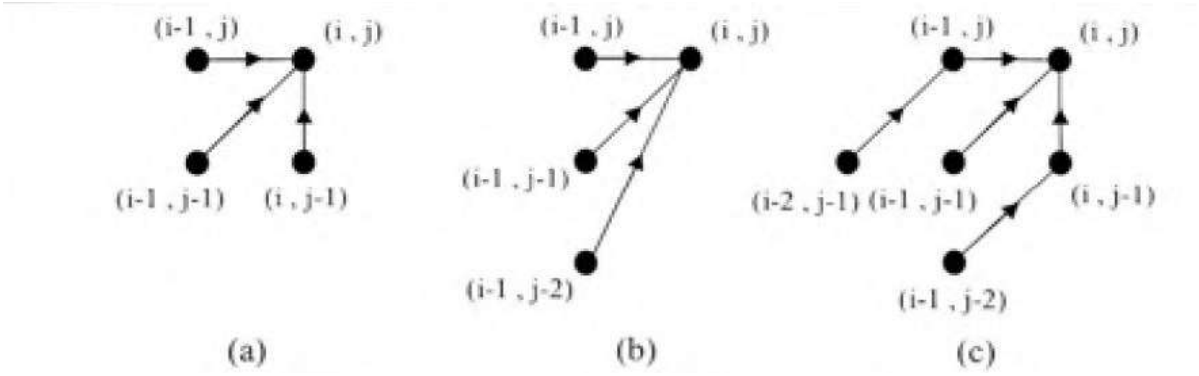


Figure III.3 Contraintes locales sur le chemin optimal

L'algorithme général de programmation dynamique peut donc être réalisé selon le schéma bloc présenté à la figure III.4 (r étant l'espace de recherche accordé autour de la diagonale, dans le but de réduire le nombre de calculs).

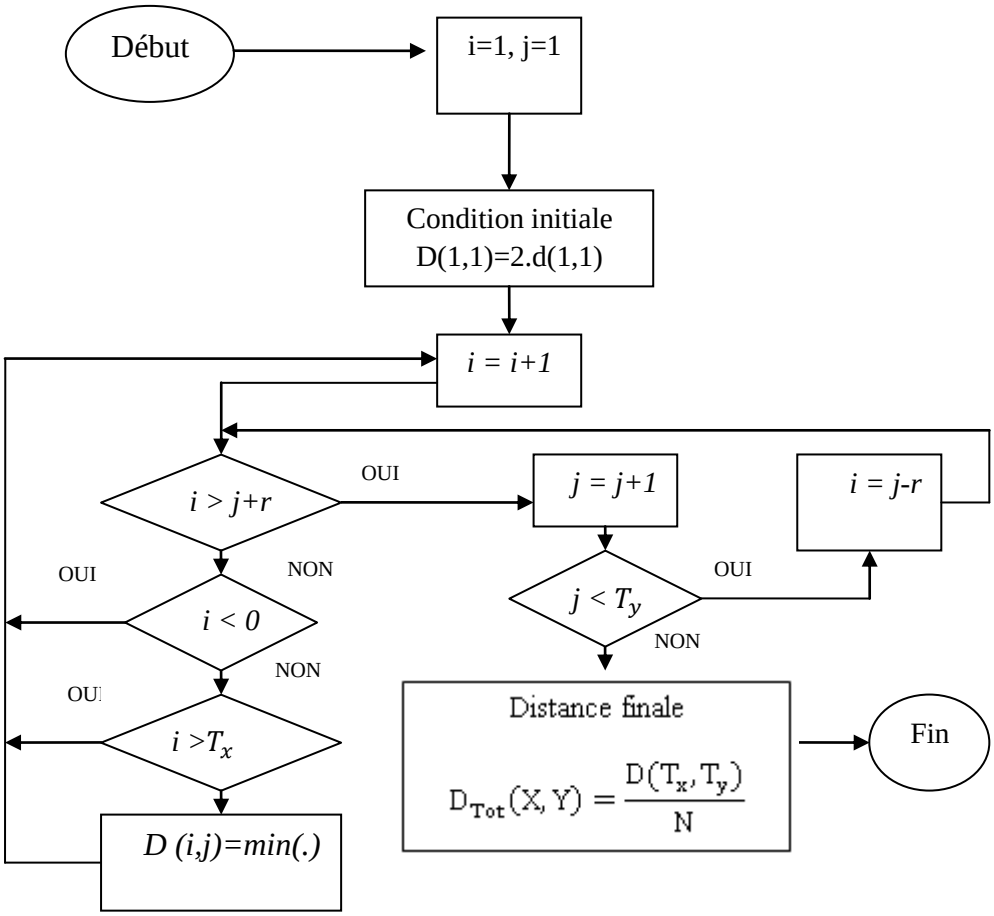


Figure III.4. Algorithme de programmation dynamique

III.4.3 Les modèles MMCs:

III.4.3.1 concept général

Les approches statistiques et les modèles probabilistes sont très utilisés, de nos jours, dans les systèmes de reconnaissance automatique de la parole. Ces approches, notamment celles basés sur les Modèles de Markov Cachés (ou HMM pour Hidden Markov Model), ont atteint des performances remarquables avec des vocabulaires de plus en plus importants et une robustesse au bruit et à la variabilité des locuteurs de plus en plus grande [28].

Pour la modélisation statistique acoustique, les modèles de Markov cachés (HMM) sont aujourd'hui utilisés dans un très grand nombre des systèmes de reconnaissance automatique de la parole. Chaque unité de parole est modélisée par un HMM. Dans le cas de petits lexiques, ces unités de parole peuvent être les mots. Dans le cas de grands lexiques, on préférera souvent utiliser des modèles de phonèmes (ou polyphones), ce qui limitera le nombre de paramètres à estimer. Dans ce dernier cas, lors de la reconnaissance, les mots

seront construits (dynamiquement) en termes de séquences de phonèmes et les phrases en termes de séquences de mots.

Les HMMs supposent que le phénomène modélisé est un processus aléatoire et inobservable qui se manifeste par des émissions elles-mêmes aléatoires. Cette approche markovienne offre une flexibilité séduisante de modélisation pour un phénomène aussi complexe que la parole.

III.4.3.2 présentation des modèles de Markov cachés:

Les modèles de Markov cachés sont apparus dans la problématique de la reconnaissance automatique de la parole dans les années 70 [30]. L'idée sous-jacente est que la parole peut être caractérisée par un processus aléatoire dont les paramètres peuvent être estimés d'une manière appropriée. Les modèles HMM ont prouvé leur efficacité dans de nombreux domaines de la reconnaissance automatique de la parole. Cependant, les modèles ont été améliorés au fil des recherches pour repousser leurs limites intrinsèques, particulièrement en intégrant des notions de corrélation entre trames et de modélisation de trajectoires.

Un modèle de Markov discret est un automate stochastique à nombre d'états fini N [28]. Un processus aléatoire se déplace d'état en état à chaque instant et on note qt l'état atteint par le processus à l'instant t.

Dans le cas des modèles de Markov discrets cachés, l'état réel qt n'est pas directement observable mais le processus émet un symbole discret après chaque changement d'état. Les observations ne sont plus univoquement liées à une seule classe bien déterminée mais sont donc des fonctions statistiques de ces classes qui ne sont plus observées directement. Ou encore, les états du modèle ne sont plus observés directement à partir des observations qui sont supposées être produites par ces états mais à travers une fonction statistique différente pour chaque classe. On a donc un processus doublement stochastique : modèle stochastique relatif au modèle de Markov sous-jacent et celui décrivant la relation entre les classes (états) et les observations. La figure III.5 présente un modèle de Markov caché ergodique, où toutes les transitions entre tous les états sont autorisées.

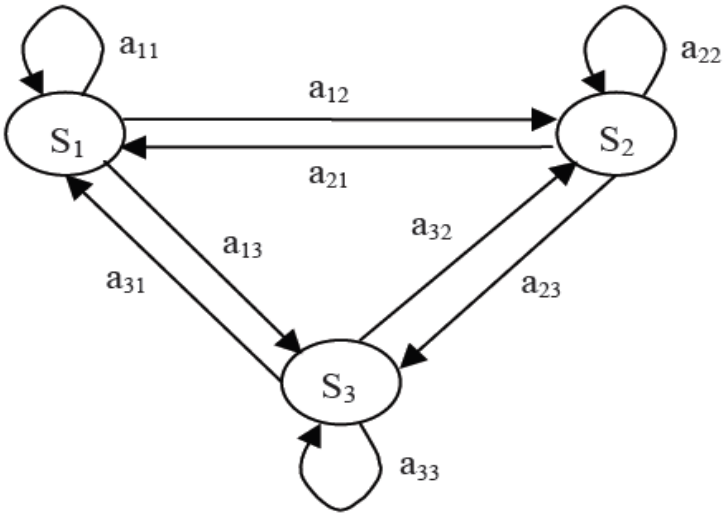


Figure III.5 Exemple de modèle de Markov caché ergodique

Concrètement, un modèle de Markov caché HMM est représenté par Φ = (*A*, *B*, p) qui est caractérisé par les éléments suivants :

- $S = \{s_1, s_2, \ldots, s_N\}$: un ensemble des états du modèle avec *N* le nombre d'états. On note *qt* l'état à l'instant *t* ;
- $O = \{o_1, o_2, \ldots, o_M\}$: un alphabet des observations avec *M* nombre fini de symboles

D'observation par état. Les symboles d'observation correspondent à chaque sortie physique du système réel qu'on modélise. On note x_t l'observation à l'instant *t* ;

- $A = \{a_{ij}\}$: une matrice des probabilités de transition entre états, dont a_{ij} est la probabilité de transition de l'état *i* à l'état *j*. On a :

$$a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i),\ 1 \le i, j \le N$$

$$\sum_{j=1}^{N} a_{ij} = 1, 1 \le i \le N \qquad (3.21)$$

- $B = \{b_i(k)\}$: une matrice des probabilités d'émission des observations dans chaque état, dont bi(k) est la probabilité d'émission de l'observation ok dans l'état si. On a :

$$b_i(k) = P(x_t = o_k \mid q_t = s_i),\ 1 \le i, j \le N \qquad (3.22)$$

- $p = \{p_i\}$: une matrice de distribution de l'état initial. On a :

$$\pi_i = P(q_0 = s_i),\ 1 \le i \le N \qquad (3.23)$$

$$\sum_{i=1}^{N} \pi_i = 1$$

Pour la modélisation acoustique, le modèle souvent utilisé est donc le modèle HMM gauche-droit (ou de Bakis), illustré par la figure 3.6, dans lequel on ne peut pas revenir à un état précédent.

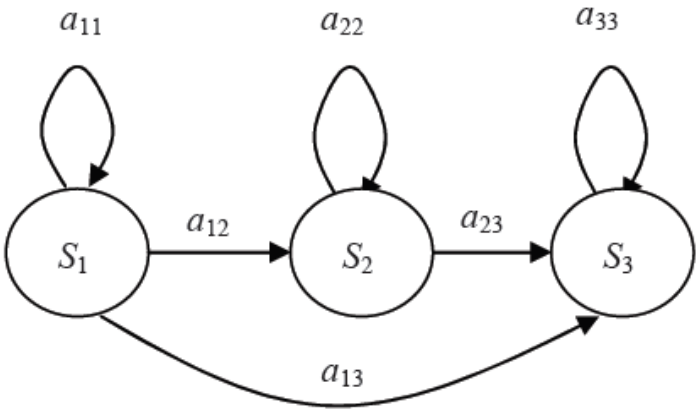


Figure III.6. Exemple de HMM à 3 états gauche-droit

III.4.3.3 Apprentissage des modèles

Le nombre d'états d'un modèle est généralement choisi de façon empirique, après une longue étude du vocabulaire concerné, tandis que les paramètres ajj et bi(k) sont obtenus au cours d'une phase d'apprentissage, à partir d'un nombre important d'énoncés du même mot. Cette phase d'entraînement, qui permet donc d'ajuster les paramètres du modèle en maximisant les probabilités d'observation jusqu'à stabilisation, est réalisée en utilisant une procédure itérative telle que l'algorithme de Baum-Welch [29].

III.4.4 La quantification vectorielle

La quantification vectorielle est une opération qui consiste en une partition de l'espace des vecteurs acoustiques en classes (voir figure III.7). Chaque classe est représentée par un vecteur particulier appelé centroïde ou noyau. Ce point représente la distance minimale intra-classe.

Le dictionnaire de prototypes appelé aussi-book qui contient l'ensemble de tous les centroïdes, est obtenu par apprentissage ou entrainement sur un grand ensemble suffisamment représentatif de données. Pour sa construction on fait appel) la méthode de Llyod-Max (ou à une de ses variantes) très connue sous le nom de l'algorithme de k-means.

La quantification vectorielle tel que présentée ci-dessus est la quantification vectorielle statistique. En effet il existe une autre forme de quantification appelée la quantification vectorielle algébrique dont le principe est d'utiliser un code-book qui présente une certaine

structure mathématique. Cependant dans les applications RAP on lui préfère la première qui représente la plus simple interprétation de la quantification vectorielle.

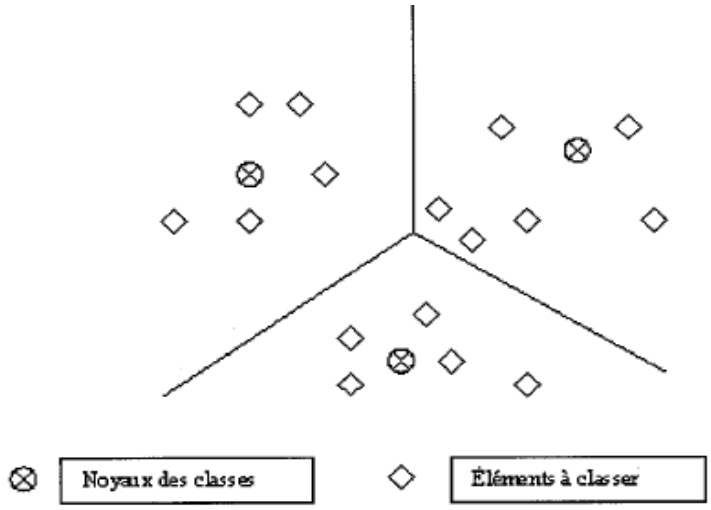


Figure III.7. Partition d'une espace en trois classes

III.4.5 Algorithme de classification

II est un dernier point important que l'on doit considérer lors de la conception d'une méthode de reconnaissance vocale, c'est la gestion du dictionnaire de références. Comme mentionné précédemment, pour être efficaces les méthodes de reconnaissance de forme travaillent à partir d'un large vocabulaire de référence, obtenu au moyen de prononciations répétées de chaque unité de parole. Que ces unités soient des mots entiers, comme pour la programmation dynamique, ou bien des phonèmes, vecteurs spectraux ou autres petits événements acoustiques, comme pour les HMM, plus le système contient de versions d'une même unité à comparer, et plus la reconnaissance a de chances de rencontrer un taux de réussite élevé, notamment dans le cas des systèmes multilocuteurs.

Malheureusement, les limitations techniques imposées par les systèmes, aussi bien en terme de rapidité de calcul que de place mémoire, nous empêchent souvent d'embarquer une trop grande quantité de références. Il est donc important de pouvoir optimiser le contenu du dictionnaire, afin de ne garder qu'un nombre restreint de versions de chaque mot, tout en conservant une grande diversité dans les représentations possibles de ce mot. Cette tâche peut être accomplie au moyen d'algorithmes de classification. Les premiers, apparus à la fin des années 70 [4], nécessitaient une intervention humaine pour guider le processus. C'est pourquoi une nouvelle classe d'algorithmes de classification automatique a été développée, dont les plus utilisés encore de nos jours sont le « Unsuper-vised cliistering Without Aver-aging » (UWA) et le « Modified K-Means » (MK-M) présentés par Wilpon et Rabiner [31].

L'algorithme MK-M, dérivé des méthodes de quantification vectorielle, est présenté par le schéma bloc de la figure III.8.

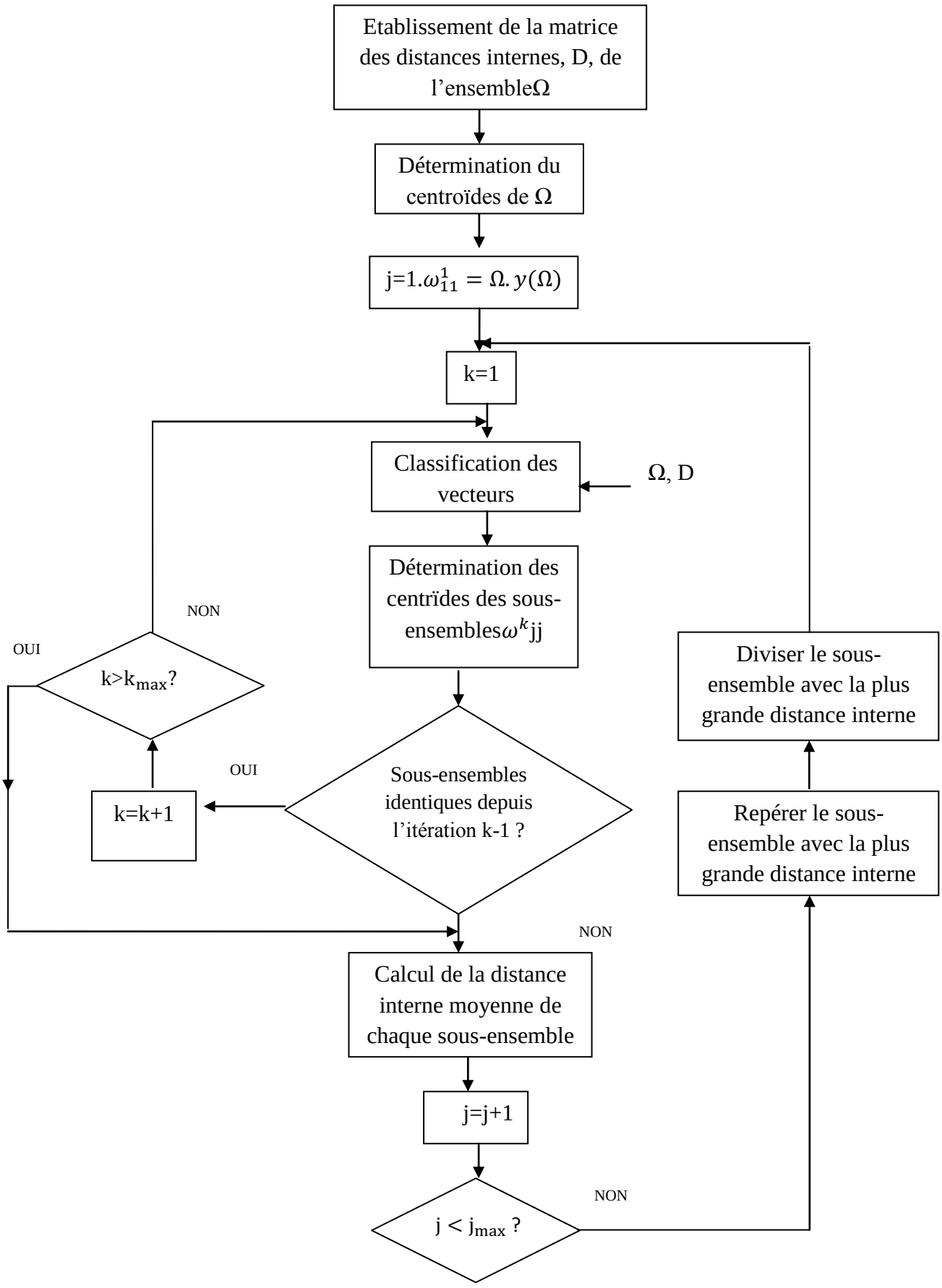


Figure III.8. Schéma bloc de la procédure de classification M-KM.

À partir d'un ensemble de plusieurs références, cet algorithme fournit un nombre de sous ensembles Jmax fixé par l'utilisateur, chacun étant représenté par son centroïde, c'est-à-dire l'élément ayant la distance la plus rapprochée de tous les autres éléments du sous-ensemble.

Les distances entre les éléments peuvent être des distances locales comme celles présentées dans la partie III.3, si l'on souhaite classer des petites unités acoustiques représentées par un simple vecteur de paramètres, ou bien une mesure de dissemblance globale obtenue, par exemple, par DTW si l'on traite avec des mots de parole complets. La figure III.9 schématise le résultat d'une classification d'un ensemble de 38 éléments en 4 sous ensembles, les centroïdes étant repérés par un petit cercle.

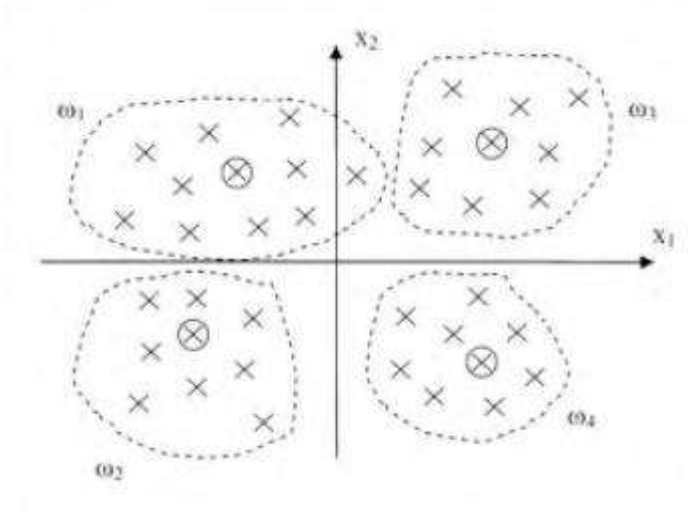


Figure III.9. Classification d'un ensemble de 38 éléments en 4 sous-ensembles.

III.5 Conclusion

En reconnaissance on distingue principalement deux grand méthode : la méthode globale et la méthode analytique, dans la première le mot (ou la phrase) à reconnaitre est comparé dans son ensemble avec des mots (ou des phrases), sans décomposition en entités plus élémentaire, alors que pour la deuxième méthode la reconnaissance est précédée d'une segmentation en unités de base tel que phonèmes, syllabes,....etc.

Pour tout système RAP, deux phases sont nécessaires: une phase d'apprentissage, et une phase de reconnaissance, le but de la première est la construction d'un dictionnaire de référence qui servira de base dans la phase de reconnaissance. Dans certains cas des algorithmes de classification automatique sont utilisé pour réaliser l'apprentissage.

Au niveau acoustique la reconnaissance d'un mot se fait par une succession de comparaisons du spectre du mot à reconnaitre avec ceux des mots de référence, pour ce faire une distance ou une mesure de dissemblance est définie d'une façon judicieuse entre deux

mots. Dans le but d'améliorer les performances de reconnaissance, une pondération adéquate est ajoutée à ces distances.

Dans le chapitre suivant sera décrit notre système de reconnaissance vocal, ainsi que son implémentation à l'aide du système de développement du Texas Instrument.

Chapitre IV

La réalisation et simulation d'un système de reconnaissance vocal

IV.1 Introduction

Notre objectif est de développer une méthode pouvant reconnaître des chiffres, isolés, prononcés par plusieurs locuteurs différents. La grande difficulté provient du fait que cette méthode doit pouvoir être utilisée sur un processeur de traitement du signal, et donc produire de bons résultats tout en satisfaisant aux contraintes de mémoire et de rapidité imposée par un tel système. Nous commencerons par présenter les différents éléments qui composent notre système ainsi que les algorithmes utilisés pour le réaliser, suivra la présentation des résultats de simulations du système de reconnaissance réalisé.

IV.2 Description du système de reconnaissance automatique réalisé

Comme illustré par le schéma bloc (figure IV.1), le processus conduisant à la reconnaissance commence par l'extraction des informations nécessaires au décodage de la parole: à partir d'un découpage en trame, une analyse par prédiction linéaire est effectuée, de laquelle on en déduit les coefficients cepstraux. Un nombre de 12 coefficients est choisi pour représenter une trame de parole, afin de disposer d'une information très précise sur le signal.

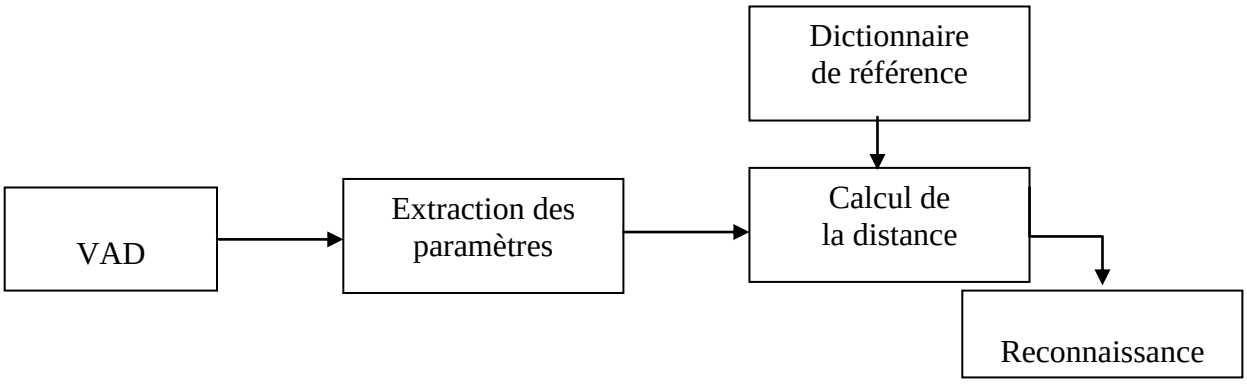


Figure IV.1: Schéma bloc du système de reconnaissance

La première phase concerne la détection d'activité vocale (VAD: Voice Activity Detection), nécessaire pour éliminer les zones de silence entourant les chiffres. Nous utiliserons la méthode travaillant à partir de l'énergie et du taux de passage par zéro. Par la suite les paramètres LPCs, ou LPCCs,ou MFCCs seront extraits pour chaque trame. Enfin le processus de la reconnaissance (DTW) sera appliqué.

IV.2.1 La détection d'activité vocale (VAD)

IV.2.1.1 introduction:

L'objectif de la détection d'activité vocale est de séparer les zones de parole des autres zones qui sont sans intérêt pour le traitement de la parole. Pour des mots prononcés dans des conditions idéales et correctement articulés, la détection des zones de parole est aisée, mais ce n'est malheureusement pas toujours le cas dans la pratique. Entre les problèmes attribués au locuteur, tels que les bits liés à l'ouverture et la fermeture de la bouche ou à la respiration, les bruits, ou la distorsion introduite parle canal de transmission, la dégradation du signal est parfois trop grande pour détecter idéalement les limites des zones de parole. Ce problème a été l'objet de nombreuses recherches, qui ont abouti à diverses approches. Nous pouvons ainsi mentionner l'algorithme de la norme G.729B de l'ITU [27], très utilisé dans les applications de téléphonie mobile, en raison de sa capacité à détecter les zones de parole en temps réel, à partir des coefficients LPC. Mais les algorithmes les plus couramment utilisés en reconnaissance de la parole utilisent plutôt des segments de parole entiers, dont ils éliminent les zones non désirables entourant la parole. C'est le cas de l'algorithme basé sur l'énergie court-terme et le taux de passage par zéro, présenté par Rabiner et Sambur en1975, et dont le déroulement général est présenté sur la figure IV.2.

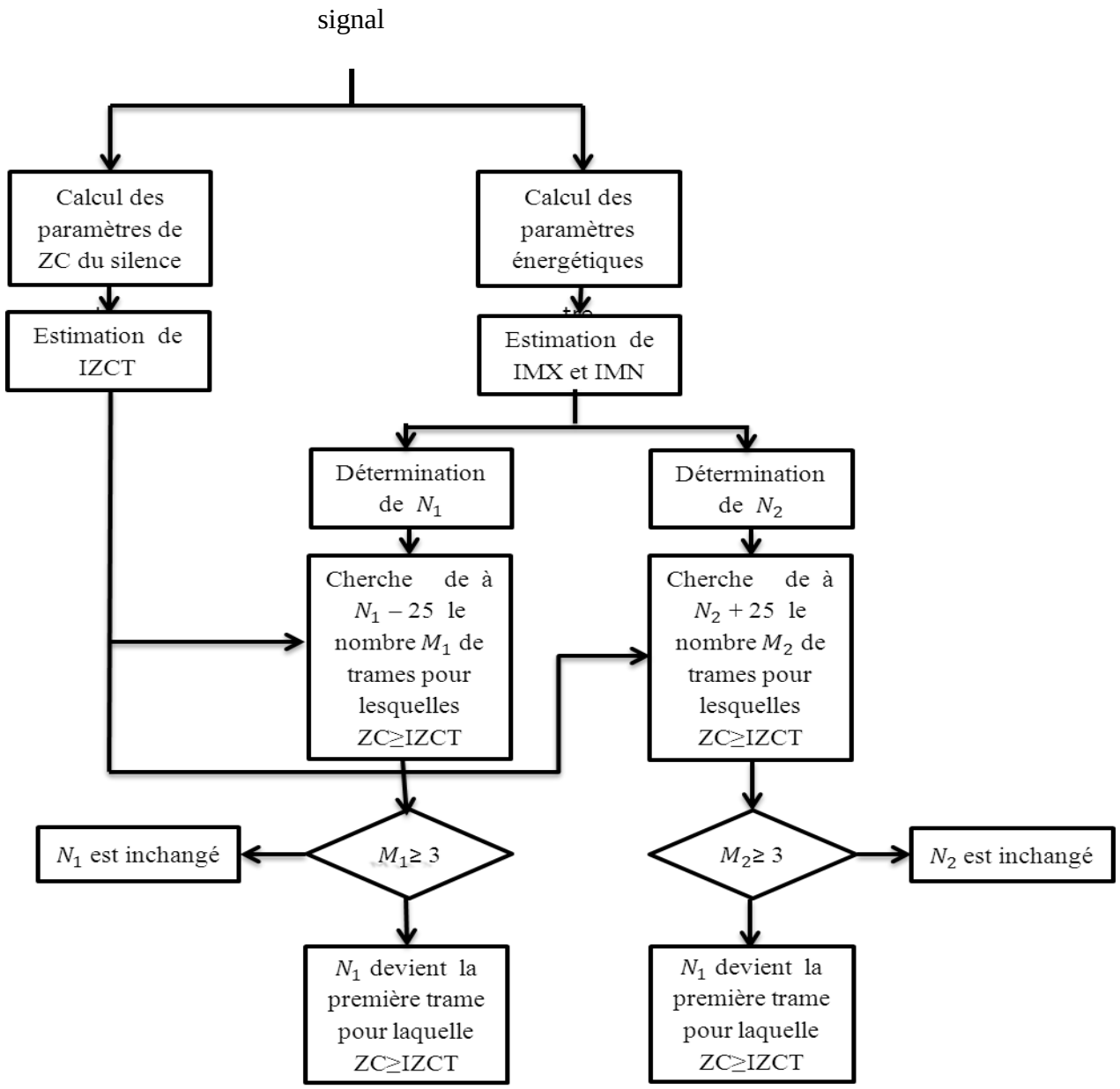


Figure IV.2: Déroulement de la détection d'activité selon Rabiner et Sambur[28]

IV.2.1.2 Description de l'algorithme utilisé

On peut résumer cet algorithme par les étapes suivantes:

a) Calcul de l'énergie et du taux de passage par zéro

Les deux paramètres, le taux de passage par zéro et l'énergie sont calculés chaque 10 ms sur des fenêtres de 10 ms.

b) Calcul des statistiques de silence [28]

Dans cette étape, on prend l'hypothèse que pendant les premières 100ms du signal il y'a absence totale de parole, alors on s'en sert pour calculer les statistiques du silence, qui vont servir pour la détermination du seuil du taux de passage par zéro *IZCT*, et les seuils haut et bas de l'énergie *ITU* et *ITL*, ces paramètres sont :

IF=25 par 10ms

I1=0.03*(*IMX* - *IMN*)+*IMN*

*I*2=4**IMN*

ITL=min(*I*1,*I*2)

ITU=5**ITL*

$IZCT=\min(IF, \overline{IZC} + 2\sigma_{IZC})$

avec :

IZCT: le seuil du taux de passage par zéro.

$\overline{IZC}$: la moyenne du taux de passage par zéro.

IMX: le maximum de l'énergie sur toute la durée de signal.

IMN: la moyenne de l'énergie du silence.

ITL: le seuil bas de l'énergie.

ITU: le seuil haut de l'énergie.

IF : étant une valeur fixée par l'utilisateur.

σ_{IZC}: l'écart-type du taux de passage par zero.

c) Détermination des points début et fin de la parole

Après la détermination des différents paramètres, on commence la recherche des points des points début et fin de la parole, cette partie de l'algorithme se fait en deux étapes:

✓ **1ère étape:**

Dans cette phase on travaille seulement avec l'énergie comme le montre la figure IV.3. Ainsi l'algorithme commence par la recherche du début de l'intervalle jusqu'à ce que le seuil inférieur de l'énergie(*ITL*) soit dépassé. Ce point alors est étiqueté le début de la parole, à moins que le niveau d'énergie chute en dessous de ce seuil(*ITL*) avant de dépasser le seuil supérieur (*ITU*). Si ceci se produit, un nouveau point sera obtenu en trouvant le premier point auquel l'énergie excède le seuil inférieur puis excède le seuil supérieur. De la même manière le point final est détecté.

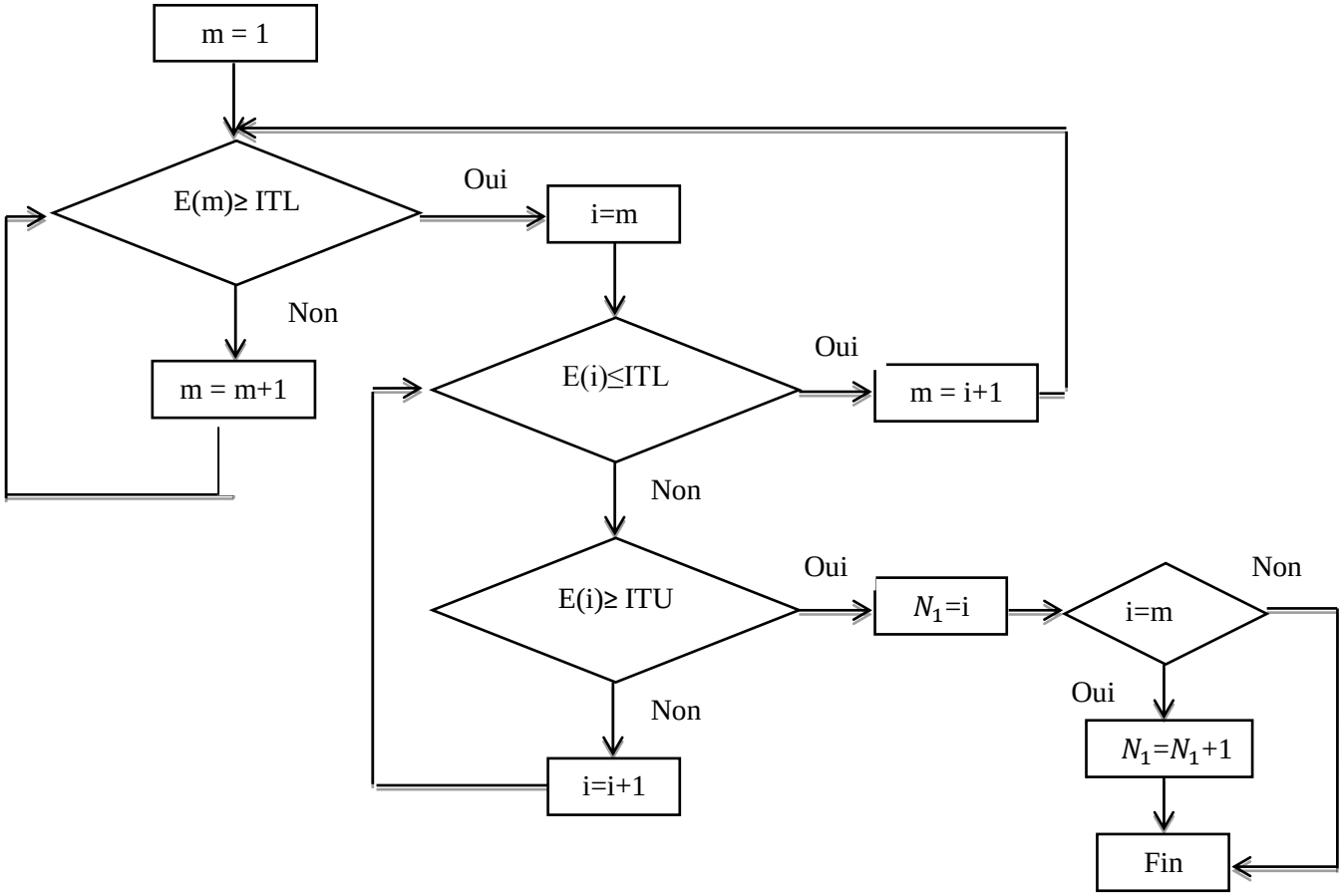


Figure IV.3: Détection énergétique de la trame de début de la zone d'activité vocale.[28]

✓ **2ème étape**

L'utilisation de l'énergie toute seule n'est pas toujours suffisante pour différencier la parole du silence, on peut supposer qu'une partie de la parole puisse être en dehors de l'intervalle trouvé en utilisant seulement les seuils d'énergie. Cet algorithme propose alors une vérification à l'aide du taux de passage par zéro pour confirmer l'intervalle trouvé où dans le cas contraire retrouver les nouveaux points début et fin de la parole.

L'algorithme examine alors les 250 ms précédant le point du début et compte le nombre d'intervalles où le taux de passage par zéro excède le seuil(*IZCT*). Si ce nombre est supérieur ou égal à trois, le point du début est placé au premier point où le seuil a été dépassé, sinon, le point de début est gardé. De la même manière le point final est recherché en effectuant la vérification sur les 250ms suivant le point final trouvé dans la 1ère étape.

IV.2.1.3L'utilité de la VAD:

Comme mentionné auparavant, la « VAD » est une étape très importante en vue d'obtenir les résultats les plus satisfaisants possibles, et de *réduire la quantité de calculs* et *la place mémoire*

utilisés pour la procédure de reconnaissance. La figure IV.4 présente bien ce problème, le chiffre 6 (pris du corpus TI-DIGIT) est entouré de larges zones de silence non nécessaire à la reconnaissance.

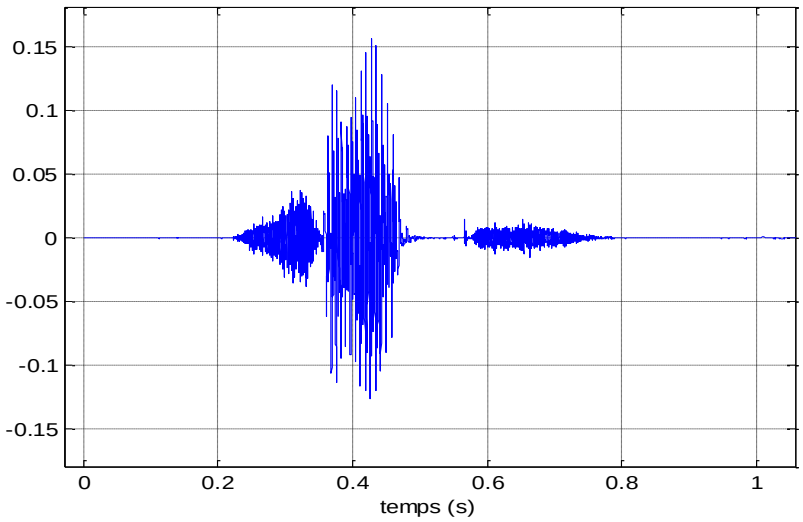


Figure IV.4: Chiffre « 6 » non isolé des zones de silence

Le chiffre« 6 » possède aussi une zone de silence intérieure qu'il est nécessaire de conserver afin de ne pas couper le chiffre en deux. La « VAD » effectue la détection à partir des limites extrêmes du signal, en s'arrêtant à la première zone de parole rencontrée, sans se soucier des éventuelles zones de silence à l'intérieur des frontières obtenues.

IV.2.2 Extraction des paramètres

Pour notre application il est important d'opter pour des paramètres pertinents, discriminants et robustes. Les coefficients cepstraux permettent un meilleur lissage du spectre vocal et sont fiables pour la reconnaissance. Pour cela, et pour une étude comparative, nous avons choisi les trois types de coefficients: les coefficients LPCs, les coefficients LPCCs obtenus grâce à une recursion depuis les coefficients LPC, et les MFCCs obtenue par l'utilisation d'une bande de filtre MEL et DCT comme il est décrit au deuxième chapitre.

IV.2.3 Le dictionnaire de référence (la phase d'apprentissage)

L'objectif de ce mémoire est la reconnaissance de chiffres isolés prononcés en anglais, allant de 0 à 9, où le zéro ayant deux prononciations possibles 'oh' et 'zero'.Pour le processus de reconnaissance, un dictionnaire est nécessaire afin de pouvoir comparer une forme inconnue avec l'ensemble des références, la base de données utilisée pour sa construction sera détaillée ultérieurement.

IV.2.4 Reconnaissance des chiffres

La dernière phase de notre système de reconnaissance vocale consiste à reconnaître le chiffre à partir des opérations décrites jusqu'ici. Le chapitre 3 nous avait permis de présenter diverses méthodes remplissant ce rôle: les modèles de Markov cachés, notamment, qui utilisent un modèle constitué d'état pour reconstruire les mots au moyen de fonctions probabilistes: ou même la DTW, dont le déroulement s'apparente quelque peu à une reconstitution continue (i.e. trame par trame) du chiffre suivant les formes des références. Ces deux méthodes ont pour objectif de repérer la forme ou le modèle se rapprochant le plus du chiffre à reconnaître. Toutes les expériences ont été réalisées avec un système de reconnaissance basé sur la DTW. La distance utilisée pour ces tests est la distance cepstrale déjà mentionnée dans le chapitre 3. Cette dernière a été choisie pour sa simplicité bien qu'elle a montré qu'elle n'est pas la plus performante [19].

IV.3 Corpus utilisé

Les données utilisées pour constituer nos références (dictionnaire) et les occurrences de test proviennent du corpus TI-DIGIT. Cette base de données a été construite en 1982, pour les besoins des laboratoires de « Texas Instruments Incorporated [33] ». Le choix de cette base permet de comparer nos résultats obtenus avec d'autres travaux. Le signal est échantillonné à 20KHzdans un environnement calme(le signal n'est pas bruité).

Ce corpus anglais est composé d'enregistrements de 225 locuteurs (hommes, femmes, enfants) répartis en deux sous-ensembles: un ensemble d'apprentissage (112 locuteurs) et un de test (113 locuteurs). Chaque locuteur a prononcé deux fois chacun des onze chiffres (de 1 à 9, plus "oh" et "zero"). Le sous-ensemble d'apprentissage (respectivement de test) contient 2464 occurrences de chiffres (respectivement, 2486 occurrences de chiffres). Il est à préciser que les locuteurs ont été sélectionnés de divers états des USA, afin de fournir une bonne représentation de l'ensemble des prononciations possibles à travers les différents accents.L'objectif d'une telle base de données est de fournir suffisamment d'échantillons de parole pour pouvoir évaluer les performances des systèmes de reconnaissance vocale multi locuteurs.

IV.4. Contexte expérimental et résultats de simulations

IV.4.1 Cas d'un système multi-locuteurs dépendants

La méthode développée pour ce mémoire est conçue pour fonctionner sur des chiffres isolés, puisque c'est de là que découle majoritairement la réussite de la reconnaissance de chiffres connectés.

En premier lieu, il est intéressant de s'assurer du succès de la reconnaissance pour un système dépendant, c'est-à-dire utilisant comme locuteurs de tests ceux employés pour constituer notre dictionnaire de références. De la base TI-DIGIT-TRN, nous avons pris un ensemble de 30 locuteurs hommes (30M), et 30 locuteurs femmes (30F). Donc, nous utiliserons 60 locuteurs pour l'apprentissage du dictionnaire. Ces mêmes locuteurs seront utilisés pour les tests de reconnaissance des chiffres, afin de pouvoir opérer les tests de façon dépendante des locuteurs d'entrainement. Ces locuteurs ont prononcé deux fois (A et B) les 11 chiffres, ce qui donne 60 × 2×11=1320 occurrences des 11 chiffres isolés. Nous avons implémenté les différents algorithmes en langage Matlab et en utilisant quelques fonctions de la « Voice Box Toolbox». Les paramètres de simulations sont présentés dans le tableau IV.1.

Type de paramétrisation utilisé ⟶ LPC ou LPCC ou MFCC
Durée de la trame ⟶ 32 ms
Fréquence d'échantillonnage ⟶ f_e= 8KHz
Coefficient de pré-accentuation ⟶ 0.93
Nombre de canaux du banc de filtres Mel ⟶ 19
Nombre de coefficients cepstraux ⟶ 12 coefficients

Tableau IV.1 : Paramètres des simulations

L'efficacité d'une méthode de reconnaissance vocale se mesure principalement par sa capacité à assurer un taux de réussite élevé, ce dernier est défini par la formule (4.1):

$$\text{Taux de réussite globale } \% = \frac{\text{nombre de chiffres reconnus}}{\text{nombre de chiffres à reconnaitres}} \text{x } 100 \qquad (4.1)$$

Le tableau IV.2 présente les résultats de la reconnaissance de chacun des 11 chiffres, incluant également le taux de réussite globale correspond à chaque simulation.

N° de simulation	Données d'apprentissage	Données de test	1	2	3	4	5	6	7	8	9	oh	zero	Réussite globale (%)
1	TRN AB 30 M + 30 F	TRN AB 30 M + 30 F	120	120	120	120	120	120	120	120	120	120	120	100
2	TRN A 30 M + 30 F	TRN A 30 M + 30F	60	60	60	60	60	60	60	60	60	60	60	100
3	TRN A 30 M + 30 F	TRN B 30 M + 30F	60	60	60	60	60	59	60	60	60	60	60	99,8485
4	TRN M AB 30 M	TRN M AB 30 M	60	60	60	60	60	60	60	60	60	60	60	100
5	TRN M A 30 M	TRN M A 30 M	30	30	30	30	30	30	30	30	30	30	30	100
6	TRN M A 30 M	TRN M B 30 M	29	30	30	29	30	29	29	30	30	30	30	98,7879
7	TRN F AB 30 F	TRN F AB 30 F	60	60	60	60	60	60	60	60	60	60	60	100
8	TRN F A 30 F	TRN F A 30 F	30	30	30	30	30	30	30	30	30	30	30	100
9	TRN F A 30 F	TRN F B 30 F	30	30	30	29	29	30	30	30	30	30	30	99,3939

Tableau IV.2 : Résultats de la reconnaissance d'un système multi-locuteurs dépendants

Le système de reconnaissance dépendant du locuteur que nous avons réalisé a montré des résultats très satisfaisants avec un taux de réussite assez élevé. Nous considérons cependant, qu'il s'agit d'un travail préliminaire qui nous permettra par la suite l'accomplissement de notre prochain objectif qui est un système de reconnaissance indépendant du locuteur pour les chiffres isolés allant de 1 à 9, plus le « zéro » et « oh » prononcés en anglais. Dans cette perspective, nous projetons de préparer un corpus de tests (TI-DIGIT-TST) dont l'étude sera effectuée ci-après.

IV.4.2Cas d'un système multi-locuteurs indépendants

Dans ce cas, l'intérêt sera de séparer l'ensemble de référence (d'apprentissage)et l'ensemble de test, afin de pouvoir opérer les tests de façon indépendante des locuteurs d'entrainement. Pour les données d'apprentissage en gardant le même ensemble défini précédemment du corpus TI-TRN, c'est-à-dire 30 locuteurs hommes (30 M) et 30 locuteurs femmes (30 F). Tandis que les tests de reconnaissance seront réalisés sur les chiffres prononcés par 30 locuteurs hommes (30M) et 30

locuteurs femmes (30F) du corpus TI-TST. Quant à l'ensemble de test, ces 60 locuteurs possèdent chacun deux versions (A et B) de chaque chiffre isolé, cela nous fournit un assortiment de 1320 chiffres à identifier, ce qui représente une quantité suffisamment importante pour évaluer l'efficacité de la méthode de reconnaissance proposée.

Le tableau IV.3 montre les résultats de la reconnaissance de chacun des 11 chiffres, incluant également le taux de réussite globale correspond à chaque simulation.

N° de simulation	Données d'apprentissage	Données de test	1	2	3	4	5	6	7	8	9	oh	Zero	Réussite globale (%)
1	TRN AB 30 M + 30 F	TST AB 30 M + 30 F	120	118	120	116	120	119	120	119	119	120	120	99,3182
2	TRN A 30 M + 30 F	TST A 30 M + 30 F	59	60	60	58	60	60	59	58	59	60	60	98,9394
3	TRN A 30 M + 30 F	TST B 30 M + 30 F	60	59	60	59	59	60	60	60	60	60	60	99,5455
4	TRN A 30 M + 30 F	TST AB 30 M + 30 F	119	119	120	117	119	120	119	118	119	120	120	99,2424
5	TRN M AB 30 M	TST M AB 30 M	60	59	60	59	60	59	60	60	59	60	60	99,3939
6	TRN M A 30 M	TST M A 30 M	29	30	30	29	30	29	29	30	30	30	30	98,7879
7	TRN M A 30 M	TST M B 30 M	30	29	30	30	30	29	30	30	30	30	30	99,3939
8	TRN M A 30 M	TST M AB 30 M	59	59	60	59	60	58	59	60	60	60	60	99,0909
9	TRN F AB 30 F	TST F AB 30 F	60	59	60	58	60	60	60	59	60	60	60	99,3939
10	TRN F A 30 F	TST F A 30 F	30	30	30	29	30	30	30	28	29	30	30	98,7879
11	TRN F A 30 F	TST F B 30 F	30	30	30	29	29	30	30	30	30	30	30	99,3939
12	TRN F A 30 F	TST F AB 30 F	60	60	60	58	59	60	60	58	59	60	60	99,0909

Tableau IV.3 : Résultat de la reconnaissance d'un système multi-locuteurs indépendants

On remarque d'après le tableau, que le test a abouti à des résultats très satisfaisants et les performances du système indépendant sont démontrées. Les chiffres correctement identifiés représentent pour chaque simulation presque 100%

IV.4.3Cas d'un système multi-locuteurs indépendants bruité

Le pouvoir discriminant est une qualité essentielle de la reconnaissance, il est impératif de montrer la fiabilité des méthodes dans des conditions réelles d'utilisation c'est-à-dire en présence de bruit dans le signal à reconnaitre. En effet, les conditions des tests précédents étaient idéales : utilisation des enregistrements dans un milieu calme (corpus TI-DIGIT). C'est pour cette raison que nous avons tout d'abord volontairement ajouté du bruit à nos échantillons et cela pour différentes valeurs de RSB.

Pour bruiter nos échantillons, nous leurs superposons à notre échantillon des bruits réels (white, pink et babble) tirés de la base de données NOISEX-92[32], ainsi d'un bruit AWGN.

L'ajout du bruit AWGN (*Additive white Gaussian noise*) de moyenne nulle et de variance σ^2en utilisant un rapport signal à bruit (RSB ou SNR) choisi tel que :

$$\boldsymbol{SNR = 10\ \log_{10}\left(\frac{P_{signal}}{P_{bruit}}\right)} \qquad (4.2)$$

Où P_{signal} et P_{bruit} représentent respectivement la puissance du signal et du bruit.

Remarque :

Le calcul du RSB ne peut se faire que sur un signal stationnaire et le signal de parole ne l'est pas. Le RSB total du signal donc est obtenu en effectuant la moyenne des RSB obtenus pour des tranches de 32 ms du signal échantillonné. C'est pourquoi, nous allons découper l'échantillon en tranches de 32 ms et calculer le RSB pour chacune de ces tranches. En réalisant une moyenne de toutes ces valeurs, nous obtiendrons le RSB global du signal.

Pour cette partie de simulations, nous avons choisi de la base TI_DIGIT les données représentés en tableau IV.4.

Donnéesd'apprentissage (sans bruit)	Données de test (sans et avec bruit)
TRN AB 15 M +15 F	TST AB 15 M + 15 F

Tableau IV.4 : choix des données

Les résultats obtenus des simulations sont illustrés dans le tableau IV.5.

N° de simulation	**Données de tests**	**SNR (RSB) (dB)**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**oh**	**Zero**	**Réussite globale (%)**
1	Clean (sans bruit)	_ _	60	60	59	58	60	60	60	56	58	60	58	98,3333
2	AWGN	20	56	58	58	56	60	60	60	56	55	59	59	96,5152
3	AWGN	10	48	48	50	57	60	60	57	56	38	52	53	87,7273
4	AWGN	5	16	15	22	26	60	60	27	35	15	18	15	46,8182
5	white	20	56	59	58	57	60	60	60	57	56	59	59	97,1212
6	white	10	49	52	50	53	60	60	59	56	41	53	55	89,0909
7	white	5	0	0	14	9	60	60	10	30	2	5	2	29,0909
8	pink	20	56	60	58	57	60	60	60	56	58	60	59	97,5758
9	Pink	10	56	59	58	56	60	60	59	57	50	59	57	95,6061
10	pink	5	20	21	38	47	60	60	49	51	17	33	34	65,1515
11	babble	20	59	60	60	57	60	60	60	58	60	60	57	98,6364
12	Babble	10	59	60	60	57	60	60	60	58	55	60	57	97,8788
13	Babble	5	56	60	58	54	59	58	59	59	52	59	56	95,4545

Tableau IV.5 : Résultats de la reconnaissance d'un système multi-locuteurs indépendants bruité

Les figures IV.5 et IV.6 montrent respectivement la réussite globale et la réussite pour les onze chiffres en fonction du RSB (20 dB, 10 dB et 5 dB) pour différents types de bruits.

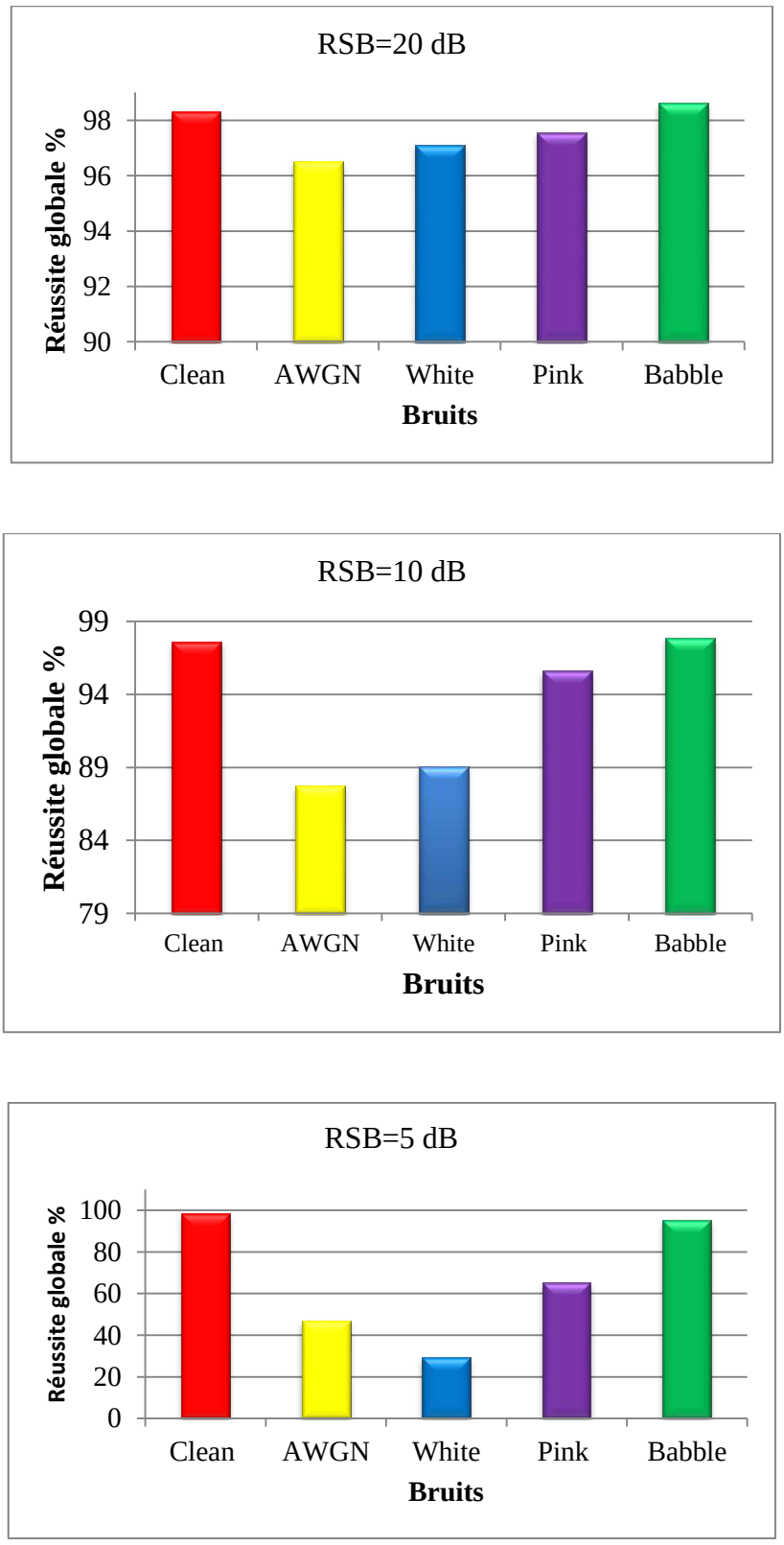
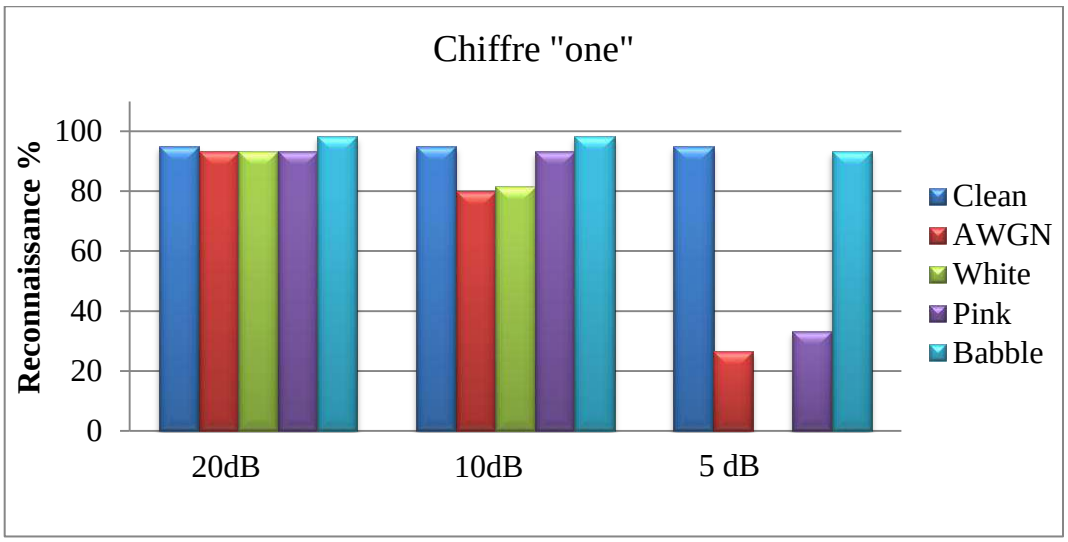
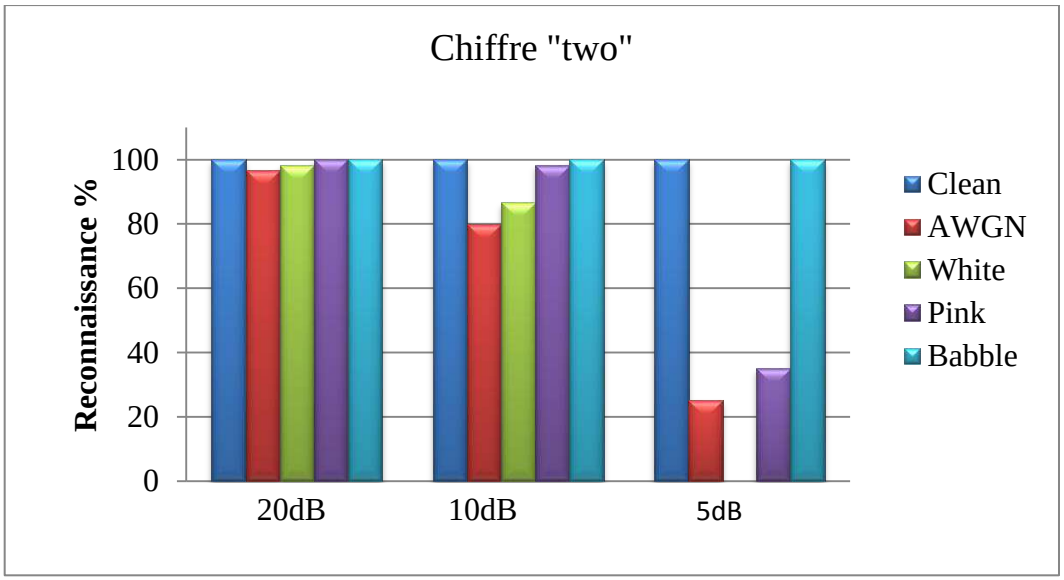
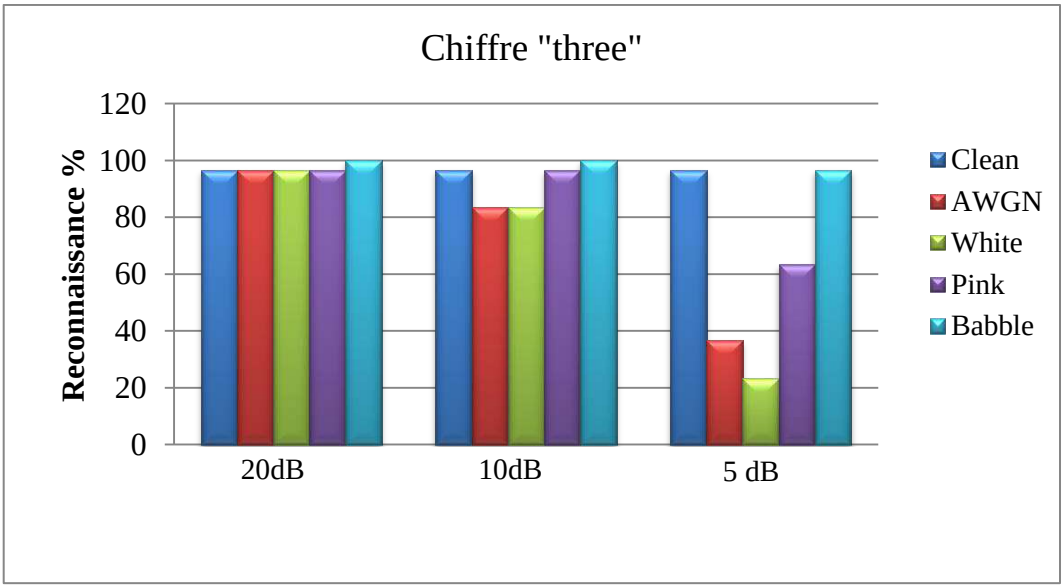
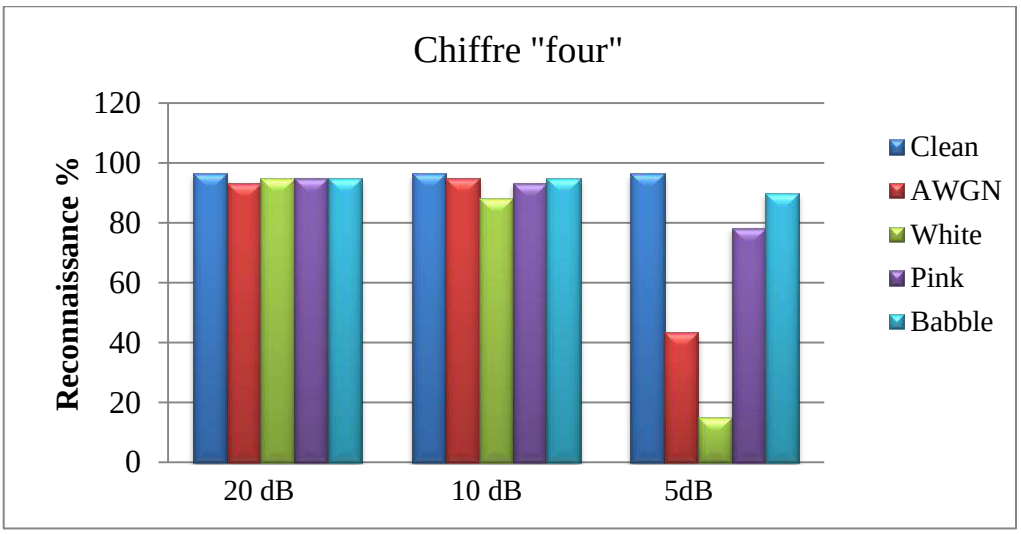


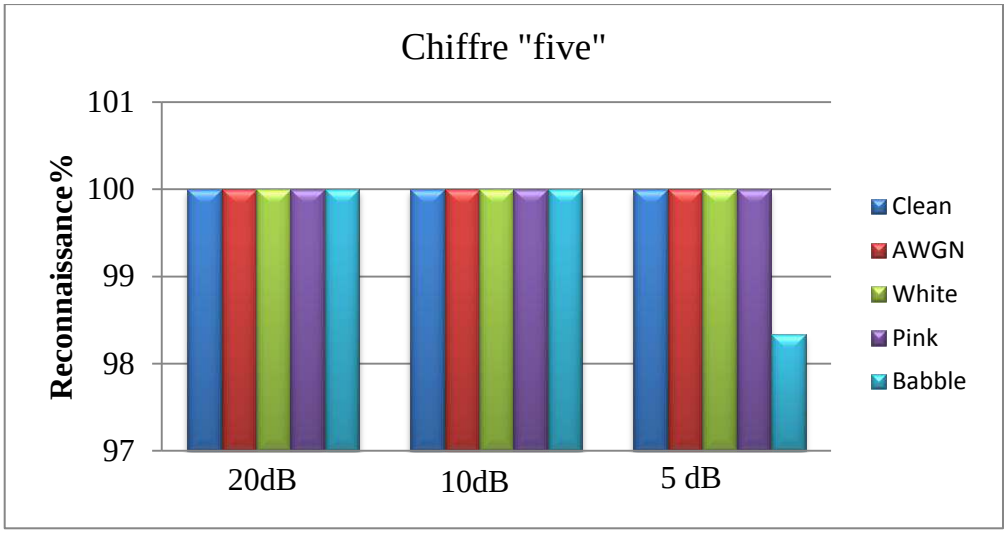
Figure IV.5 : Réussite globale en fonction du RSB (20 dB, 10 dB et 5 dB) Pour différents types de bruits

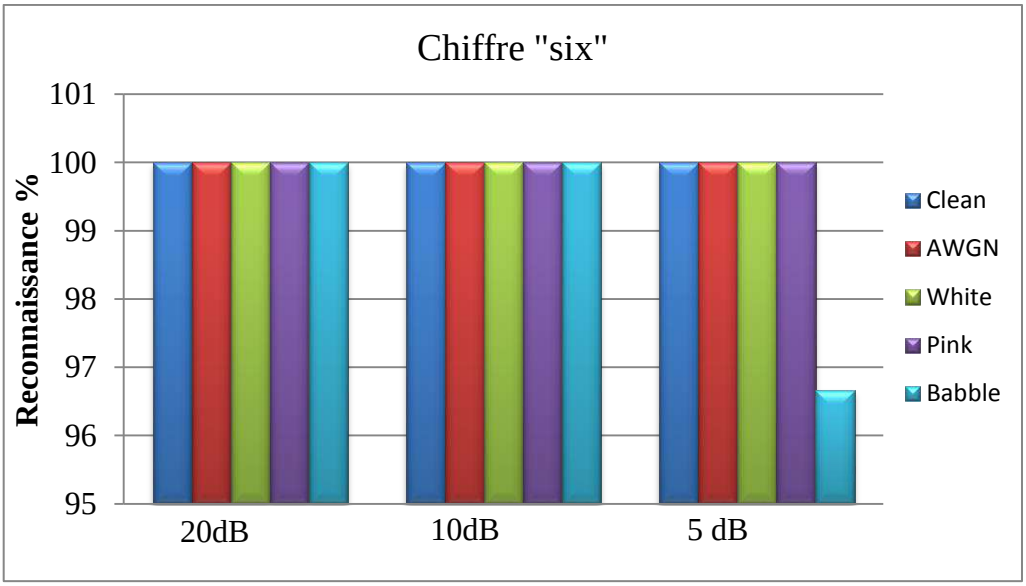

Chiffre "one"
Reconnaissance %
100
80
60
40
20
0
20dB
10dB
5 dB
Clean
AWGN
White
Pink
Babble

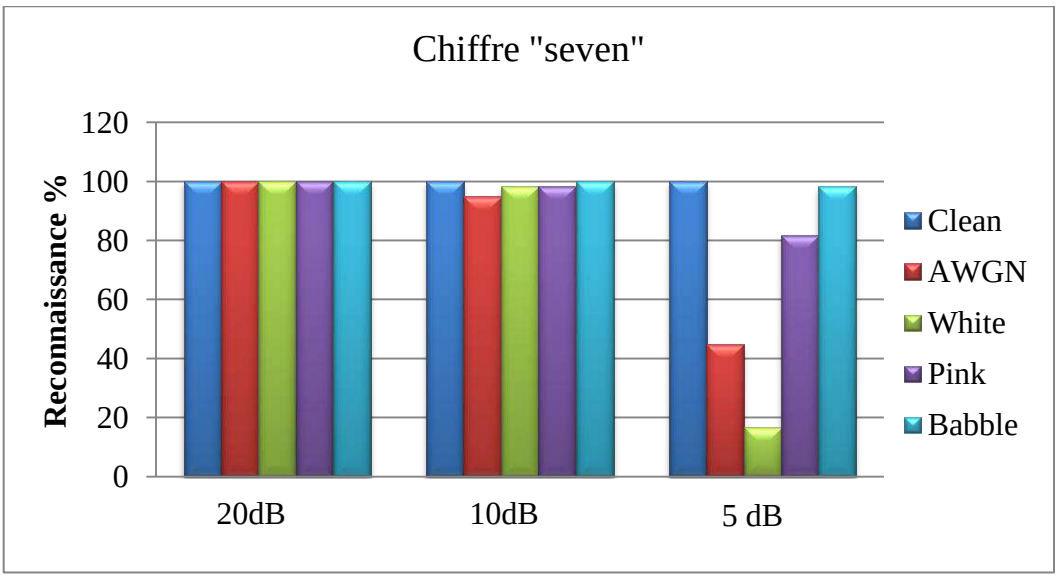

Chiffre "two"
Reconnaissance %
100
80
60
40
20
0
20dB
10dB
5dB
Clean
AWGN
White
Pink
Babble

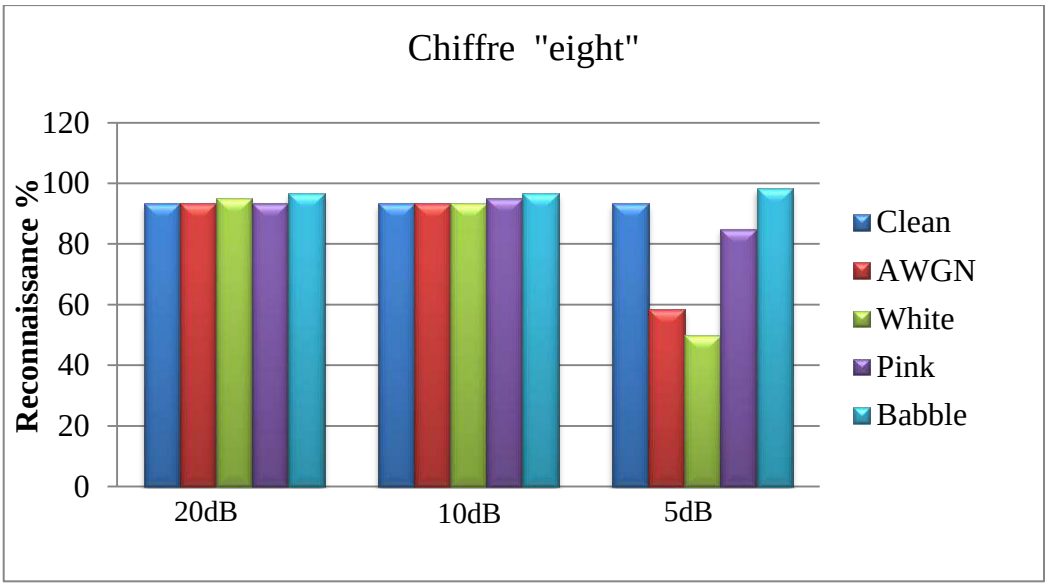

Chiffre "three"
Reconnaissance %
120
100
80
60
40
20
0
20dB
10dB
5 dB
Clean
AWGN
White
Pink
Babble

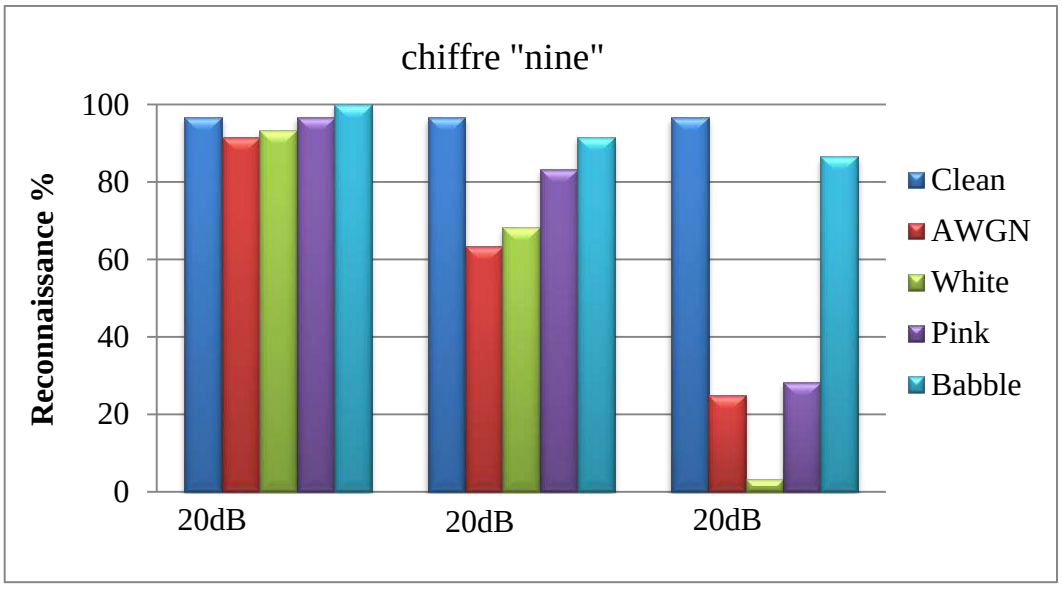

Chiffre "four"
Reconnaissance %
120
100
80
60
40
20
0
20 dB
10 dB
5dB
Clean
AWGN
White
Pink
Babble

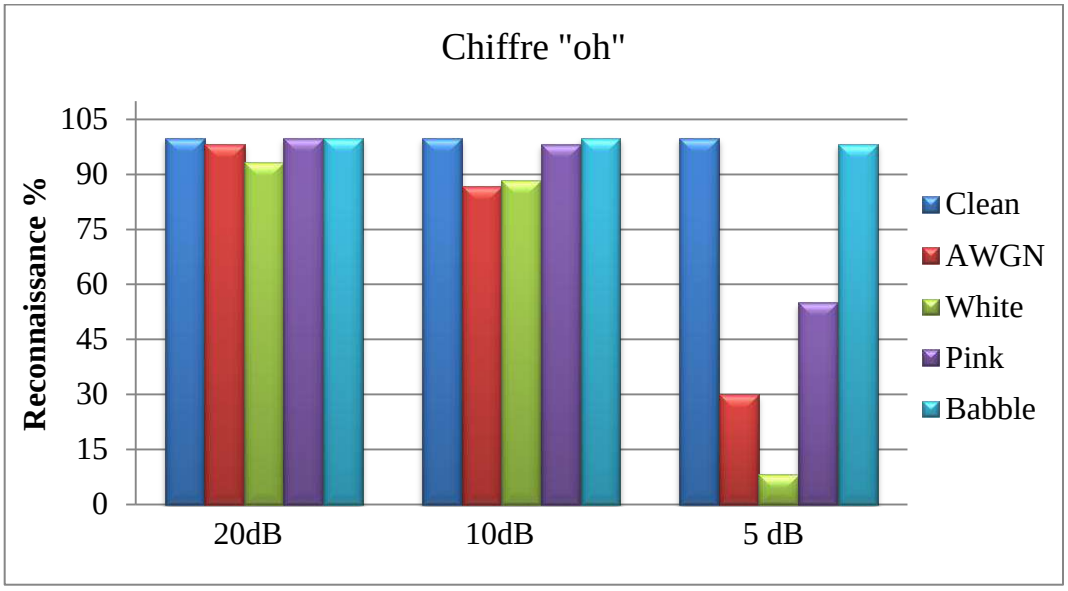

Chiffre "five"
Reconnaissance%
101
100
99
98
97
20dB
10dB
5 dB
Clean
AWGN
White
Pink
Babble


Chiffre "six"
Reconnaissance %
101
100
99
98
97
96
95
20dB
10dB
5 dB
Clean
AWGN
White
Pink
Babble


Chiffre "seven"
Reconnaissance %
120
100
80
60
40
20
0
20dB
10dB
5 dB
Clean
AWGN
White
Pink
Babble


Chiffre "eight"
Reconnaissance %
120
100
80
60
40
20
0
20dB
10dB
5dB
Clean
AWGN
White
Pink
Babble


chiffre "nine"
Reconnaissance %
100
80
60
40
20
0
20dB
20dB
20dB
Clean
AWGN
White
Pink
Babble



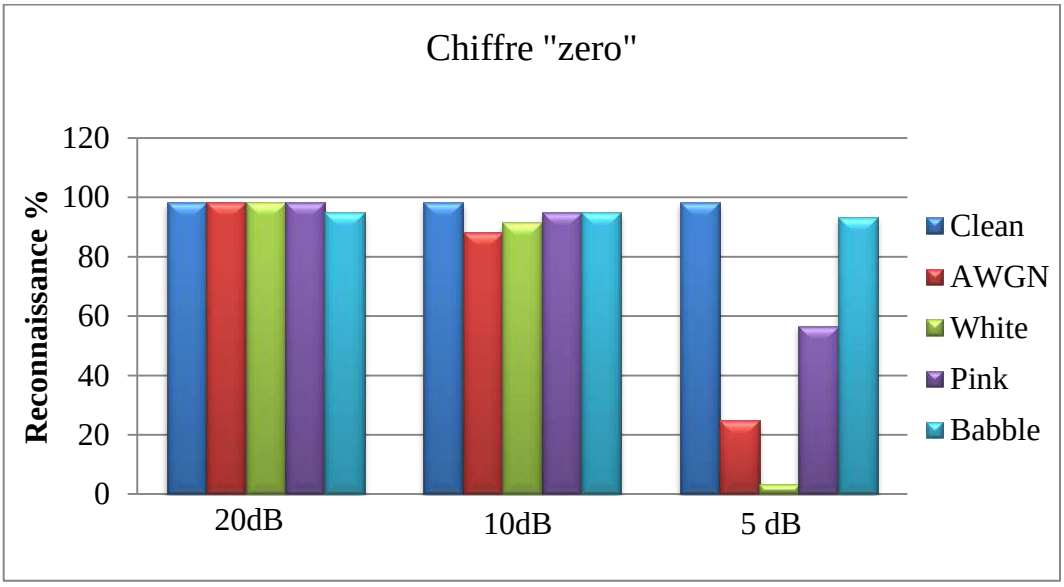


Figure IV.6 : Réussite de la reconnaissance des onze chiffres en fonction du RSB (20 dB, 10 dB et 5 dB) pour différents types de bruits

IV.4.4 Cas d'un système multi-locuteurs indépendants bruité avec plusieurs techniques de Paramétrisation spectrale

En téléphonie numérique fixe, le codage PCM nécessite un taux de transfert de 56/64kO (8KHz x7/8bits). Pour la téléphonie mobile, ce taux s'avère être trop élevé; de nouveaux codec sont donc été développés. Parmi ceux basés sur la technique LPC, le débit varie entre 12,2 kO/s (EFR) et 4,75 kO/s (AMR plus bas débit). L'utilisation de LPC pour la RAP évite l'implémentation d'une nouvelle analyse et donc nécessite moins de ressources mémoire. Pour évaluer la possibilité de cette solution, nous avons comparé différents algorithmes de paramétrisation (LPC, LPCC et MFCC) dans le cadre spécifique de la reconnaissance en systèmes embarqués. Les 12 coefficients auxquels on ajoute l'énergie (soit 13 coefficients) ont été utilisés pour le décodage. Comme déjà mentionné dans le deuxième chapitre, les coefficients LPCC sont calculés à partir des coefficients LPC, ceci permet de minimiser l'accroissement du cout de calcul, car ils sont obtenus par une simple récursivité.

Pour les simulations nous avons utilisé les mêmes données de la partie précédente. Les résultats obtenus sont illustrés dans le tableau IV.6.

SNR (dB)	Technique utilisée	1	2	3	4	5	6	7	8	9	oh	zero	Réussite globale (%)
CLEAN (sans bruit)													
_ _	MFCC-DTW	60	60	59	58	60	60	60	56	58	60	58	98,3333
_ _	LPCC-DTW	59	58	60	58	59	56	58	57	56	60	57	96,6667
_ _	LPC-DTW	52	54	56	49	53	57	54	56	49	53	44	87,4242
AWGN													
20	MFCC-DTW	56	58	58	56	60	60	60	56	55	59	59	96,5152
	LPCC-DTW	56	55	59	55	59	58	57	57	50	59	55	93,9394
	LPC-DTW	34	50	53	41	52	59	31	56	28	51	39	74,8485
10	MFCC-DTW	48	48	50	57	60	60	57	56	38	52	53	87,7273
	LPCC-DTW	37	44	52	50	59	59	53	52	28	51	44	80,1515
	LPC-DTW	17	23	41	15	57	59	26	50	9	33	22	53,3333
5	MFCC-DTW	16	15	22	26	60	60	27	35	15	18	15	46,8182
	LPCC-DTW	0	0	8	2	60	59	11	33	1	2	2	26,9697
	LPC-DTW	0	0	4	0	52	60	3	24	1	0	0	21,8182

WHITE													
20	MFCC-DTW	56	59	58	57	60	60	60	57	56	59	59	97,1212
	LPCC-DTW	56	56	59	57	59	58	58	56	50	60	55	94,5455
	LPC-DTW	34	50	52	36	55	59	34	57	22	51	33	73,1818
10	MFCC-DTW	49	52	50	53	60	60	59	56	41	53	55	89,0909
	LPCC-DTW	45	47	52	52	59	59	56	52	31	53	47	83,7879
	LPC-DTW	19	28	43	18	59	56	27	51	10	37	23	56,2121
5	MFCC-DTW	0	0	14	9	60	60	10	30	2	5	2	29,0909
	LPCC-DTW	0	0	8	2	60	59	9	38	2	3	1	27,5758
	LPC-DTW	0	0	4	0	50	60	2	26	1	1	0	21,8182
PINK													
20	MFCC-DTW	56	60	58	57	60	60	60	56	58	60	59	97,5758
	LPCC-DTW	57	58	59	57	59	56	58	55	54	60	57	95,4545
	LPC-DTW	36	56	56	41	52	56	41	57	35	53	41	79,3939
10	MFCC-DTW	56	59	58	56	60	60	59	57	50	59	57	95,6061
	LPCC-DTW	54	56	59	55	59	59	58	54	48	60	57	93,7879
	LPC-DTW	29	52	50	33	54	59	41	58	20	48	30	71,8182
5	MFCC-DTW	20	21	38	47	60	60	49	51	17	33	34	65,1515
	LPCC-DTW	19	16	46	32	60	59	48	51	12	31	24	60,3030
	LPC-DTW	12	19	28	11	57	59	31	49	7	13	9	44,697
BABBLE													
20	MFCC-DTW	59	60	60	57	60	60	60	58	60	60	57	98,6364
	LPCC-DTW	58	58	60	54	59	57	55	58	56	60	58	95,9091
	LPC-DTW	52	54	58	45	51	57	49	57	41	53	45	85,1515
10	MFCC-DTW	59	60	60	57	60	60	60	58	55	60	57	97,8788
	LPCC-DTW	59	58	60	53	57	56	55	56	58	60	58	95,4545
	LPC-DTW	48	57	55	43	51	59	44	58	40	53	43	83,4848
5	MFCC-DTW	56	60	58	54	59	58	59	59	52	59	56	95,4545
	LPCC-DTW	54	58	58	52	56	56	53	55	53	59	59	92,8788
	LPC-DTW	40	54	52	40	51	55	37	58	35	52	45	78,6364

Tableau IV.6 : Résultats de la reconnaissance pour différentes techniques de paramétrisation

Pour mieux évaluer les performances et l'influence de chaque technique de paramétrisation à la reconnaissance vocale, on utilise la représentation graphique.

La figure IV.7 représente les résultats de la réussite globale de reconnaissance pour différentes types de paramétrisation (LPC, LPCC et MFCC) dans le cas de la base de données TI-DIGIT, c'est à dire dans un milieu calme (clean).

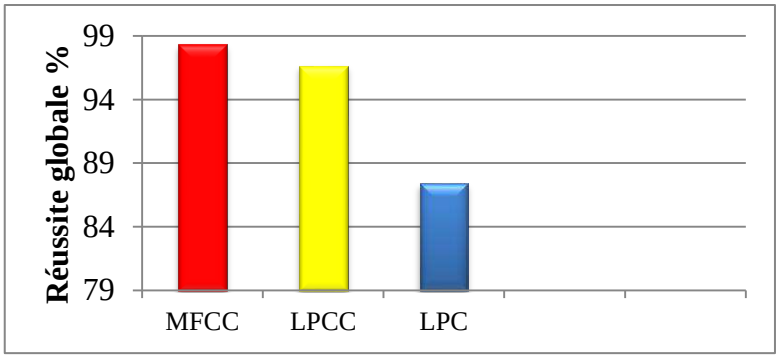


Figure IV.7 : Comparaison de la réussite globale en fonction de différents types de paramétrisation en milieu calme (clean)

On remarque que les deux techniques MFCC et LPCC donnent de bons résultats devant la technique LPC, mais cette dernière peut être considérée acceptable vu sa simplicité d'implémentation.

La figure IV.8 illustre le résultat de la réussite globale de reconnaissance pour différentes types de paramétrisation (LPCC, LPC, MFCC) dans le cas des bruits AWGN, white, pink, babble avec plusieurs valeurs de RSB (20 dB, 10 dB et 5 dB). Pour les valeurs supérieur à 20 dB, les résultats sont similaires aux valeurs obtenus en milieu calme (clean), tandis que pour les valeurs inférieurs à 5dB donnent des mauvaises résultats.

Dans un milieu bruité, les résultats (Tableau IV.6 et Figure IV.8) montrent que les coefficients LPCC sont beaucoup plus performants que les coefficients LPC classiques. La paramétrisation MFCC basée sur une analyse en banc de filtres du signal offre de meilleures performances. Néanmoins, cette méthode de paramétrisation est complexe et nécessite plus de ressources.

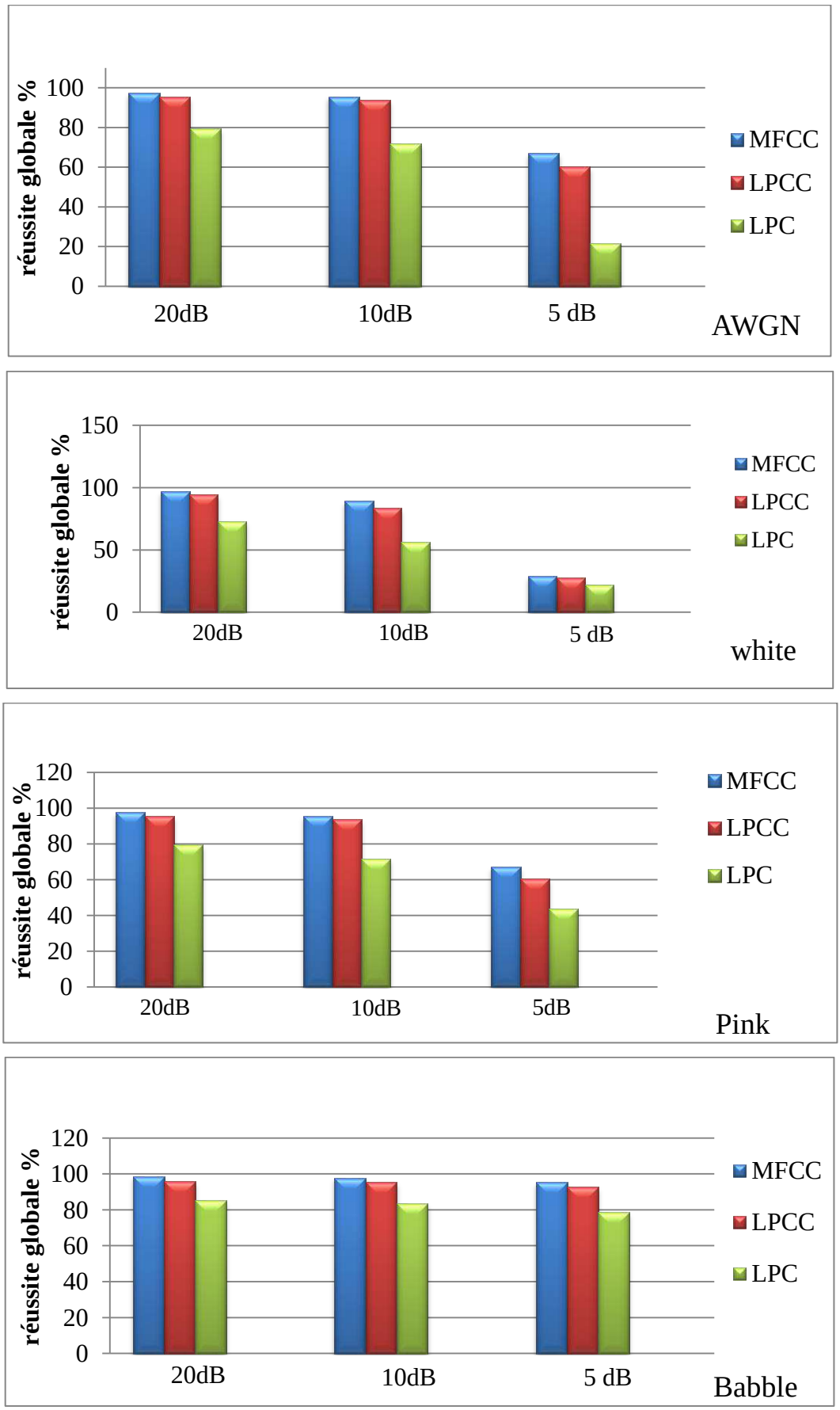


Figure IV.8: Comparaison de la réussite globale en fonction du RSB pour différents types de paramétrisation avec différents types de bruits

IV.5Conclusion

Dans ce chapitre, nous avons présenté et implémenté sous Matlab un système de reconnaissance vocale, en utilisant plusieurs techniques de paramétrisation, d'estimer leur performance et de vérifier quelles étaient leurs sensibilités au différents types de bruit.

Dans ce travail, nous nous sommes concentrés sur la reconnaissance des chiffres isolés pour un système embarqué (par exemple un téléphone portable).Dans ce contexte, le système de reconnaissance doit respecter certaines contraintes tellesque la place mémoire et les ressources de calcul.

Nous avons montré que les coefficients cepstraux LPCC issus des LPC (par recursion) permettent un bon rapport performance /cout (calcul et mémoire) .Néanmoins, la paramétrisation basée sur une analyse en banc de filtres(MFCC) reste plus performante.

CONCLUSION GENERALE

Dans ce travail nous avons proposé un système de reconnaissance dépendant/indépendant des locuteurs, notre objectif était de développer une méthode de reconnaissance automatique robuste pouvant reconnaitre des chiffres isolés, prononcés par plusieurs locuteurs différents à partir d'un système possédant des ressources limités.

Une simple procédure commence par l'extraction des informations nécessaires au décodage de la parole, à partir d'un découpage en trame, une analyse par prédiction linéaire est effectuée, de laquelle on en déduit les coefficients cepstraux afin de disposer d'une information très précise du signal.

Pour le processus de reconnaissance, un dictionnaire était nécessaire afin de pouvoir comparer la forme inconnue avec l'ensemble de référence, les tests de reconnaissance ont été effectués sous Matlab, sur la base de données TIDIGIT de Texas Instrument, sur des chiffres isolés prononcés en anglais, allant de 0 à 9 dans notre cas.

Dans la première partie des simulations, toutes les expériences ont été réalisées avec un système de reconnaissance basé sur les techniques MFCC et DTW, les résultats obtenus montrent clairement que la reconstitution de la forme rapprochant au chiffre à reconnaître est presque 100 %.

Dans une deuxième partie de simulations, nous avons implémenté plusieurs techniques de paramétrisation (LPC, LPCC et MFCC), en ajoutant différents types de bruits avec plusieurs niveaux du SNR. Les résultats trouvés montrent que les coefficients LPCC issus des LPC permettent une meilleure performance/ cout (calcul et mémoire) tandis que la paramétrisation basée sur une analyse en banc de filtre MFCC reste la plus performante.

Pour les perspectives futures, on peut envisager l'élargissement de cette méthode sur la reconnaissance automatique d'un vocabulaire limitée de mot (chiffres connectés). Pour cela on propose une procédure fondée sur la segmentation en syllabes et le seuillage d'énergie.

Bibliographie

[1] http://outilsrecherche.over-blog.com/pages/Notes_131_Lappareil_Phonatoire_Humain-3083095.html [consulter en 2011].

[2] RENE BOITE. MURAT KUNT **«** Traitement de la parole**»**.1ère édition.1987. Lausanne: Presses polytechniques. Romandes. p 280.

[3] CHRISTINE MEUNIER. Laboratoire parole et langage –CNRS - .France. Janvier 2010.

[4] CHARPENTIER CHRISTOPHE « Conception d'une méthode robuste de reconnaissance de la parole pour un système embarqué ». Mémoire présenté à l'école de technologie supérieure, Université du QUEBEC ,2008.

[5] CALLIOPE «La parole et son traitement automatique ». Paris. Masson. 717p.1989.

[6] A.LANDERY, R. RENARD, Didier 1977 « éléments de phonétique » ,Bruxelles.

[7] JEAN.HENNERBERT « traitement de la parole » université de Fribourg, Suisse, SE2006.

[8] RENE BOITE, HERVE BOURLARD, THIERRY DUTOIT, JOEL HANCQ et HENRI LEICH, « Traitement de la parole », presses polytechniques et universitaires romande,1999.

[9] http://www:cyberphon.ish-lyon.cnrs.fr/partie3/p3c1:htm

[10] T .DUTOIT. « Introduction au traitement automatique de la parole », Faculté polytechnique de Mons, Belgique, 2000.

[11] G.VON BEKESY, Experiments in Hearing,McGraw-Hill,New York, 1960.

[12] A. DRYGAJLO « Traitement de la parole », EPFL, Lausanne, 2005.

[13] ALAIN GOYE. La perception auditive , Ecole Nationale Supérieur des Télécommunication, ENST, Paris, 2002.

[14] www.sante.univ-nantes.fr/med/ticem/ressources/1367.pdf [consulter en 2012]

[15] JUNQUA, J.C. et J.P. Haton. « Robustness in Automatic Speech Recognition, 1ère éd,. Norwell : Kluwer Academic Publisher, 440p. 1996.

[16] Dr.Mohamed DEBYECHE, Analyse du signal de la parole en vue de sa reconnaissance » PG: Communication Parlée, LCPTS: Laboratoire de Communication Parlée et Traitement des Signaux, Université Houari Boumediene.

[17] T.PARSONS, « Voice and Speech Processing», McGraw-Hill, 1986.

[18] C-S.GARGOUR, « Traitement numériques des signaux », Ecole de technologie supérieure, 2001.

[19] KHALED ZAABI, Implémentation d'une méthode de reconnaissance de la parole sur le processeur de traitement numérique du signal, TMS320C6711, Ecole de technologie supérieure, Université de QEBEC, juin 2004.

[20] T-F.QUATIERI, « Discrete-Time Speech Signal Processing Principles and Practice », Prentice Hall PTR.

[21] KUNT, MURAT. « Traitement numérique des signaux ». 2ème éd. Lausanne : Presses polytechniques romandes, 402p. 1984.

[22] G. BAUDOIN et J.-F. BERCHER, « TRANSFORMÉE DE FOURIER DISCRÈTE , École Supérieure d'Ingénieurs en Électrotechnique et Électronique, version 0.1, Novembre 2001.

[23] Z. HAMAIZIA, M. BEDDA, « Analyse et synthese de la parole arabe », Octobre 2011.

[24] ITAKURA, F. et S. SAITO, «Analysis synthesis telephony based on the maximum likelihood method». Ininter-n. Congr. Acoust. Tokyo. 1968.

[25] J.MAKHOU, «linear prediction: Atutorial review», proceeding of the IEEE, vol.3, pp561-580, 1975.

[26] YASSIN MAMI, « Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence», Ecole Nationale Supérieure des Télécommunication. Paris, 2003.

[27] S.DAVIS and P.MERMELSTEIN, « Comparison of representations for monosyllabic word recognition in continuously spoken sentences», IEEE transactions on, Acoustics, Speech, and Signal Processing, vol.28, pp.357-366, 1980.

[28] L.Rabiner and B.Juang, «Fundamentals of Speech Recognition », Prentice Hall,1993.

[29] BOUM. L. «An inequality and associated masimization technique in statistical estimation for probabiliste function of Markov Process ». Inequalities, Vol .3 ,p .18, 1972.

[30] J .K .BAKER, Stochastic modeling for automatic speech understanding », speech Recognition , Academic Press ,PP .521-542, 1975.

[31] WILPON .J et L RABINER « A modified clustering algorithm for use in isolated word recognition » IEEE Trans ASSP . Vol .3. n 33, p 587-594, 1985.

[32] Spib.rice.edu/spib/select_noise. [consulter en 2013].

[33] www.ee.columbia.edu/~dpwe/sounds/tidigits/ [consulter en 2013].