الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira -
Tasdawit Akli Muḥend Ulḥağ - Tubirett -

وزارة التعليم العالي والبحث العلمي
جامعة أكلي محند أولحاج
-البويرة-

جامعة البويرة

Faculté des Sciences et des Sciences Appliquées

كلية العلوم والعلوم التطبيقية

Référence : ...../MM/2021

المرجع:...../م م/ 2021

Mémoire de Master

Présenté au

Département : Génie Électrique

Domaine : Sciences et Technologies

Filière : Télécommunications

Spécialité : Systèmes des Télécommunications

Réalisé par :

MESBAH Randa

Et

RAFED Nacira

Thème

Modélisation acoustique pour un système de reconnaissance

Soutenu le:**02/07/2023**

Devant la commission composée de :

Mr :	AYAD Mouloud	Prof.	Univ. Bouira	Président
	REZKI Mohamed	M.C.A	Univ. Bouira	Rapporteur
	SAIDI Mohammed	M.A.A	Univ. Bouira	Examinateur

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي
جامعة أكلي محند أولحاج - البويرة

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira
Tasdawit Akli Muḥend Ulḥağ - Tubirett

نموذج التصريح الشرفي الخاص بالالتزام بقواعد النزاهة العلمية لإنجاز بحث.

انا الممضي اسفله،

السيد(ة).....الصفة: طالب، استاذ، باحث .....

الحامل(ة) لبطاقة التعريف الوطنية: .....والصادرة بتاريخ.....

المسجل(ة) بكلية :العلوم و العلوم التطبيقية..... قسم:..... الهندسة الكهربائية.....

والمكلف(ة) بإنجاز اعمال بحث(مذكرة، التخرج، مذكرة ماستر، مذكرة ماجستير، اطروحة دكتوراه).

عنوانها: .....

.....

تحت إشراف الأستاذ(ة): .....

أصرح بشرفي اني ألتزم بمراعاة المعايير العلمية والمنهجية الاخلاقيات المهنية والنزاهة الاكاديمية المطلوبة في انجاز البحث المذكور أعلاه.

التاريخ:.....

توقيع المعني(ة)

رأي هيئة مراقبة السرقة العلمية:

النسبة:itin % 23

الامضاء:

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي
جامعة أكلي محند أولحاج - البويرة

Ministère de l'Enseignement Supérieur
et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira
Tasdawit Akli Muḥend Ulḥağ - Tubirett

نموذج التصريح الشرفي الخاص بالالتزام بقواعد النزاهة العلمية لإنجاز بحث.

انا الممضي اسفله،

السيد(ة)..... الصفة: طالب، استاذ، باحث .....

الحامل(ة) لبطاقة التعريف الوطنية: ..... والصادرة بتاريخ.....

المسجل(ة) بكلية :العلوم و العلوم التطبيقية..... قسم:..... الهندسة الكهربائية.....

والمكلف(ة) بإنجاز اعمال بحث(مذكرة، التخرج، مذكرة ماستر، مذكرة ماجستير، اطروحة دكتور

عنوانها: .....

.....

.....

تحت إشراف الأستاذ(ة): .....

أصرح بشرفي اني ألتزم بمراعاة المعايير العلمية والمنهجية الاخلاقيات المهنية والنزاهة الاكاديمية المطلوبة في انجاز البحث المذكور أعلاه.

التاريخ:.....

توقيع المعني(ة)

رأي هيئة مراقبة السرقة العلمية:

النسبة: % itin

الامضاء:

Dédicaces 1

Je dédie

Ce modeste travail à mes chers parents qui ont toujours été à mes coté pour me soutenir et m'encourager durant ces longues années d'études offrants un énorme soutien moral, affectif et financier, à mon cher frère Mouloud et chères sœurs Sarah, Dounia ,Liza et Malak , a ma cousine Tina et a ma cher copine Asma et à toute ma famille et mes amis ,Sans oublier mon amie et ma binôme Nacira.

Randa

Dédicaces 2

Je suis heureuse de dédier ce travail à mes chers parents,

Source d'amour et de tendresse qui m'encouragent à affronter la vie et à poursuivre mes études pour réussir, A tous mes chers frères et sœurs, Azzeddine louanes, Menad Fahima, Wassila, qui m'ont entouré de leur aide et de leur amour, au mari de ma sœur, Mohamed, aux enfants de ma sœur, la joie de la maison, Nihal Ashwaq Islam, à toute ma famille et mes amis, A ma cher binôme Randa, Tout l'amour et le respect pour son aide et sa sincérité pour terminer ce travail.

Nacira

Remerciements

Tout d'abord, nous remercions Allah le tout puissant de nous avoir donné le courage et la patience nécessaires à mener ce travail à son terme.

Nos sincères remerciements en vont vers notre encadreur M. REZKI Mohamed, pour l'aide compétente qu'il nous a apportée et pour sa patience avec nous. Sans oublier M. BELABBACI, pour son aide tout au long de ce mémoire.

Que les membres de jury trouvent, ici, l'expression de notre gratitude pour l'honneur qu'ils nous font en prenant le temps de lire et d'évaluer ce travail.

Pour finir, nous souhaitons remercier toute personne ayant contribué de près ou de loin à la réalisation de ce travail

Résumé

Notre projet décrit plusieurs méthodes de reconnaissance automatique de locuteur, nous avons enregistré une base des données qui se compose en 60 fichier sonores (femmes et hommes). Ce système se compose de deux phases (apprentissage et test).

Ces méthodes comprennent l'étude des applications des certains algorithmes tels que la MFCC (mel-scale fréquence cepstral coefficient), PCA (Principal component analysis), la modélisation vecteur histogramme et l'algorithme issu de l'intelligence artificielle (le SVM Support Vecteurs Machines) afin de faire une reconnaissance de genre (homme-femme) à travers l'empreinte de leur speech (voix).

Pour cela on a utilisé pas mal de méthodes simples dont la plus connue est le pitch, MFCC, l'énergie, PCA mais ça donne des résultats plus au moins acceptables cependant on a trouvé à travers l'application des méthodes hybrides qu'on peut améliorer le taux de reconnaissance du genre.

Apres l'analyse des résultats qu'on a eus, on a trouvé que la meilleure combinaison possible (pour nous techniques choisis) est l'hybridation entre le MFCC et le PCA, Ce qui a permet l'augmentation de la précision.

Mots clés :

Identification automatique, MFCC, Pitch, PCA, Audio, Algorithme combinée, SVM.

Table des Matières

Remerciements	I
Résumé	II
Table des Matière	III
Liste des Figures	IV
Liste des Tableaux	V
Listes des Acronymes et Symboles	VI
Introduction Générale	1

Chapitre I : Production et perception de la parole

I.1 Introduction	02
I.2 Traitement automatique de la voix	02
I.3Définition de la parole	02
I.4L'appareille vocalique (phonatoire)	02
I.4.1Appareil respiratoire (soufflet)	03
I.4.2Larynx	03
I.4.3Cavite supra-glottique	03
I.5L'audition humaine (l'oreille humaine)	03
I.5.1L'oreille externe	04
I.5.2 L'oreille moyenne	04
I.5.3L'oreille interne	04
I.6 Paramètre du signal de parole	04
I.6.1 Hauteur	04
I.6.2 Le Niveau	05
I.6.3 Le Timbre	05
I.6.4L'Amplitude	05
I.7 Perception de la parole	05
I.8 Classification des sons de la parole	06
I.8.1Un son voisée	06
I.8.2Un Son non-voisée ou dévoisé	06
I.9Modélisation de mécanisme de la production de la parole	07
I.10Propriétés statistiques du signal parole	08
I.11 Traitement de la parole	08
I.11.1 Analyse et synthèse	09
I.11.2 Codage	09
I.11.3 Reconnaissance	09
I.12 Reconnaissance automatique de locuteur	10
Généralité	10
I.13 Domaine d'application de la technologie de RAL	10
I.14 les objectifs principaux de RAP	11
I.15 les différente taches en RAL	11
I.15.1 Vérification (authentification) automatique du locuteur (VAL)	11
I.15.2 Identification automatique du locuteur (IAL)	12
I.16 Structure d'un système de reconnaissance automatique de locuteur	12
I.16.1 L'apprentissages	12
I.16.2 Test	12
I.17 Parametrisation du signal de parole	13
I.18 Modélisation	13
I.19 La décision	13

I.20 Conclusion 13

Chapitre II : Traitement numérique du signal vocal

II.1 Introduction 14
Généralité sur les signaux vocaux 14
II.2 Définition d'un signal 14
II.3Définition d'un bruit 14
II.4 Le traitement de signal 14
II.5 Les fonctions principales du traitement de signal 15
II.5.1Cree 15
II. 5.1.1 La synthèse 15
II. 5.1.2 La modulation et le changement de fréquence 15
II.5.2 L'Analyse 15
II.5.2.1 La Détection 15
II.5.2.2 Identification 15
II.5.3 Transformer 16
II.5.3.1 Filtrage 16
II.5.3.2 Codage /Compression 16
II.6 Classification du signal 16
II.6.1 Classification phénoménologique 16
II. 6.1.1 Les signaux déterministes 17
II.6.1.2 Les signaux aléatoires 17
II.6.2 Classification énergétique 17
II.6.2.1 Les signaux a énergie finie 17
II.6.2.2 Les signaux a puissance moyenne finie 17
II.6.3 Classification morphologique 17
II.6.3.1 Les signaux analogique 17
II.6.3.2 Les signaux quantifiès 17
II.6.3.3 Les signaux échantillonnès 17
II.6.3.4 Les signaux numériques 17
II.7 Composition d'un SRAP 18
II.8 Analyse du signal parole 18
II.9 Le prétraitement du signal 19
II.9.1 La numérisation 19
II.9.1.1L'échantillonnage 19
II.9.1.2La quantification 20
II.9.1.3Le codage 20
II.10Les différentes méthodes d'analyse du signal de parole 20
II.10.1La méthode directe 21
II.10.2La méthode indirecte 21
II.10.2.1L'analyse temporelle 21
II.10.2.1.1Analyse par l'énergie 21
II.10.2.1.2Taux de passage par zéro (TPZ) 21
II.10.2.1.3L'autocorrélation 21
II.10.2.2Analyse fréquentielle 22
II.10.2.2.1Analyse par spectrogramme 22
II.10.2.2.2 Analyse par la transformée de Fourier à court terme (TFCT) 22
II.10.2.2.3 La transformée de Fourier rapide(FFT) 23
II.10.2.2.4 Analyse cepstrale 23
II.10.2.2.4.1 Les coefficients LPCs 24

II.10.2.2.4.2 Les coefficients MFCCs 24
II.10.2.2.4.2.1 Préaccentuation 25
II.10.2.2.4.2.2 Fenêtrage 25
II.10.2.2.4.2.3 FFT 25
II.10.2.2.4.2.4 Application de l'échelle de Mel 25
II.10.2.2.4.2.5 Logarithme(Log) 26
II.10.2.2.4.2.6 Transformé en cosinus discrète (DCT) 26
II.10.2.3 Analyse temps-fréquentielle 26
II.10.2.3.1 Transformée en ondelettes 26
II.10.2.3.2 La transformée en ondelettes discrète 27
II.11 Les différentes méthodes de reconnaissance automatique de locuteur 27
II.11.1 Le modèle de Markov Cache(HMM) 28
II.11.1.1 Evaluation de la vraisemblance 29
II.11.1.2 Le décodage 29
II.11.1.3 L'apprentissage 29
II.11.2 Alignement temporelle dynamique ou(DTW) 29
II.11.3 Machine a vecteur de supporte ou(SVM) 30
II.11.4 Quantification vectorielle(VQ) 32
II.11.4.1 L'algorithme K-means 32
II.11.4.1.1 Initialisation 32
II.11.4.1.2 Affectation 32
II.11.4.1 .3 Mise à jour 33
II.11.4.1.4 Tests d'arrêt 33
II.11.5 Le modèle mélange de gaussien ou(GMM) 33
II.11.5.1 Approche GMM-UBM (Gaussien Mixture Model-universel Background Model) 34
II.11.5.2 Estimation du modèle du monde UBM 34
II.11.5.3Estimation du modèle du locuteur par l'adaptation MAP 35
II.11.6 Compensation de l'effet de la variabilité 36
II.11.6.1 Principal component analysis ou PCA 36

II.12 Calcule le scores avec la méthode CSS (Cosine Similarity Scoring) 38
II.13 Conclusion 38

Chapitre II : Traitement numérique du signal vocal

III.1 Introduction 39
III.2 Description de la base de données 39
III.3 Environnement de travail 40
III.4 protocoles de travail 40
III.5 Répartition naïve 40
III.6 Méthodologie proposée pour l'extraction des paramètres désirés 41
III.7 Expérimentations et résultats 42
III.7.1Extraction des paramètres acoustiques 42
III.7.2Modélisation 42
III.7.2.1Modélisation avec la technique basée sur les vecteurs histogrammes 42
III.7.2.1.1 Influence du nombre de coefficients MFCC 43
III.7.2.1.2 Influence des paramètres dynamiques 43
III.7.2.1.3 Influence de nombre de projection PCA 44
III.7.2.1.4Influence de nombre MFCC, Pitch de projection PCA 44
III.7.2.2 Modélisation par SVM Support Victor Machine 45

III.8 conclusion 46
Conclusion Générale 47
Références bibliographiques 48

Liste des Figures

Fig. I.1 : Schéma de l'appareille phonatoire 03
Fig. I.2 : Production de son de la parole 03
Fig. I.3 : Composition de l'oreille humaine 04
Fig. I.4 : Processus de production de la parole dans le cas son voisé 06
Fig. I.5 : Processus de production de la parole dans le cas son non voisé 07
Fig. I.6 : Exemple de signal de son voise(haut) et son non voise(bas) 08
Fig. I.7 : Traitment de la parole 09
Fig. I.8 : La tâche de vérification automatique du locuteur 12
Fig. I.9 : La tâche de d'identification automatique du locuteur 12
Fig. I.10 : La strecture d'un système de verefication du locuteur 13

Fig. II.1: Création de signal 15
Fig. II.2: Modulation de signal 15
Fig. II.3: Identification d'un signal voise et non voise 16
Fig. II.4: Classification des signaux phénoménologiques 16
Fig. II.5: Classification des signaux déterministes 17
Fig. II.6: Classification des signaux 17
Fig. II.7: Classification morphologies des signaux 18
Fig. II.8: Composition de base des SRAP 18
Fig. II.9 : èchantillionnage 20
Fig. II.10: quantificatin 20
Fig. II.11: Analyse homomorphie de la parole 24
Fig. II.12: Diagramme en blocs de LPC 24
Fig.II.13: Diagramme en blocs deMFCC 25
Fig.II.14: Modèle HMM a 4 ètats 29
Fig.II.15: Exemple d'un model DTW 30
Fig.II.16 :Classification SVMlinèaire 30
Fig. II.17 : Classification SVM hyperplan 31
Fig. II.18 **:** Processus de rangement des objets 31
Fig. II.19 : Quantification vectorielle d'un échantillon de dimension 2 32
Fig. II.20 : Mélange de gaussiennes(GMM) construit en utilisant des paramètres acoustiques issus de plusieurs enregistrements 34
Fig. II.21 : Architecteur du système RA à base de GMM-UBM 34
Fig. II.22 : Adaptation MAP d'un modèle GMM-UBM 36
Fig. II.23 : Analyse en Composantes Principales 38

Fig.III.1 : Procédure étape par étape de la méthodologie d'extraction des paramètres. 41
Fig. III.2 : Schéma représentant la Structure du système de base 43
Fig. III.3 : Classification SVM poly-karnel 46

Liste des Tableaux

Tab.III.1. Représentation des fichiers audio 40
Tab.III.2. Représente le taux d'influence de nombre 43
Tab.III.3 : Représentation de l'influence des paramètres 44
Tab.III.4. L'influence de PCA 44
Tab.III.5 : Représentation l'influence des MFCC, Pitch, PCA 45
Tab. III.6 : Représente le taux de reconnaissance par déférente combinaison SVM 45

Listes des Acronymes et Symboles

- **Acronymes**

CSS	Cosine Similarity Scoring
DCT	Transformée en Cosinus Discret
DWT	Transformée Ondelette Discret
DT-CWT	Transformée en Ondelette Continue Discret dans le Temps
DTW	Dynamique Time Warping
EM	Expectation Maximization
FFT	Transformée de Fourier Rapide
GMM	Gaussien Mixture Model
GMM-UBM	Gaussian Mixture Model-Universal background Model
HMM	Model de Markov Cache
IAL	Identification Automatique de Locuteur
LPC	Linear Prédictive Coding
MFCC	Mel Frequency Cepstral Coefficients
PCA	Principal Component Analysis
RAP	Reconnaissance Automatique Parole
RAL	Reconnaissance Automatique Locuteur
SVM	Support Vecteur Machines
SRAP	Système Reconnaissance Automatique de la Parole
TF	transformée de Fourier
TPZ	Taux de Passage Par Zero
TFCT	transformée de Fourier à court terme
VQ	Quantification Vectorielle
VAL	Verification Automatique de Locuteur

Symboles

$\boldsymbol{f_e}$**:** La fréquence d'échantillonnage.

$\boldsymbol{f_{max}}$**:** La fréquence maximale du signal.

$\boldsymbol{T_e}$**:** Période d'échantillonnage.

$\emptyset_{xx}$: Autocorrélation.

$\boldsymbol{f_0}$: La frequence fondamental.

Introduction Générale

L’usage de la reconnaissance automatique du genre a une grande importance particulièrement dans le domaine biomédical, ou des télécommunications ou dans n’importance quel domaine qui nécessite le groupage hommes-femmes tel que les plages privées, les parkings.

Dans ce cadre, notre travail porte sur la modélisation d'un système de reconnaissance vocale, qui se compose de deux phases : la partie apprentissage (identification) qui sert à ajouter une base de données on donnera les caractéristiques de chaque personne, et la phase de test (vérification) qui le système effectuera la détection de la voix d'une femme ou d'un homme.

Une introduction générale a été fournie pour la recherche de l’algorithme le plus significatif (précis) afin de faire cette reconnaissance pour réaliser cette objectif (reconnaissance de genre) notre manuscrit sera divisé (organiser) en trois chapitres principaux s’exposent comme suit :

- Le premier chapitre concerne la présentation de la production et perception de la parole, généralité et différentes taches de la reconnaissance automatique de locuteur qui ont représenté par l’identification et la vérification.
- Dans le deuxième chapitre, nous retrouvons le détail des techniques qui opèrent dans les trois blocs, de prétraitement, d’extraction, à savoir celle des coefficients scepstraux à échelle Mel (Mel Frequency Cepstral Coefficient : MFCC), la PCA (Principal Component Analysis), la modélisation Vecteur par histogramme et SVM (Support Vector Machine), choisies pour notre étude.
- Le dernier chapitre montre la base de données 60 voix (femmes et hommes) qui sera notre matière d’étude comprenant les tests et résultats, pour finir avec une implémentation d’une méthode de fusion des scores pour plus de robustesse.
- Le mémoire se termine par une conclusion générale présentant l’issue du travail réalisé dans ce mémoire.

Chapitre I :

Production et perception de la parole

I.1 Introduction :

La parole est un moyen de communication et un langage incarné de l'homme qui sera adressé à l'interlocuteur, peut-être lui-même, mentalement, ou par écrit. Elle permet d'exprimer les besoins, les pensées, les sentiments, les douleurs et les aspirations du locuteur [01].

Un large éventail de technologies est disponible, allant des systèmes de reconnaissance de la parole spécialisés pour un seul locuteur à ceux capables de reconnaître des centaines de milliers de mots. De plus, de nombreux services requièrent désormais une identification vocale pour accéder aux boîtes vocales, aux services d'abonnement et aux consultations de comptes bancaires, entre autres exemples [02].

I.2 Traitement Automatique de la voix :

La voix, employée instinctivement par les êtres humains pour reconnaître une personne, représente un comportement modulable qui peut être affecté par des pathologies, du stress ou des fluctuations émotionnelles. De plus, elle peut être altérée intentionnellement par le locuteur. Malgré tout, elle conserve des traits distinctifs significatifs qui permettent d'identifier l'orateur même s'il tente de falsifier sa voix [03].

I.3 Définition de la parole :

La parole est un flux sonore en constante évolution, caractérisé par une énergie limitée et une non-stationnarité. Sa structure est à la fois complexe et variable dans le temps [04], et peut être considérée [02] :

- Périodique (plus exactement pseudopériodique) pour les sons voisés,
- Aléatoire pour les sons fricatifs,
- Impulsionnelle dans les phases explosives des sons occlusifs.

I.4 L’appareil vocalique (phonatoire) :

L'appareil phonatoire englobe tous les organes impliqués dans la production de la parole, ainsi que les muscles responsables de leur mouvement. Ils permettent la production des phonèmes, ou de sons propres à la langue parlée [05]. Ce son est produit par des organes utilisés par les humains et est défini comme suit [06] :

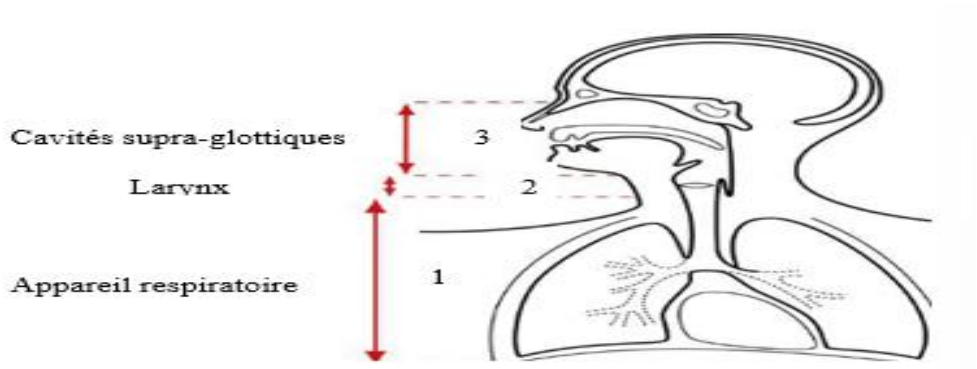

Fig.I.1 : Schéma de l'appareil phonatoire

I.4.1 Appareil respiratoire (soufflerie) : commence au nez et à la bouche, se poursuit dans les voies respiratoires du cou et du thorax et se termine dans les alvéoles pulmonaires [07].Le souffle est un circuit respiratoire commence par deux cavités séparées par le pli de l'aile.

I.4.2 Larynx : est le vibrateur de l'appareil phonatoire, situé entre l'appareil respiratoire et les cavités de résonance [08], composée par des cordes vocales.

I.4.3 Cavités supra-glottiques : Les cavités de résonance jouent un rôle essentiel dans la distinction des voyelles et des consonnes au sein de la parole. Composée de pharynx, cavité buccale et nasale [09].

Les trois étapes de la production des sons de parole sont rappelées dans la figure ci-dessous [10] :

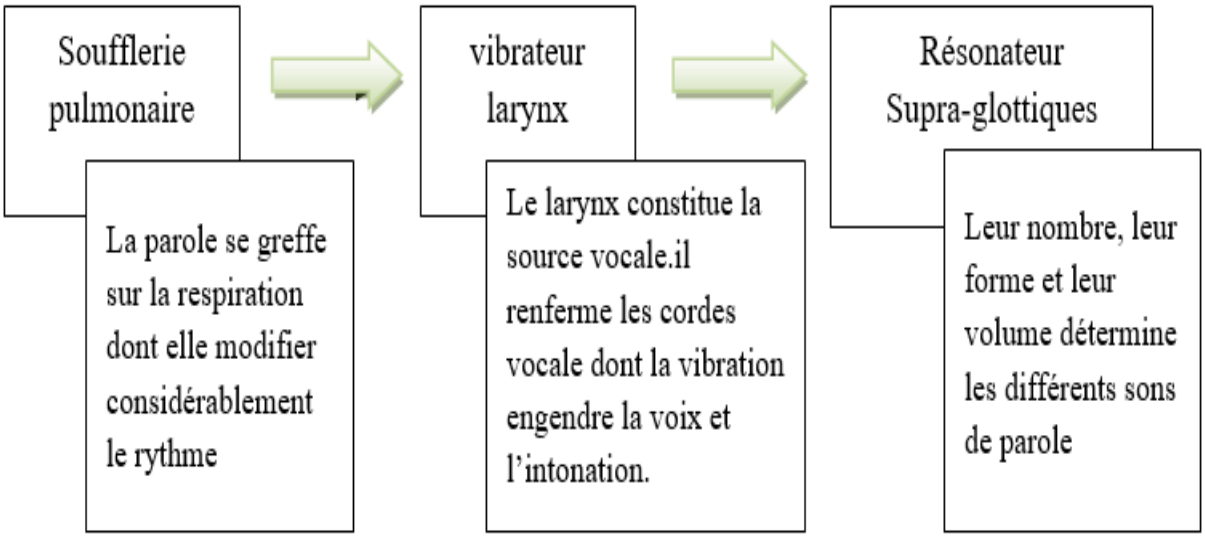

Fig. I.2 : Production de son de la parole [10]

I.5 L'audition humaine (L'oreille humaine) :

L'oreille humaine : est la faculté permettant aux individus de percevoir et donner un sens aux sons qui les entourent [11]. L'oreille enregistre les ondes sonores et les convertit en impulsions électriques. La structure anatomique de l'oreille est composée de trois zones [12]. Chaque zone a une fonction différente :

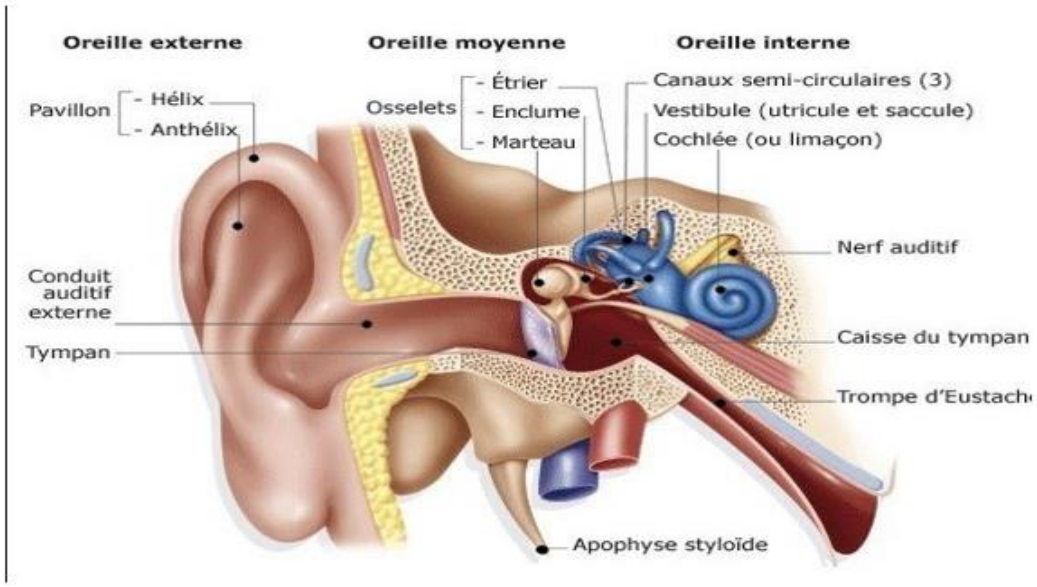

Fig. I.3 : Composition de l'oreille humaine.

I.5.1 L'oreille externe : Recueille les vibrations sonores et les oriente vers le tympan, elle est constituée du [13] : pavillon**,** conduit auditif externe et tympan.

I.5.2 L'oreille moyenne : se trouve entre l'oreille externe et l'oreille interne, restant invisible depuis l'extérieur de l'oreille [12], sa fonction consiste à capter les vibrations sonores dans l'air et les convertir en vibrations à travers les structures de l'oreille, permettant ainsi la transmission du son. [14].

I.5.3 L'oreille interne : est la partie la plus profonde du système auditif [12]. Au sein de cette partie, on trouve la cochlée abritant les cellules sensorielles essentielles à la perception des sons, le vestibule qui joue un rôle dans l'équilibre, ainsi que le point de départ du nerf auditif qui transmet les signaux sonores sous forme d'impulsions électriques vers le cerveau [15].

I.6 Paramètres du signal de parole :

Le signal vocal est caractérisé par quatre paramètres, Un son se différencie ainsi d'un bruit : hauteur, timbre, intensité et durée [16].

I.6.1 Hauteur :

La hauteur correspond au sens de l'ouïe associé à la fréquence de vibration des cordes vocales, également appelée larynx basique ce varie selon les caractéristiques physiques de l'orateur. La variabilité est importante et dépend du sexe et de l'âge de l'individu [17] :

- Autour de 100 Hz pour un homme

- Autour de 200 Hz pour une femme,
- Autour de 300 Hz pour un enfant.

Ainsi, la fréquence des sons permet au système auditif de différencier une voix grave (ou basse) d'une voix aiguë (ou haute) et de reconnaître l'identité d'un locuteur [17] [28].

I.6.2 Le niveau :

Dans un environnement calme, le niveau moyen de la parole sur une distance de 1 mètre est compris entre 60 et 65 dB SPL ("Sound Pressure Level"). L'énergie contenue dans le signal vocal change en fonction de sa composition fréquentielle. En fait, la majeure partie de l'énergie est contenue dans la basse fréquence, approximativement de 300 Hz à 700 Hz, dans la zone où retentit le premier formant de la voyelle. Les consonnes contenues dans les fréquences moyennes à hautes (environ 2000 Hz) ne sont pas très dynamiques. A partir de ces faits, on peut expliquer que les hautes fréquences jouent un rôle dans l'intelligibilité des informations vocales, tandis que les basses fréquences fournissent l'énergie nécessaire pour obtenir des niveaux sonores audibles et intelligibles [28].

I.6.3 Le timbre :

Comme la hauteur, le timbre joue également un rôle déterminant dans la reconnaissance du locuteur. Il correspond à la couleur d'un discours personnel. Cette coloration est le résultat de la modulation et du filtrage de la source sonore (voix, larynx basique) à travers la cavité d'air supérieure de la gorge. L'effet de ces résonateurs sur l'onde fondamentale du col produit un faisceau harmonique spécifique et individuel. Il offre des indices au système auditif concernant des aspects du locuteur tels que son âge, son sexe et son origine [28].

I.6.4 L'amplitude :

L'amplitude d'un son représente la variation maximale de pression de l'air due aux oscillations, ce qui correspond au volume sonore. On peut objectivement exprimer l'amplitude d'une vibration en calculant les variations de pression de l'air (mesurées en Micron Bar et converties en watt/cm2). Cependant, on utilise plus couramment une unité de mesure relative, le décibel (dB), pour évaluer l'intensité d'un son [28].

I.7 Perception de la parole :

La perception auditive humaine désigne le mécanisme par lequel les individus sont capables d'interpréter et de comprendre les sons utilisés dans le langage. La partie externe de l'oreille capte le son. Cette pression sonore est amplifiée par le milieu. La partie de l'oreille qui est dédiée à la détection de l'équilibre et de la position envoie également des impulsions à travers. Ces impulsions

sont envoyées à la partie vestibulaire du système nerveux central. L'oreille humaine peut généralement entendre des sons avec des fréquences comprises entre 20 Hz et 20 kHz. Bien que la sensation d'ouïe nécessite une partie auditive intacte et fonctionnelle de système nerveux central ainsi qu'une oreille qui fonctionne, la surdité humaine (insensibilité extrême à sonore) survient le plus souvent en raison d'anomalies de l'oreille interne plutôt que des nerfs ou des voies du système auditif central [18].

I.8 Classification des sons de la parole :

La décomposition simplifiée du signal de parole fait apparaître deux types de sons :
Voisés tel que les voyelles et non voisés comme certaines consonnes. La différence entre ses deux sons [16] est que :

I.8.1 Un son voisé : est un son produit en faisant vibrer les cordes vocales [19] qui a la particularité de produire, en plus de sa fréquence fondamentale, un spectre riche en harmoniques [16].

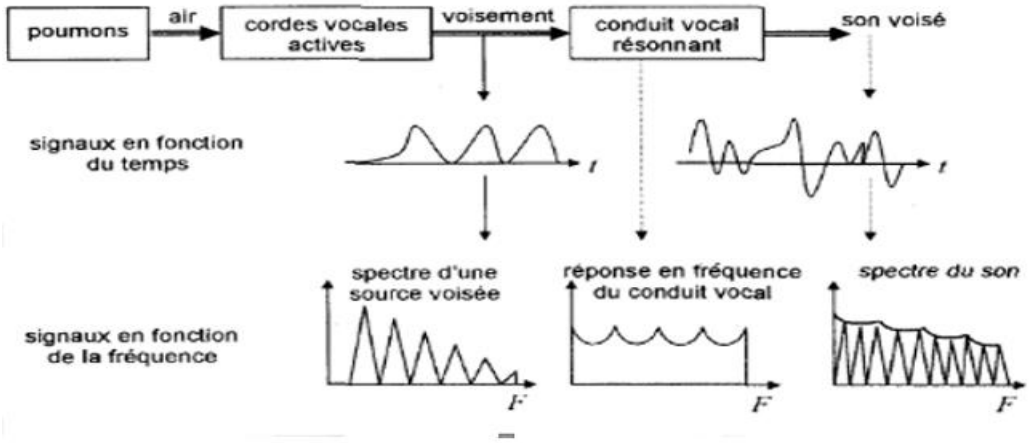

Fig. I.4 : processus de production de la parole dans le cas son voisé

I.8.2 Un son non-voisé ou dévoisé est un son produit sans faire vibrer les cordes vocales [19] et est le son de l'air s'écoulant de l'air des poumons (bruit), son spectre est similaire au bruit blanc [16]

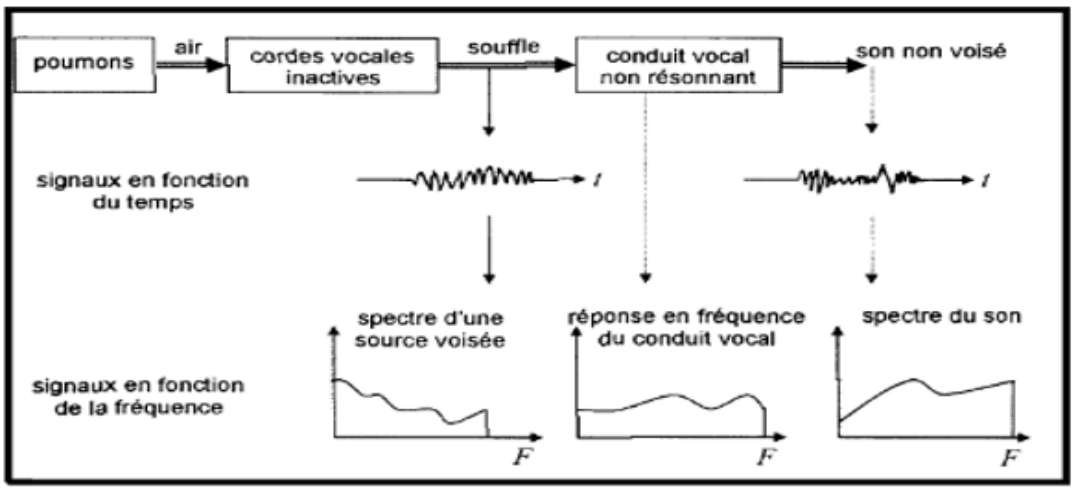

Fig. I.5 : Processus de production de la parole dans le cas son non voisé

I.9 Modélisation du mécanisme de la production de la parole :

Il convient de noter que chaque parole est le fruit d'une succession d'actions des poumons, du larynx et du conduit vocal, tous agissant de façon indépendante les uns des autres permettant à l'être humain de produire beaucoup de sons différents.

Fonctionnalités audio :

Le signal vocal audio comprend différents types d'informations sur le locuteur, par exemple :

"Haut niveau » Caractéristiques telles que l'accent, le contexte, le style de parole, l'état émotionnel de l'orateur, etc.

" Bas niveau" telles que la hauteur (la fréquence fondamentale des vibrations des cordes vocales), Intensités de fréquence, fréquences modales et leurs largeurs de bande, corrélations spectrales, spectre à courte portée et d'autres.

La quantité de données générées lors de la production de la parole est assez importante alors que là les caractéristiques essentielles de la parole générée évoluent assez lentement, nécessitent relativement moins de données pour représenter les caractéristiques du discours et de la personne qui l'a prononcé [18].

On peut représenter une source sonore en utilisant un modèle composé de séries d'impulsions périodiques pour les sons voisés, ou de bruit blanc pour les sons non voisés. Ces signaux excitent ensuite un filtre appelé filtre tous-pôles, dont les éléments représentent les caractéristiques du conduit vocal. [19].

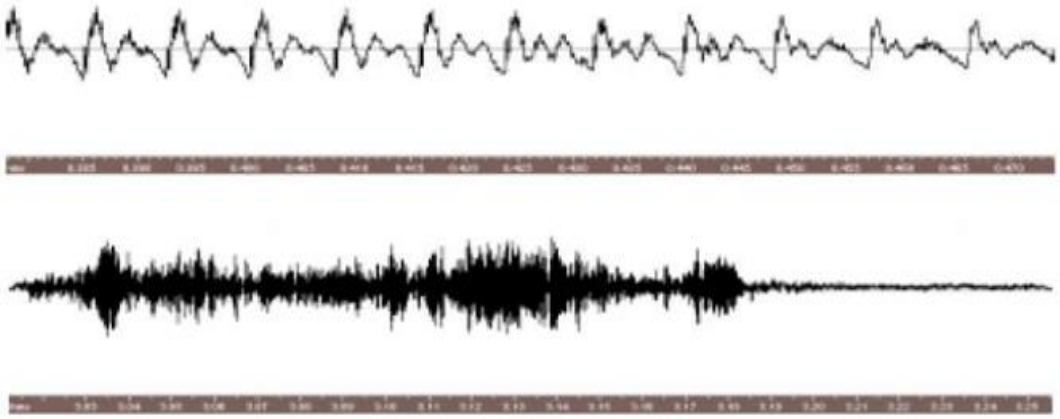

Fig. I.6 : Exemple de signal de son voise (haut) et son non voise (bas).

I.10 Propriétés statistiques du signal parole :

Le signal de parole est une manifestation spécifique d'un processus aléatoire non stationnaire, ce qui signifie que ses propriétés statistiques évoluent dans le temps. On suppose une quasi-stationnarité sur des intervalles de temps allant de 10 à 35 ms. [20].

I.11Traitement de la parole :

Traitement automatique de la parole (étude de signale de la parole) : est Ensemble de disciplines technologiques visant à capter, transmettre, identifier et synthétiser la parole [21]. Ces disciplines incluent spécifiquement la reconnaissance vocale, la synthèse vocale, l'identification du locuteur et la vérification du locuteur [22].

Le système automatique de traitement de la parole est représenté et résumé dans le schéma ci-dessous, par des tâches très importantes.

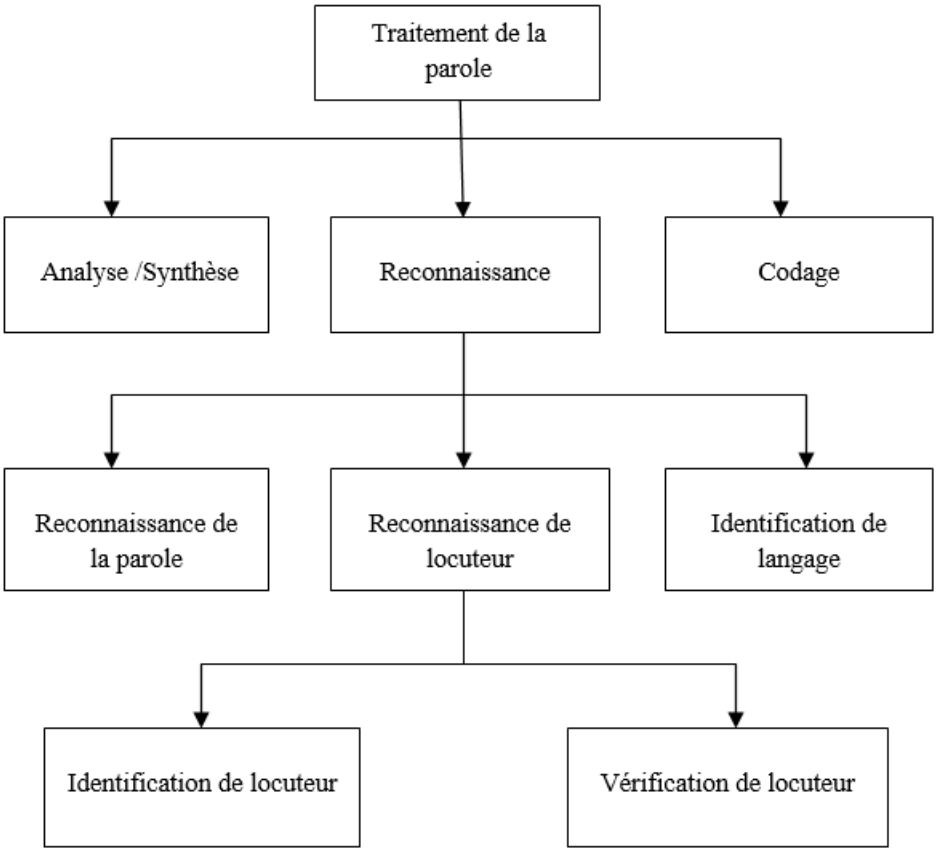

Fig. I.7 : Traitement de la parole [35]

I.11.1 Analyse et synthèse :

Les analyseurs de parole ont pour objectif de mettre en évidence les caractéristiques du signal de parole tel qu'il est produit, voire tel qu'il est perçu. Ils sont utilisés soit comme composants essentiels de systèmes de codage, de reconnaissance ou de synthèse de la parole, soit en tant qu'outils autonomes pour des applications spécialisées, telles que l'aide au diagnostic médical ou l'étude des langues, par contre les synthétiseurs sonores permettent de créer de la parole artificielle à partir d'un texte [02].

I.11.2 Codage :

Les codeurs jouent un rôle essentiel en permettant la transmission ou le stockage de la parole avec un débit réduit. Il est évident que pour obtenir des résultats optimaux dans ces tâches, il est crucial de prendre en considération les caractéristiques du signal étudié [02].

I.11.3 Reconnaissance :

La reconnaissance de la parole est une technique visant à reconnaître dans une suite de signaux sonores les phonèmes et les phrases prononcées par un locuteur, elle est basée sur une représentation paramétrique du signal [17]. Elle dépose un signal de parole et le déduis en une information qui est composée de 3 catégorie :

- Reconnaissance de la parole : analyser la voix capte pour la transcrire sous la forme d'un texte par une machine [23].
- Reconnaissance de la langue : détecter, identifier et reconnaitre les caractères (l'alphabet) dans les mots courants dans une langue [24].
- Reconnaissance de locuteur : qui est l'identification vérifié que la voix analysée correspond bien à la personne qui est sensée de produire [16].
- La vérification de locuteur : L'objectif est d'identifier parmi un nombre défini et limité de locuteurs préétablis, celui qui a produit le signal analysé [16].

I.12 Reconnaissance automatique du Locuteur

Les humains rêvent depuis longtemps de pouvoir manipuler des machines pour les rendre plus intelligentes. Malgré les efforts déployés, il n'a pas atteint l'objectif souhaité. La question pose Qu'est-ce qu'un système de reconnaissance vocale et comment peut ont créé un modélisateur d'un système de reconnaissance acoustique ?

Généralité :

La reconnaissance automatique du locuteur (RAL) est une branche du traitement du signal de la parole qui vise à identifier les caractéristiques vocales distinctives ou l'identité d'un individu à partir de sa voix. Ces caractéristiques sont utilisées pour créer une signature vocale qui permet d'authentifier la voix de chaque personne [25].

Les facteurs mentionnés peuvent en effet avoir un impact sur la performance des systèmes de reconnaissance automatique de la parole (RAL). Voici comment ces facteurs peuvent influencer les performances des systèmes de RAL [19] :

- L'état pathologique du Locuteur (maladie, émotion, …)
- Vieillissement.
- Facteurs socioculturels.
- Locuteurs non coopératifs.
- Conditions de prise de son.
- Bruit ambiant, …

I.13 Domaine d'applications des technologies de RAL :

La RAL est un ensemble de technique et technologie qui associent traitement du signal, intelligence artificielle et traitement de langage naturelle. Cette technique elle est utilisée dans plusieurs domaines [46] :

- Divertissement comme Shazam est un logiciel propriétaire de reconnaissance musicale de chansons.

- Marketing et Commerce tel que les télévisions, les hauts parleurs a commande vocale, téléphone intelligent sers a analyse du contenue sonore.
- Finance et Investissement : utilise les techniques de traitement de la voix et des algorithmes d'analyse les sentiments.
- Banque, Assurance et Institution financières.
- Sécurité : la voix peut être considérée comme une empreinte capable de détecter quelqu'un.
- Médicine par exemple la détection de Verus Covid 19 à partir de la voix. La performance de cette technologie de reconnaissance vocale dépend de plusieurs facteurs : la qualité sonore de l'enregistrement, les donnes d'entrainement utilisée pour la modélisation acoustique, le bruit …etc.

I.14 Les objectifs principaux de RAP :

La reconnaissance vocale a connu une utilisation considérable, en particulier dans les services à usage général tels que les services de télécommunications (utilisation des téléphones portables), afin d'atteindre les objectifs suivants [26] :

- Améliorer la fiabilité des systèmes en passant d'un mode de fonctionnement indépendant du locuteur à un mode totalement sécurisé et étroitement lié au locuteur.
- Accroître l'interactivité des systèmes homme-machine en intégrant le module de reconnaissance vocale à ces systèmes.
- Rendre la phase de reconnaissance de la parole robuste, en particulier dans des environnements bruyants.
- Tester l'adaptation de la reconnaissance vocale sur des applications réelles et avec un vaste vocabulaire.

I.15 Différentes tâches en RAL :

Les systèmes de Reconnaissance Automatiques du Locuteur (RAL) sont déterminés en deux taches : l'Identification Automatique du Locuteur (IAL) et la Vérification Automatique du Locuteur (VAL).

Dans cette section, nous allons décrire les principales tâches de la RAL qui sont l'IAL et la VAL [19].

I.15.1 Vérification (authentification) Automatique du Locuteur (VAL) :

La Vérification Automatique du Locuteur (VAL) vise à vérifier l'identité déclarée par un individu en comparant un signal vocal avec un modèle de référence préalablement appris par le système pour le locuteur supposé. Un système de VAL possède donc deux entrées : une identité déclarée et un échantillon de test. Le résultat de cette comparaison est considéré comme une mesure de similarité, qui est ensuite comparée à un seuil d'acceptation. Lorsque la mesure de similarité dépasse ce seuil, l'individu est accepté, sinon il est rejeté [25].

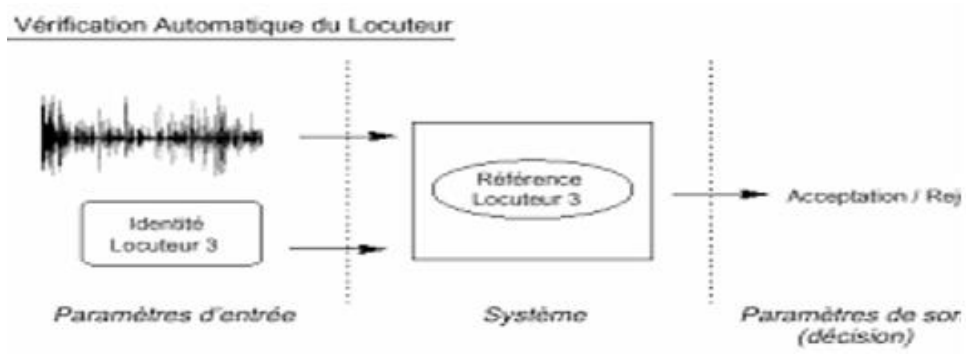

Fig. I.8 : La tâche de vérification automatique du locuteur

I.15.2 Identification Automatique du Locuteur (IAL)

L'Identification Automatique du Locuteur (IAL) est le processus qui vise à déterminer l'identité d'un locuteur dans un signal vocal (signal de test) en se basant sur un ensemble de locuteurs préalablement enregistrés dans le système. Le système évalue la similarité entre ce signal et les modèles de chaque locuteur dans la base de données. Deux conditions d'identification sont établies : le milieu fermé et le milieu ouvert [25].

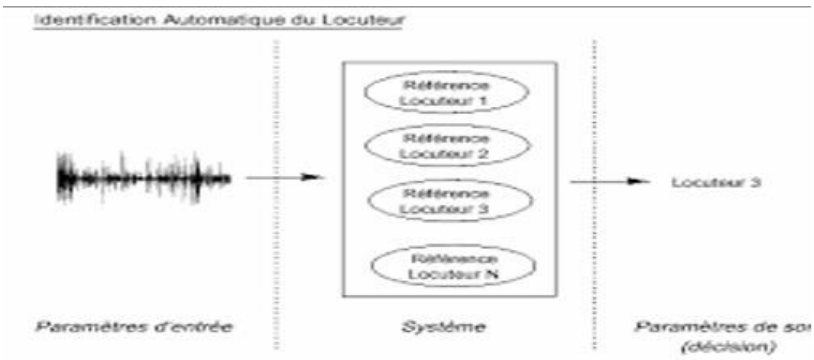

Fig. I.9 : La tâche de d'Identification Automatique du Locuteur

Note : l'identification et la vérification du locuteur sont les taches les plus essentielles dans la reconnaissance automatique de locuteur (RAL) [27].

I.16 Structure d'un système de reconnaissance automatique du locuteur :

La reconnaissance automatique du locuteur se décompose en trois phases principales. Tout d'abord, il y a la phase de paramétrisation qui consiste en une analyse acoustique. Ensuite, vient la phase de modélisation, suivie de la phase de décision. De plus, un système de reconnaissance automatique du locuteur possède deux modes de fonctionnement [25] :

I.16.1 Apprentissage : Dans ce cas, un modèle est calculé pour chaque locuteur "client" du système, qui servira ensuite de référence pour les futures tâches de reconnaissance [25].

I.16.2 Test : Dans ce processus, une étape cruciale est la reconnaissance (vérification, identification, etc.). À la suite de cette étape, le système génère une réponse : une identité dans le

cas de la tâche d'identification, ou une décision d'accès/rejet dans le cas de la vérification [25]. Le schéma ci-dessus illustre la structure d'un système de vérification de locuteur :

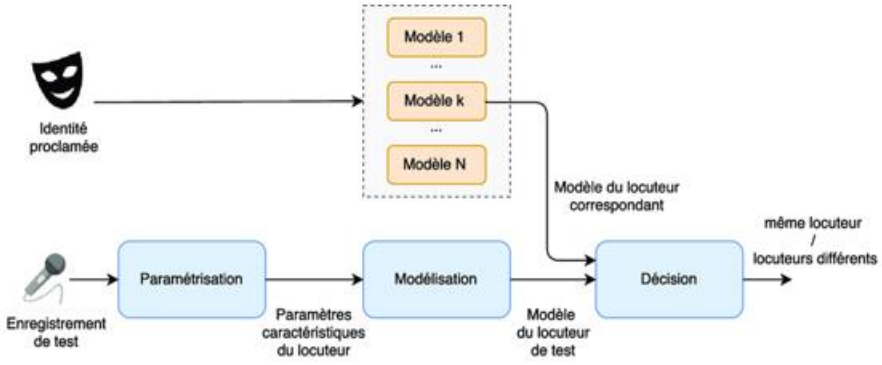

Fig. I.10 : La structure d'un système de vérification du locuteur

I.17 Paramétrisation du signal de parole :

L'objectif de cette étape est de recueillir des paramètres distinctifs de la parole d'un individu spécifique [19]. Ces données renferment des informations complexes, souvent redondantes et mêlées à du bruit. Le module de paramétrisation chargé du traitement du signal acoustique doit atteindre plusieurs objectifs [02] :

- Isoler le signal du bruit
- Extraire les informations pertinentes pour la reconnaissance
- Convertir les données brutes en un format directement exploitable par le système.

I.18 Modélisation :

Le module de modélisation utilise les données provenant du module de paramétrisation pour générer la représentation d'une personne [02]. Il permet une modélisation précise des caractéristiques spectrales des voix des locuteurs et sa mise en œuvre est relativement facile [25].

I.19 La décision :

Le stade de décision concerne l'identification finale du locuteur. La stratégie employée à ce stade dépendra étroitement de la méthode de modélisation sélectionnée ainsi que du type d'application : identification ou vérification [25].

I.20 Conclusion :

Au cours de ce chapitre, nous avons examiné le processus de production de la parole, ainsi que la structure générale d'un système de Reconnaissance Automatique de la Parole (RAL) et les diverses tâches associées à ce domaine, ainsi que les étapes correspondantes. Dans le prochain chapitre, nous explorerons les différents outils de traitement de la parole et présenterons également quelques méthodes de paramétrisation du signal vocal.

Chapitre II :

Traitement numérique du signal vocal

II.1 Introduction :

Le premier étage d'un système RAP est d'analyser et paramétriser le signal [26]. L'analyse de la parole revêt une importance cruciale dans toutes les applications de synthèse, de codage et de reconnaissance. Elle joue un rôle essentiel en fournissant une description détaillée du signal acoustique et en extrayant les paramètres pertinents [29]. Dans ce chapitre on va comprendre la base du traitement de la parole et on s'appuie sur les caractéristiques spécifiques du signal de parole pour déterminer la méthode d'extraction des paramètres la plus adaptée.

Généralités sur les signaux vocaux

II.2 Définition d'un signal

Un signal est une manifestation physique de l'information, qui est transmise de sa source à son destinataire [44]. La théorie du signal vise à fournir une description mathématique de ces signaux, permettant ainsi l'analyse, la conception et la caractérisation des systèmes de traitement de l'information [02].

II.3 Définition d'un bruit

Le bruit est un phénomène physique qui vient perturber ou gêner la transmission ou l'interprétation d'un signal [02]. Le niveau sonore est mesuré en décibels (dB). Afin de prendre en compte la perception réelle par l'oreille humaine, on utilise le décibel pondéré A, abrégé dB(A).

• 0 dB(A) correspond au bruit le plus faible perceptible par l'oreille humaine.
• 50 dB(A) représentent le niveau habituel lors d'une conversation.
• 80 dB(A) est le seuil de nocivité pour une exposition de 8 heures par jour.
• 120 dB(A) correspondent à un niveau de bruit provoquant une sensation douloureuse.

Lorsque les niveaux sonores sont très élevés, l'oreille humaine ne perçoit pas les bruits de la même manière. Pour tenir compte de cet effet, on utilise le décibel pondéré C, noté dB(C) [30].

Il convient de noter que le sonomètre est l'instrument de mesure de base utilisé pour évaluer le bruit [30].

II.4 Le traitement de signal

La théorie du signal se consacre à la modélisation mathématique des signaux [02]. Elle est une discipline technique qui s'appuie sur les connaissances en électronique, informatique et physique appliquée, et vise à élaborer ou interpréter les signaux porteurs d'informations. Son champ

d'application englobe les domaines de la transmission et de l'exploitation des informations véhiculées par ces signaux [44].

II.5 Les fonctions principales du traitement de signal

Les principales fonctions du traitement de signal sont :

II.5.1 Cree : Élaboration de signaux

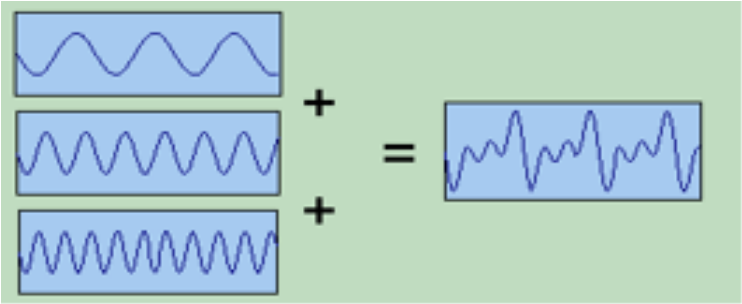

Fig. II.1 : Création de signal

II.5.1.1 La synthèse : création de signaux par combinaison de signaux élémentaires [31]

II.5.1.2 La modulation et le changement de fréquence : Ces moyens sont principalement utilisés pour ajuster un signal aux caractéristiques fréquentielles d'une voie de transmission, d'un filtre d'analyse ou d'un dispositif d'enregistrement [31].

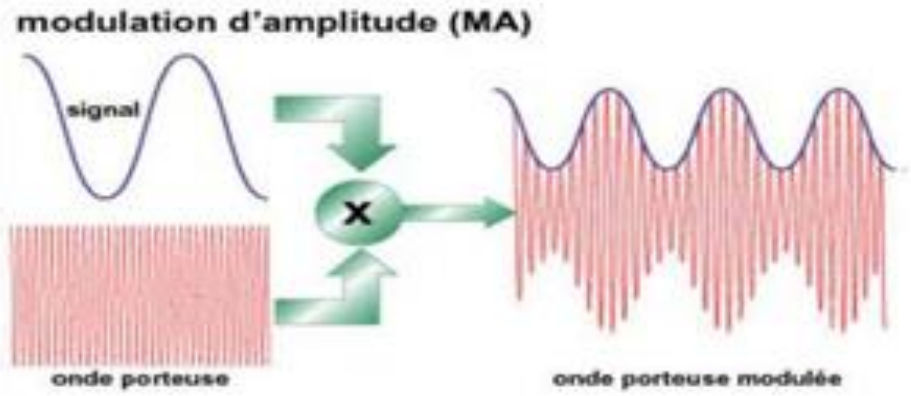

Fig. II.2 : Modulation de signal

II.5.2 L'analyse : Analyser : Interprétation des signaux

II.5.2.1 La détection : isoler les composantes utiles d'un signal complexe, extraction du signal d'un bruit de fond [31].

II.5.2.2 Identification : Le classement du signal implique l'identification de pathologies sur un signal ECG, la reconnaissance de la parole, et d'autres applications similaires [31].

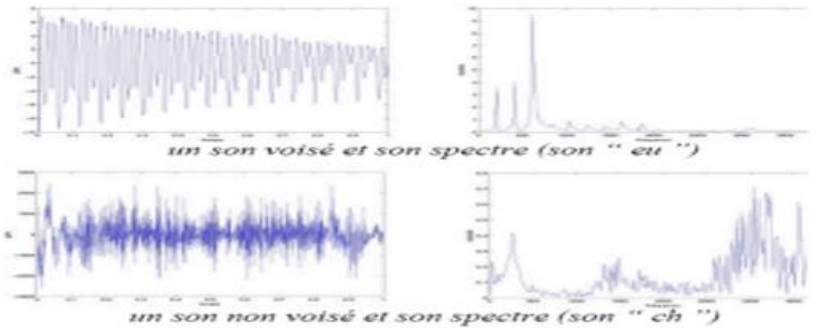

Fig. II.3 : Identification d'un signal voise et non voise

II.5.3 Transformer : adapter un signal aux besoins

II.5.3.1 Filtrage : élimination de certaines composantes

- Détection de craquements sur un enregistrement,
- Détection de bruit sur une image,
- Annulation d'écho, etc.

II.5.3.2 Codage/compression (Jpeg, mp3, mpeg4, etc.) : outre sa fonction de traduction en langage numérique, est utilisé soit pour lutter contre le bruit de fond, soit pour tenter de réaliser des économies de largeur de bande ou de mémoire d'ordinateur [31].

II.6 Classification de signal

On peut envisager plusieurs modes de classification pour les signaux suivant leurs propriétés [32].

II.6.1 Classification phénoménologique

On considère la nature de l'évolution du signal en fonction du temps. Il apparaît deux types de Signaux. Les signaux déterministes (ou certains) et les signaux aléatoires (probabilistes) [02] [32].Le signal phénoménologique est représenté dans le schéma ci-dessus [45] :

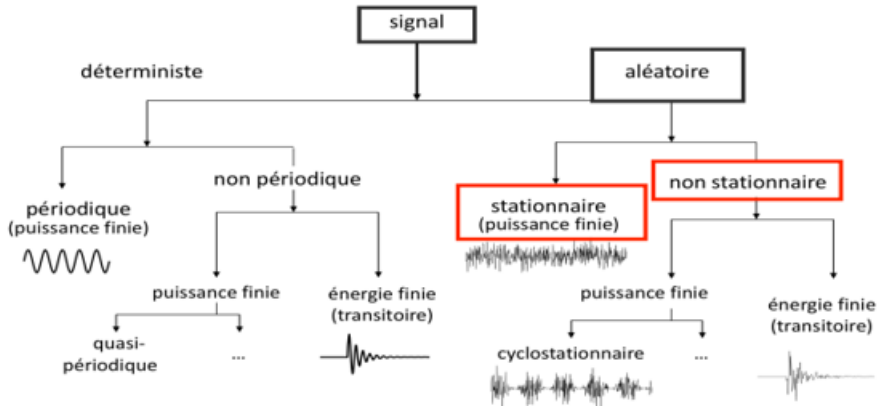

Fig. II.4 : Classification des signaux phénoménologique

II.6.1.1 Les signaux déterministes : Certains signaux, tels que les signaux périodiques, les signaux transitoires, les signaux pseudo-aléatoires, etc., peuvent être parfaitement modélisés par une fonction mathématique qui représente leur évolution dans le temps [32].

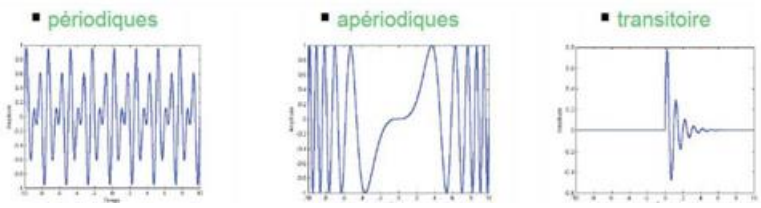

Fig. II.5 : Classification des signaux déterministes

II.6.1.2 Les signaux aléatoires :(probabilistes) Leur comportement temporel est caractérisé par une imprévisibilité. Afin de les décrire, il est nécessaire de recourir à leurs propriétés statistiques. Lorsque ces propriétés statistiques demeurent constantes dans le temps, on les qualifie de signaux stationnaires [32].

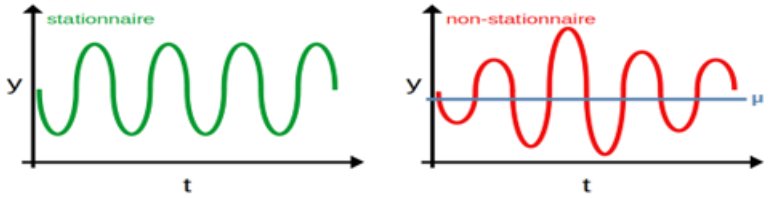

Fig. II.6 : Classification des signaux

II.6.2 Classification énergétique

On considère l'énergie des signaux. On distingue :

II.6.2.1 Les signaux à énergie finie : il possède une puissance moyenne nulle et une énergie finie.

II.6.2.2 Les signaux à puissance moyenne finie : il possède une énergie infinie et sont donc physiquement irréalisable [32].

II.6.3 Classification morphologique

Le temps joue un rôle crucial dans la classification des signaux. Le traitement numérique des signaux permet de distinguer différentes catégories de signaux, ce qui conduit à l'identification de quatre classes distinctes [02] [32] :

II.6.3.1 Les signaux analogiques dont l'amplitude et le temps sont continus

II.6.3.2 Les signaux quantifiés dont l'amplitude est discrète et le temps continu

II.6.3.3 Les signaux échantillonnés dont l'amplitude est continue et le temps discret

II.6.3.4 Les signaux numériques dont l'amplitude et le temps sont discrets

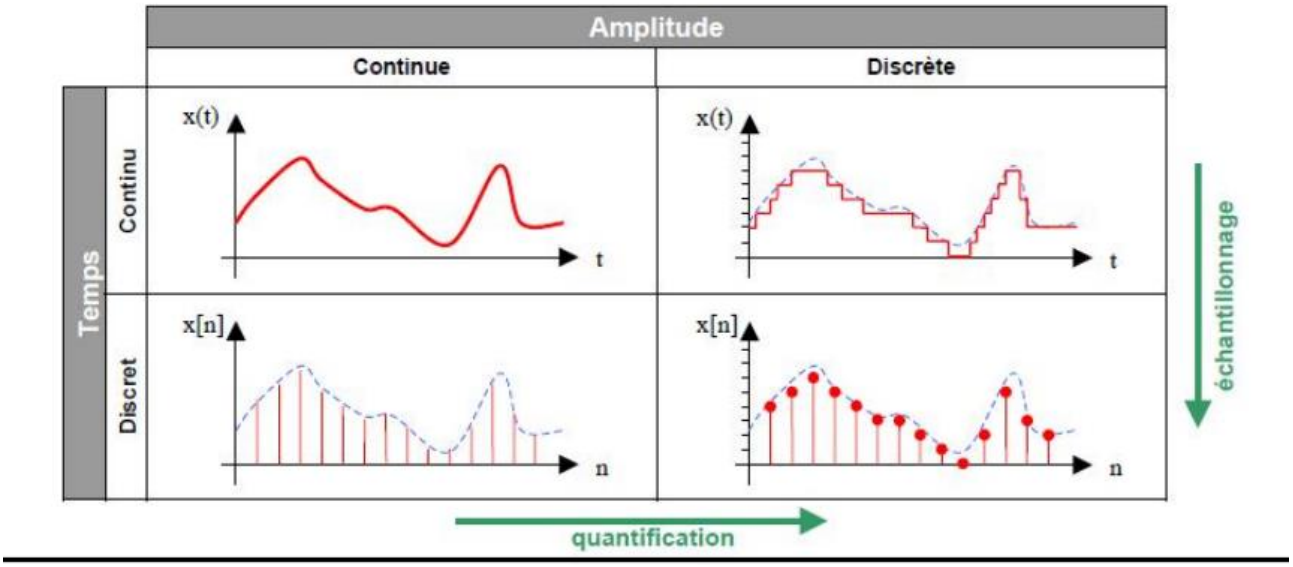

Fig. II.7 : Classification morphologique des signaux

II.7 Composition d'un SRAP :

Un programme typique de reconnaissance automatique de la parole est compose des étapes suivantes [49] :

- Prétraitement du signal qui inclut la numérisation, l'échantillonnage, la quantification et le codage
- Extraction des paramètres caractéristique (MFCCs, LPCs).
- Méthodes de classification et de reconnaissance (HMM, DTW, SVM, AQ, GMM)
- Evaluation de la reconnaissance.

Fig. II.8 : Composition de base des SRAP [49]

II.8 Analyse du signal parole

L'analyse et la synthèse sont deux activités complémentaires. L'analyse permet de fournir une description détaillée du signal acoustique, que la synthèse utilise ensuite pour le reproduire. L'analyse acoustique revêt une importance primordiale dans le traitement du signal sonore pour la

réalisation de systèmes de synthèse, de compréhension ou de reconnaissance de la parole de haute qualité. Cette étape consiste à extraire à partir du signal vocal un ensemble de paramètres pertinents, discriminants et robustes qui permettent de le représenter de manière précise [34].

II.9 Le prétraitement du signal

Le prétraitement acoustique vise à extraire les informations essentielles du signal audio afin de faciliter la reconnaissance de la parole. Cette étape implique le calcul d'une séquence de vecteurs numériques contenant les informations pertinentes pour le traitement automatique du langage parlé [39].

Le signal vocal est de nature analogique (continue en temps), en revanche, les systèmes de traitement sont des systèmes numériques, d'où la nécessité de cette phase [19].

II.9.1 La numérisation

La numérisation consiste à convertir un signal analogique contenant un nombre infini d'amplitudes en un signal numérique contenant un nombre fini de valeurs [40] .

Le passage de l'analogique au numérique repose sur trois étapes [32] :

- L'échantillonnage pour rendre le signal discret.
- La quantification pour associer à chaque échantillon une valeur.
- Le codage pour associer un code à chaque valeur.

II.9.1.1 l'échantillonnage

L'échantillonnage est l'opération qui consiste à prélever des échantillons du signal a temps continu x(t) pour obtenir un signal temps discret x (nT_e), cela implique la conversion du signal en une séquence de nombres afin de faciliter sa mémorisation, sa transmission ou son traitement ultérieur [41].

Le signal analogique est découpé en échantillons. Le nombre d'échantillons par seconde représente la fréquence d'échantillonnage fe (Hz), qui est elle-même l'inverse de la période d'échantillonnage Te. Le choix de la fréquence d'échantillonnage doit être adéquate (prélever assez de valeurs pour ne pas perdre l'information contenue) c'est-à-dire respecter le théorème de Shannon [19].

$$f_e \geq 2 f_{max} \qquad \text{II.1}$$

f_e : La fréquence d'échantillonnage

f_{max}: La fréquence maximale du signal

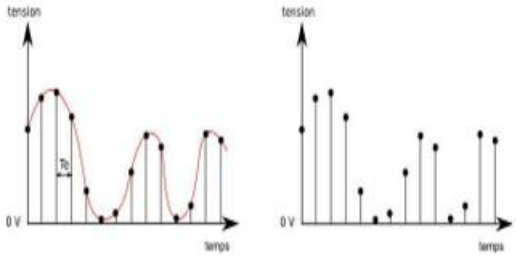

Fig. II.9 : échantillonnage.

II.9.1.2 la quantification

La quantification d'un signal consiste à affecter une valeur numérique à chaque échantillon prélevé [19].

La quantification d'un signal consiste à attribuer des valeurs d'amplitude à des intervalles réguliers sur une échelle prédéfinie. Chaque échantillon correspond alors à un nombre binaire unique. L'utilisation d'une quantification surn bits permet d'utiliser 2^n valeurs différentes. La quantification a pour effet d'arrondir l'amplitude de chaque échantillon à l'une de ces 2^n valeurs. Cette opération arrondit l'amplitude de chaque échantillon à l'une de ces valeurs. Ainsi, le nombre de bits de quantification détermine la précision en amplitude ou la dynamique de la conversion, tandis que la fréquence d'échantillonnage influence la précision temporelle de la conversion [42].

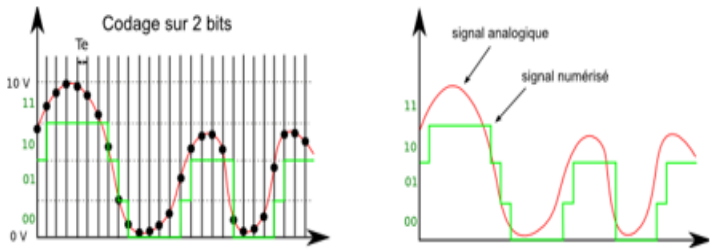

Fig. II.10 : quantification

II.9.1.3 Le Codage

On appelle codage la transformation des différentes valeurs quantifiées en langage binaire qui permet le traitement du signal sur machine [19]. C 'est à dire, il consiste à associer à un ensemble de valeurs discrètes un code composé d'éléments binaires [32] [43].

Les codes les plus connus : code binaire, code DCB, code Gray…etc.

II.10 Les différentes méthodes d'analyse du signal de parole

Il existe deux approches distinctes pour l'analyse : l'approche directe et l'approche indirecte. L'approche directe se concentre sur les grandeurs anatomiques, tandis que l'approche indirecte se penche sur les signaux électriques correspondants [29].

II.10.1 La méthode directe

Cette approche implique l'étude du comportement des organes du système vocal tels que la glotte, le larynx et les cordes vocales. Pour réaliser cette analyse, différentes techniques sont utilisées.

L'objectif principal est d'obtenir une représentation de la fréquence fondamentale. Les techniques utilisées comprennent les suivantes [29] :

- La glottographie photoélectrique
- La glottographie électrique
- La glottographie ultrasonore

II.10.2 La méthode indirecte

Nous nous intéressons particulièrement à cette méthode car elle permet de traiter le signal électrique capté par un microphone. En fonction des paramètres à extraire, il est possible de traiter ou d'analyser le signal soit temporellement, soit fréquentielle ment [29]. Dans la suite, nous allons expliquer les principes des différentes méthodes utilisées.

II.10.2.1 L'analyse temporelle

Elle génère des paramètres en traitant la représentation temporelle du signal.

II.10.2.1.1 analyse par l'énergie

Selon le type et l'amplitude du son produit, L'énergie à court terme sert à détecter les silences, elle est élevée pendant l'activité vocale.

Les sons voisés présentent une valeur d'énergie plus grande à celle des sons non voisés [29].

L'énergie E en décibel (dB) est définie par :

$$E = 10Log \sum_{n=0}^{N-1} |x(n)|^2 \qquad \text{II.2}$$

II.10.2.1.2 Taux de passage par zéro (TPZ)

Le taux de passage par zéro Pour un signal échantillonné, il y a passage par zéro lorsque deux échantillons successifs sont de signes opposés, il est particulièrement utile pour distinguer une zone voisée d'une zone non voisée [29]. Le taux de passage par zéro à court terme peut être estimé par la formule suivante :

$$\boldsymbol{Zc = \frac{1}{2N} \sum_{n=0}^{N-1} \left| sgn\left(x(n)\right) - sgn\left(x(n-1)\right) \right|} \qquad \text{II.3}$$

Le calcul de l'indice TPZ est très simple. Cette caractéristique est utilisée pour classer les signaux en sons voisés et non voisés.

Les sons non voisés ont un taux de passage par zéro supérieur à celui des sons voisés [19].

II.10.2.1.3 L'autocorrélation :

L'autocorrélation est la convolution du signal avec lui-même, et son estimation Elle est défini par la formule suivante :

$$\emptyset_{xx}(K) = \frac{1}{N-K}\sum_{n=0}^{N-1-K}[x(n) \times x(n+K)] \quad \text{II.4}$$

L'idée d'utiliser la fonction d'autocorrélation est de déterminer à quel point deux échantillons successifs d'un signal sont similaires. Parmi ses autres applications, on peut s'y référer pour estimer la fréquence de la fréquence fondamentale (ou pitch) [19].

II.10.2.2. Analyse fréquentielle

Nous passons maintenant à la méthode la plus couramment utilisée pour le traitement du signal vocal, qui est l'analyse fréquentielle. Cette méthode permet d'extraire les propriétés spectrales d'un signal vocal, qui revêtent une grande importance pour la perception auditive. L'analyse spectrale vise à caractériser la distribution de l'énergie ou de la puissance d'un signal en fonction de la fréquence [29]. Parmi les outils utilisés pour l'analyse spectrale du signal vocal, on peut citer :

II.10.2.2.1 Analyse par Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle d'un signal, sous la forme d'un spectrogramme. Le spectrogramme est une représentation tridimensionnelle

- Le temps est représenté sur l'axe des abscisses.
- La fréquence sur l'axe des ordonnées.
- L'énergie apparaît sous la forme de niveau de gris pour un temps et une fréquence donnée.

Il existe deux types de spectrogrammes :

- Les spectrogrammes à bandes larges : ils sont générés en utilisant des fenêtres de courte durée (généralement 10 ms). Ces spectrogrammes mettent en évidence l'enveloppe spectrale du signal, également connue sous le nom de formants, où les périodes de sonorité sont visibles sous la forme de bandes verticales plus sombres [19].
- Les spectrogrammes à bandes étroites : (moins utilisés) sont obtenus avec des fenêtres de 30

À 40 ms, et ils offrent une bonne résolution au niveau de la fréquence, où les harmoniques du signal dans les zones voisées apparaissent comme des bandes de fréquence horizontales [29].

II.10.2.2.2 Analyse par la transformée de Fourier à court terme (TFCT)

La transformée de Fourier (TF) tout court, a suscité beaucoup d'intérêt et reste un outil essentiel dans le traitement du signal

En réalité, l'analyse de Fourier permet de déterminer les différentes fréquences présentes dans un signal, c'est-à-dire son spectre de fréquences. Cependant, elle ne fournit pas d'informations sur le moment précis où ces fréquences ont été émises, ce qui signifie que la transformation de Fourier

donne une vision globale plutôt que locale du signal. Cette perte de localité constitue un inconvénient lors de l'analyse de signaux non stationnaires ou quasi-stationnaires, tels que les signaux vocaux. La transformation de Fourier est définie par :

$$\mathbf{x(f)} = \int_{-\infty}^{+\infty} \mathbf{x(t)e^{-j2\pi ft}\,dt} \qquad \text{II.5}$$

La TF discrète :

$$\boldsymbol{x(K)} = \sum_{\mathbf{n=0}}^{\mathbf{N-1}} \mathbf{x(n)e^{-j\frac{2\pi kn}{N}}} \qquad \text{II.6}$$

x(n) : signal échantillonnée

N : nombre de point de la suit temporel x(n)

Afin de réduire le temps de calcul de la TFD, on applique la transformée de Fourier rapide (FFT) [19].

II.10.2.2.3 La transformée de Fourier Rapide (FFT)

La transformée de Fourier rapide est simplement une TFD calculée à partir d'un algorithme qui passe du domaine temporel vers le domaine spectral où le signal parole (résultat de convolution de la source par le conduit vocale) devient un simple produit. La FFT est algorithme rapide qui est définie pour un signal x de N échantillons par [35] :

$$\mathbf{X_n} = \sum_{\mathbf{K=0}}^{\mathbf{N-1}} \mathbf{X_K\, e^{\frac{-2j\pi kn}{N}}} \qquad \text{II.7}$$

Pour $0 \leq K \leq N-1$

Le résultat obtenu est considéré comme étant le spectre du signal.

II.10.2.2.4 Analyse Cepstrale

Le spectre donné par la FFT contient des renseignements sur la source et le conduit vocal, mais leur intermodulation fait qu'il est difficile de mesurer la f0 (fréquence fondamentale) et des fréquences des formants qui caractérisent respectivement la source et le conduit.

Le lissage cepstre est une méthode visant à séparer la contribution de la source et celle du conduit par dé convolution. Pour cela nous supposons que le signal vocal (n) est produit par un signal excitateur (n) (source glottique) traversant un système linéaire passif de réponse impulsionnelle (n) (conduit). Avec ces suppositions, nous pouvons écrire

$$\hat{\mathbf{x}}(\mathbf{n}) = \mathbf{g}(\mathbf{n}) \otimes \mathbf{b}(\mathbf{n}) \qquad \text{II.8}$$

L'opérateur de convolution ($\otimes$) correspond à un opérateur d'addition (+). Cette transposition homomorphe se fait sur les étapes suivantes :

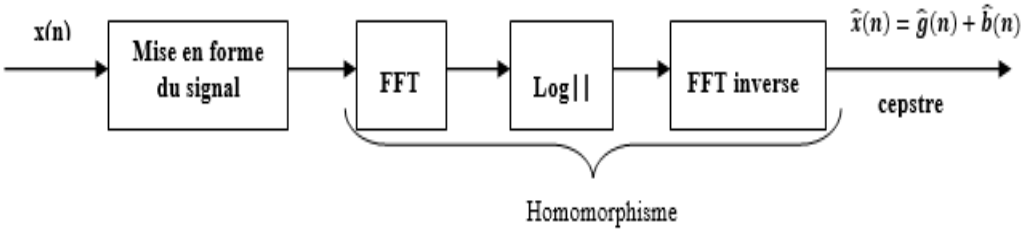

Fig. II.11 : Analyse homomorphique de la parole

$\hat{x}(n)$ sont les coefficients cepstraux approchés, prenant des valeurs dans un domaine pseudo-temporel réel appelé domaine fréquentiel. La structure de la parole et les hypothèses sur la source d'excitation et le conduit vocal

$\hat{g}(n)$ Se réduit théoriquement à une séquence d'impulsions de période n 0 (n 0 correspond à la fréquence fondamentale f 0).

$\hat{b}(n)$ Décroît rapidement avec n (en 1/n) et devient négligeable lorsque n>n 0 [19].

Il existe deux principaux types de coefficients cepstraux

II.10.2.2.4.1 Les coefficients LPCs

Le Linear Prédictive Coding (LPC) est une méthode de codage et de représentation de la parole qui repose sur la notion que le système phonatoire peut être modélisé par un filtre linéaire. Ce filtre est excité par un train d'impulsions pour les sons voisés et par un signal aléatoire pour les sons non voisés. L'objectif est de prédire le signal à l'instant n en utilisant les échantillons précédents (p échantillons). Le schéma ci-dessus illustre le diagramme de bloc du LPC [04] [36].

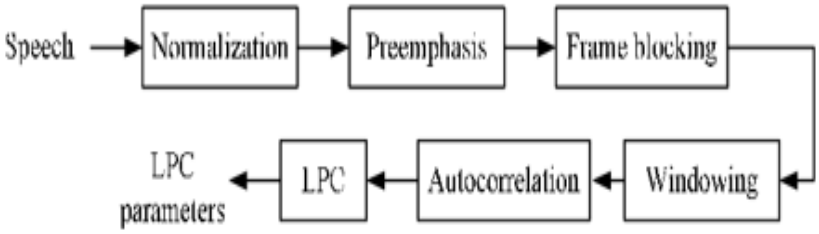

Fig. II.12 : Diagramme en blocs de LPC

II.10.2.2.4.2 Les coefficients MFCCs

Les coefficients MFCC (Mel Frequency cepstral coefficients) sont basés sur une échelle de perception non-linéaire qui correspond à la distribution fréquentielle de l'oreille humaine. Le modèle de l'échelle de Mel défini précédemment sera utilisé dans cette méthode [35] [18]. Les principales étapes de calcul de ces coefficients sont illustrées dans la figure suivante

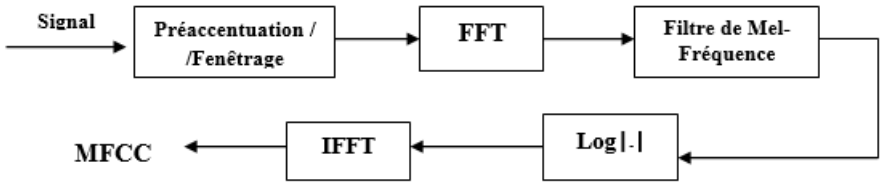

Fig. II.13 : Diagramme en blocs de MFCC

II. 10.2.2.4.2.1 Préaccentuation

On effectue une préaccentuation du signal au début du traitement [19].Le signal de parole continu est découpé en trames de N échantillons, où les trames sont séparées par une distance M (M < N). La première trame est formée des N premiers échantillons. La deuxième trame commence M échantillons après la première trame et chevauche N-M échantillons. De manière similaire, la troisième trame commence 2M échantillons après la première trame (ou M échantillons après la deuxième trame) et chevauche N - 2M échantillons. Ce processus se répète jusqu'à ce que toute la parole soit prise en compte dans une ou plusieurs trames. Les valeurs typiques pour N et M sont N = 256 et M = 100 [18].

II. 10.2.2.4.2.2 Fenêtrage

Dans cette étape on utilise une fenêtre de type Hamming pour décomposer le signal en un ensemble de segments d'échantillons

Afin de minimiser la distorsion spectrale lors de la transformation du domaine temporelle vers le domaine fréquentiel, on effectue un fenêtrage qui tend à rendre le signal nul au début et à la fin de chaque trame. La fenêtre de Hamming [35] :

$$\mathbf{w\,(n) = 0.54 - 0.46(cos(\frac{2\pi n}{N-1}))0 \;\leq\; n \;\leq\; N-1} \qquad \text{II.9}$$

II.10.2.2.4.2.3 FFT

Faste Fourier Transformé est la prochaine étape de traitement, qui convertit chaque image de N échantillons du domaine temporel au domaine fréquentiel. La FFT est un algorithme rapidepour implémenter la Transformée de Fourier Discrète (DFT) qui est définie sur l'ensemble de N échantillons (xn), comme suit :

$$\mathbf{X_k = \sum_{N+0}^{N-1} x_n\, e^{-\frac{j2\pi kn}{N}} \quad K=0,\ 1,\ 2\ \ldots\ N\text{-}1} \qquad \text{II.10}$$

Le résultat après cette étape est souvent appelé Spectrum [18].

II.10.2.2.4.2.4 Application de l'échelle de Mel

Le spectre présente plusieurs fluctuations, et afin de réduire la taille des vecteurs cepstraux on ne s'intéresse seulement à l'enveloppe du spectre. Pour réaliser ce lissage nous multiplions le spectre par un blanc de filtre tenant compte de la réponse acoustique que l'oreille humaine.

Un banc du filtre est une série de filtres à bande passante repartie d'une façon équidistante dans l'échelle de Mel. La formule approchée suivante pour calculer les Mels pour une fréquence donnée f en Hz [35] :

$$\mathbf{Mel(f) = 2595Log_{10}\left(1 + \frac{f}{700}\right)} \qquad \text{II.11}$$

II.10.2.2.4.2.5 Logarithme(Log)

Nous prenons le logarithme de cette enveloppe et nous multiplions chaque coefficient par 20 afin d'obtenir l'enveloppe spectrale en dB [35].

Le logarithme s'appliqué pour transformer la multiplication en addition [19].

II.10.2.2.4.2.6 Transformé en cosinus discrète (DCT)

Dans cette dernière étape, nous reconvertissons le spectre log Mel dans le temps. Le résultat s'appelle le Mel coefficients de cepstre de fréquence (MFCC). Pour revenir à l'espace temporel, nous utilisons la transformée de Fourier inverse (FFT inverse). Étant donné que nous ne travaillons qu'avec la partie réelle du signal, la transformée cosinus discrète (DCT) peut être utilisée pour inverser facilement la transformation. Si nous définissons 'K' comme le nombre de filtres et 'L' comme le nombre de coefficients souhaités, les coefficients MFCC seront obtenus en utilisant ces paramètres [18] :

$$\hat{c}(n) = \sum_{k=1}^{k} (\log \hat{E}_k) \cos[\, n(k - \tfrac{1}{2}) \tfrac{\pi}{k}] \qquad \textbf{Ou n=1, 2 ...L} \qquad \text{II.12}$$

$\hat{E}_k$: L'énergie a la sortie des filtre k=1,2,…k

On note que le coefficient $\hat{c}_0$ a été écarté, cela en raison du fait qu'il représente l'énergie moyenne dans la trame de la parole et ne contribue pas de manière significative dans les applications RAP. [19]

II.10.2.3 Analyse temps-fréquentielle

II.10.2.3.1 Transformée en ondelettes

La transformée en ondelette est un opérateur qui remplace la sinusoïde de la transformée de Fourier par une famille de translations et de dilatations d'une même fonction appelée ondelette [29]. L'analyse d'un signal par l'ondelette est effectuée à l'aide d'une fonction spécifique appelée "ondelette mère" Ψ. Le terme "ondelette" est utilisé car cette fonction est oscillante, semblable à une onde. Elle est positionnée dans le domaine temporel afin de sélectionner la partie du signal à traiter [37].

Si x (t) est une fonction réelle de variable réelle la transformée en ondelettes de f est :

$$\mathbf{g\ (a,b)}=\frac{\mathbf{1}}{\sqrt{\boldsymbol{a}}}\int_{\boldsymbol{t}=-\infty}^{\boldsymbol{t}=+\infty}\boldsymbol{x}(\boldsymbol{t})\boldsymbol{\psi}_{\boldsymbol{a,b}}\,(\boldsymbol{t})\boldsymbol{dt} \qquad \text{II.13}$$

a≠0 La fonction $\psi_{a,b}$(t) est obtenue par translation et dilatation d'une fonction particulière appelée ondelette mère :

$$\boldsymbol{\psi}_{\mathbf{a,b}}(\mathbf{t})=\boldsymbol{\psi}\left(\frac{\mathbf{t-b}}{\mathbf{a}}\right) \qquad \text{II.14}$$

b détermine la position et a donné l'échelle.

Cas d'un signal : a est la fréquence et b le temps.

L'objectif principal de l'analyse en ondelettes est d'ajuster la taille de la fenêtre d'analyse en fonction des variations du signal étudié. L'ondelette est une fonction qui constitue un autre outil de traitement du signal, caractérisée par sa localisation en temps et en fréquence [29].

II.10.2.3.2 La transformée en ondelette discrète

La transformée en ondelettes discrètes (DWT) permet de décomposer le signal en un ensemble d'ondelettes mutuellement orthogonales, ce qui constitue la principale différence par rapport à la transformée en ondelettes continue ou à son implémentation en séries discrètes dans le temps, parfois appelée transformée en ondelettes continue discrète dans le temps (DT-CWT).

L'ondelette peut être construite à partir d'une fonction d'échelle qui décrit les propriétés d'échelle du signal. La condition d'orthogonalité avec les translations discrètes implique certaines contraintes mathématiques, telles que l'équation de dilatation.

$$\emptyset(\mathbf{x})=\sum_{\mathbf{k}=-\infty}^{+\infty}\mathbf{a_k}\,\emptyset(\mathbf{S_x}-\mathbf{k}) \qquad \text{II.15}$$

Où S est un facteur d'échelle (généralement égal à 2). De plus, il est nécessaire de normaliser la zone entre les fonctions et d'assurer l'orthogonalité par translation de la fonction d'échelle, ce qui signifie que nous avons [38] :

$$\int_{-\infty}^{+\infty}\emptyset(\mathbf{x})\emptyset(\mathbf{x}+\mathbf{l})\mathbf{dx}=\boldsymbol{\delta}_{\mathbf{0,l}} \qquad \text{II.16}$$

II.11 Les différentes méthodes de reconnaissance automatique de locuteur

Âpres avoir vu les différentes méthodes d'analyse de signal vocal, maintenant nous allons voir les différentes méthodes de reconnaissance ou de comparaison nous allons citer quelques méthodes reconnus dans le domaine de reconnaissance qui sont :

- HMM : le modèle de Markov cache.
- DTW : Alignement temporelle dynamique.
- SVM : Machine à vecteur de support.

- VQ : quantification vectorielle.
- GMM : le modèle mélange de Gaussien.

II.11.1 Le Modèle de Markov Cache : ou

HMM Hidden Markov Models : sont des approches stochastique qui utilise la probabilité a la place de la distance ou le signal de la parole est représenter par une séquence d'états d'observation .le principe de reconnaissance d'un mot avec HMM consiste à trouver un modèle qui reconstitue le mot avec une grande probabilité.

Un modèle de markovʎ (A, B, **π**) est un automate probabiliste d'états finis constitue de N états.

Un processus aléatoire se déplacé d'état en état à chaque instant, et on note q_t les numéros de l'état atteint par le processus l'instant t .L'état réelq_t du processus n'est pas directement observable (cache), mais le processus émit âpres chaque changement d'état un symbole discret O_t qui appartient a un alphabet fini de $n_v$symbole V=$\{V_M\}, 1 \leq M \leq n_v$. Dans le cas d'un processus markovien de premier ordre, la probabilité de passe de l'état i a l'état j a l'instant t et d'emmètre le symbole v_k ne dépend ni de la tempe s, ni des états au instants précédentes .un modèle de Markov Cache [49] ou HMM est alors défini par

- Un ensemble S=$\{S_1, S_2, \ldots . S_N\}$ de N états ou un état es défini a listant t par :$q_t \in S$
- Un ensemble S=$\{V_1, V_2, \ldots . V_M\}$ qui contient M symboles d'observations observation à l' instant t note par :$O_t \in V$
- La matrice de probabilité A=$\{a_{ij}\}$ou a_{ij} est la probabilité de passage de l'état i vers l'état on a :

$$\boldsymbol{a_{ij}}\textbf{= p (}\boldsymbol{q_{t+1}}\textbf{=}\boldsymbol{S_j}|\boldsymbol{q_t = S_j} \qquad \textbf{avec } \boldsymbol{1 \leq i \; et \; j \leq N}\textbf{1} \qquad \text{II.17}$$

- la matrice de probabilité B=$\{b_j\}$ou b_j est la probabilité d'observation d'un symbole $v_k$sachant qu'on est q l'état j on a :

$$\boldsymbol{b_j}\textbf{(k)= p (}\boldsymbol{O_t}\textbf{=}\boldsymbol{V_k}|\boldsymbol{q_t = S_j}\textbf{)} \qquad \textbf{avec } \textbf{1}\leq \boldsymbol{j \leq N}\textbf{et } \textbf{1}\leq \boldsymbol{j \leq M} \qquad \text{II.18}$$

- la probabilité initiales π=$\{\pi_1, \pi_{2,\ldots}\pi_N\}$ ou π_i est la probabilité que le modèle commence par l'état i on a

$$\boldsymbol{\pi_i}\textbf{= p (}\boldsymbol{q_i}\textbf{=}\boldsymbol{S_i}\textbf{),} \qquad \textbf{avec } \boldsymbol{1 \leq i \leq N}\textbf{1} \qquad \text{II.19}$$

Un exemple de Markov cache a 4 états est représenter sur la Figure ci-dessus [49] :

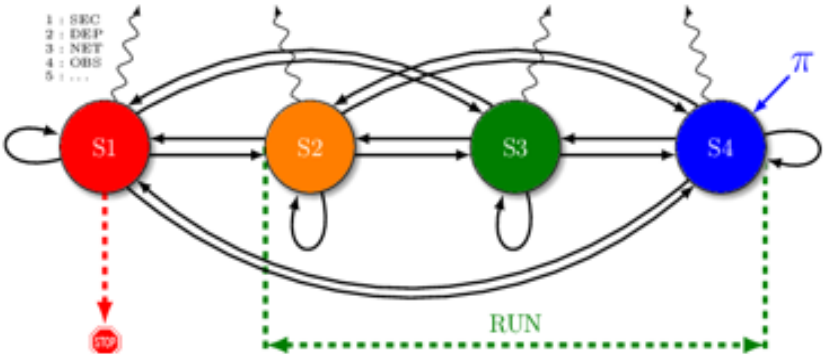

Fig. II.14 : Modèle HMM a 4 états

Pour utiliser la méthode des HMMs, il faut résoudre les trois problèmes fondamentaux suivants :

II.11.1.1 Evaluation de la vraisemblance :

L'estimation de la probabilité que la séquence d'observations ait été générée par un modèle est essentielle. Lorsqu'il y a plusieurs modèles disponibles, cette évaluation permet de sélectionner le modèle le plus probable [49].

II.11.1.2 Le Décodage :

La recherche de la séquence d'états d'un modèle ayant produit les observations. La séquence cachée de plus forte probabilité est déterminée par l'algorithme de Viterbe [49].

II.11.1.3 L'apprentissage :

L'apprentissage des paramètres d'un modèle. À partir d'un modèle donne a priori et d'observations supposées émises par ce modèle, on cherche les probabilités de transition et d'émission maximisant la vraisemblance des observations on utilise l'algorithme Baum Welche.

La résolution du problème d'évaluation de la vraisemblance consiste à calculer la mesure moyenne d'adéquation entre une séquence d'observations et un modèle. Cela permet ensuite de déterminer le meilleur modèle en utilisant la règle de Bayes. En ce qui concerne le décodage, résoudre ce problème permet de segmenter les séquences en recherchant la séquence d'états présentant la plus grande vraisemblance. Enfin, l'apprentissage doit permettre d'adapter automatiquement un modèle de Markov caché (HMM) à un ensemble spécifique de données [49].

II.11.2 Alignement Temporelle Dynamique ou DTW :

DTW Dynamics Time Warping est un moyen de comparer deux séquences, généralement temporelles, qui ne sont pas parfaitement synchronisées [46]. Elle repose sur le principe que chaque mot est représenté par une prononciation de référence [16].Il s'agit de synchroniser temporellement une séquence de vecteurs de paramètres de test avec une séquence de vecteurs d'apprentissage. Dans ce contexte, le modèle de locuteur est simplement constitué de l'ensemble des vecteurs de paramètres obtenus après avoir paramétrées les signaux d'apprentissage. Une mesure de distance est

calculée entre les vecteurs d'apprentissage et de test, puis elle est moyennée sur l'ensemble de la séquence [48].

La mesure de distance entre deux vecteurs caractéristiques x et y ce calcule avec une simple distance euclidienne donnée par :

$$\mathbf{D(x,y)}=\sqrt{\sum_j^p (X_j Y)^2{}_j} \qquad \text{II.20}$$

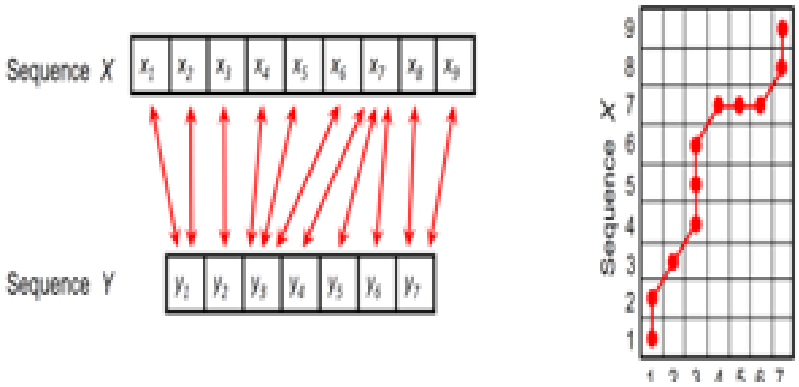

Fig. II.15 : Exemple d'un model DTW

La programmation dynamique est une méthode exclusivement utilisée en mode dépendant du texte. C'est une approche rapide qui fournit des résultats relativement bons. Cependant, elle est très sensible à la qualité de l'alignement, en particulier au choix du point de départ [48].

II.11.3 Machine à vecteur de support ou SVM

Les SVM (Support Vecteur Machines), ou machines à vecteurs de support, sont basées sur le concept de plans de décision qui définissent les limites de décision. Un plan de décision est un plan qui sépare un ensemble d'objets appartenant à différentes classes. Une illustration schématique de cet exemple est présentée ci-dessous. Dans cet exemple, les objets appartiennent à la classe verte ou rouge. La ligne de séparation établit une limite où tous les objets à droite sont rouges et tous les objets à gauche sont verts. Tout nouvel objet (représenté par un cercle blanc) qui tombe à droite est étiqueté comme appartenant à la classe rouge (ou à la classe verte s'il tombe à gauche de la ligne de séparation) [47].

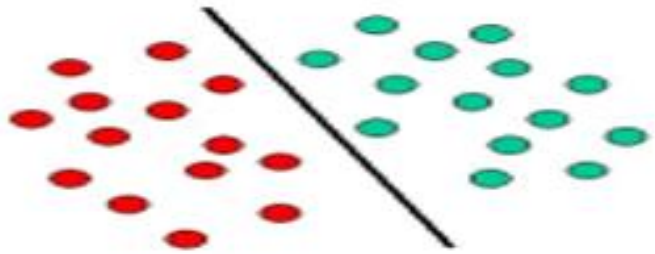

Fig. II.16 : Classification SVM linéaire

Cet exemple représente un cas classique de classification linéaire, où un classificateur sépare un ensemble d'objets en leurs groupes respectifs (rouge et vert dans ce cas) à l'aide d'une ligne. Cependant, la plupart des tâches de classification ne sont pas aussi simples, et des structures plus complexes sont souvent nécessaires pour effectuer une séparation optimale. Cela signifie classifier correctement de nouveaux objets (cas de test) en se basant sur les exemples disponibles (cas d'entraînement). Cette situation est illustrée ci-dessous. Comparé au schéma précédent, il est clair qu'une séparation complète des objets rouges et verts nécessiterait une courbe plus complexe qu'une simple ligne. Les tâches de classification basées sur la recherche d'une ligne de séparation pour distinguer les objets appartenant à différentes classes sont connues sous le nom de classificateurs hyperplans. Les SVL's sont particulièrement adaptées à ce type de tâches. [47].

Figure II.17 : Classification SVM hyperplan

L'illustration ci-dessous met en évidence l'idée fondamentale derrière les SVM (Support Vector Machines). Ici, nous observons les objets d'origine (à gauche du schéma) qui sont ensuite mappés, c'est-à-dire réarrangés, à l'aide d'un ensemble de fonctions mathématiques appelées noyaux. Ce processus de réarrangement des objets est appelé cartographie (transformation). Il est important de noter que dans ce nouvel espace de paramètres, les objets mappés (à droite du schéma) sont linéairement séparables. Par conséquent, au lieu de construire une courbe complexe (comme dans le schéma de gauche), tout ce que nous avons à faire est de trouver une ligne optimale qui peut séparer les objets verts et rouges [47].

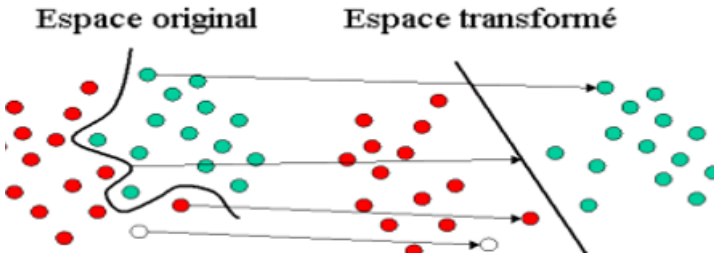

Fig. II.18: Processus de rangement des objets

II.11.4 Quantification Vectorielle :

VQ vecteur quantization est une technique de compression de donnes, qui consiste à coder de manière efficace des échantillons représenter par plusieurs valeurs (vecteurs).Ce codage se fait de la manière suivante :

On divise l'espace en classe adaptées al ensemble des échantillons et on calcule un représentant pour chaque classe. Ce représente appelé centriole ou noyau, représente la distance minimale intra-classes. L'ensemble des noyaux est appelé dictionnaire ou cood-book. Imaginons du 'on a un ensemble d'échantillon, chacun représente par un couple de valeur (x_1, x_2) pour quantifier un échantillon X, on lui attribue les valeurs du représentant le plus proche [49]

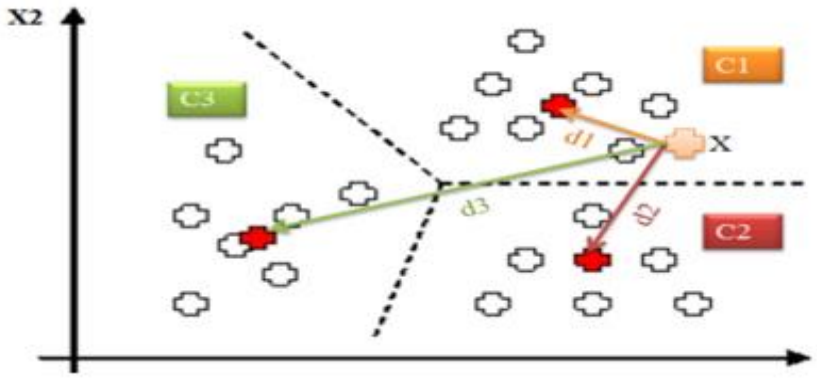

Fig. II.19 : Quantification vectorielle d'un échantillon de dimension 2

Cette figure montre que le vecteur X appartient à la classe C1, car il est plus proche du noyau de cette classe ($d_1 < d_2 < d_3$) distance euclidienne.

Malgré la simplicité du codage par quantification vectorielle, la conception d'un dictionnaire est plus compliquée et a donné lieu à de nombreux algorithmes, l'un des plus utilises est l'algorithme K-means [49].

II.11.4.1 L'algorithme K –means

L'algorithme K –means consiste à définir d'une manière itérative L classes à partir de L-vecteurs de paramètre qui constitué l'ensemble d'apprentissage. Chaque classe est concertée autour d'un noyau, le dictionnaire des références est constitué de l'ensemble des noyaux des diverse classe, l'algorithme est décrit comme suite [49] :

II.11.4.1.1 Initialisation : le nombre de classe M est choisi a priori, alors on procède à leur initialisation d'une manière aléatoire avec n_i noyauc (mots),$1 \leq i \leq M$.

II.11.4.1.2 Affectation : affecter chaque élément $x_k(1 \leq K \leq L)$, de l'ensemble d'apprentissage, a chacune des classe en utilisant la loi du K plus proche voisin (avec k=1), qui consiste a choisir le noyau le plus proche pour chaque élément x_k :

$$\mathbf{x_k} \in \mathbf{ss_i d(x_k, n_i)} \leq \mathbf{d(x_k, n_j), avec j \neq i et 1 \leq j \leq M} \qquad \text{II.21}$$

d : la mesure de distorsion (il s'agit de la distance euclidienne dans la plupart des cas) [49].

II.11.4.1.3 Mise à jour : calcule des nouveaux noyaux des classes, afin de minimiser la distorsion au sein de chaque classe :n_i $est\ defini\ par$[49]:

$$\mathbf{n_i}\ \frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{n=1}}^{\mathbf{N}}\mathbf{X_n} \qquad \text{II.22}$$

Ou N est le nombre d'élément de la classeC_i.

II.11.4.1.4 Tests d'arrêt : si les contenus des classes restent stables et inchangés entre deux itération consécutives alors fin de l'algorithme, sinon retour a l'étape 2.

Il est noté que cet algorithme converge vers un optimum local, qui dépend des valeurs initiales des noyaux des classes [49].

II.11.5. Le Modèle Mélange de Gaussien ou GMM

Le modèle de mélange de gaussiennes (GMM - Gaussian Mixture Model) est une approche utilisée pour représenter un téléphone portable. Dans ce modèle, le téléphone est modélisé comme une combinaison pondérée de M gaussiennes multidimensionnelles. Chaque gaussienne, notée gi, est utilisée pour représenter un ensemble de classes acoustiques. Chaque gaussienne gi est caractérisée par son poids w_m, un vecteur moyen m de dimension d, et une matrice de covariance Σm de dimension D x D. La fonction de densité de probabilité correspondante peut être exprimée par :

$$\mathbf{P(x|\gamma)}=\sum_{m=1}^{M} w_m N(x|\mu_m,\Sigma_m) \qquad \text{II.23}$$

$$N(x|\mu_m,\Sigma_m)=\frac{1}{(2\pi)^{\frac{D}{2}}|\sum_m^{\frac{1}{2}}|}e^{(-\frac{1}{2}(x-\mu_m)T^t\sum_m^{-1}(x-\mu_m)} \qquad \text{II.24}$$

$$\sum_{m=1}^{M} w_m=1 \qquad \text{II.25}$$

L'apprentissage du modèle GMM comprend l'utilisation de l'ensemble de données d'apprentissage X = {X1, X2, ..., XT} pour estimer tous les paramètres. Ce type d'apprentissage nécessite généralement la technique d'estimation du maximum de vraisemblance MLE (Maximum Likelihood Estimation).

Le principal inconvénient de cette technique est le nombre de signaux d'apprentissage requis pour une bonne estimation des paramètres du modèle [50].

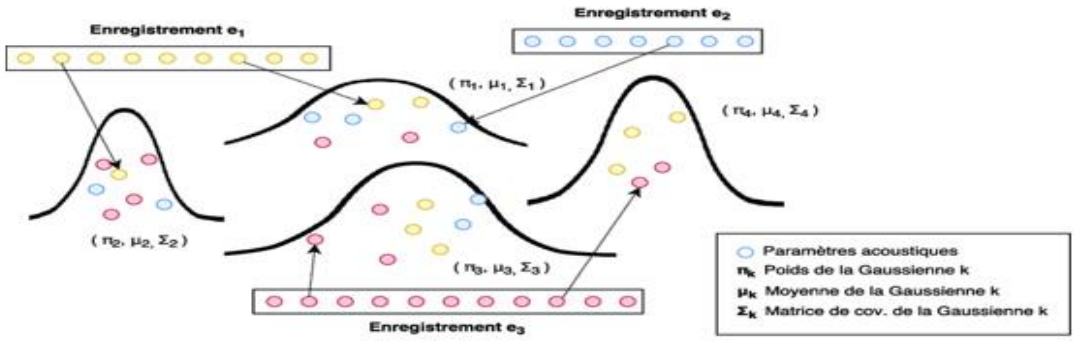

Fig. II.20 : Mélange de Gaussiennes (GMM) construit en utilisant des paramètres acoustiques issus de plusieurs enregistrements

II.11.5.1Approche GMM-UBM (Gaussien Mixture Model-Universel Background Model) :

Pour résoudre le problème de données insuffisantes, une version améliorée du modèle GMM a été développée. Cette approche, connue sous le nom de GMM-UBM (Universal Background Model), est utilisée pour modéliser les téléphones mobiles. Voici la description et le schéma de principe de la méthode GMM-UBM [50] :

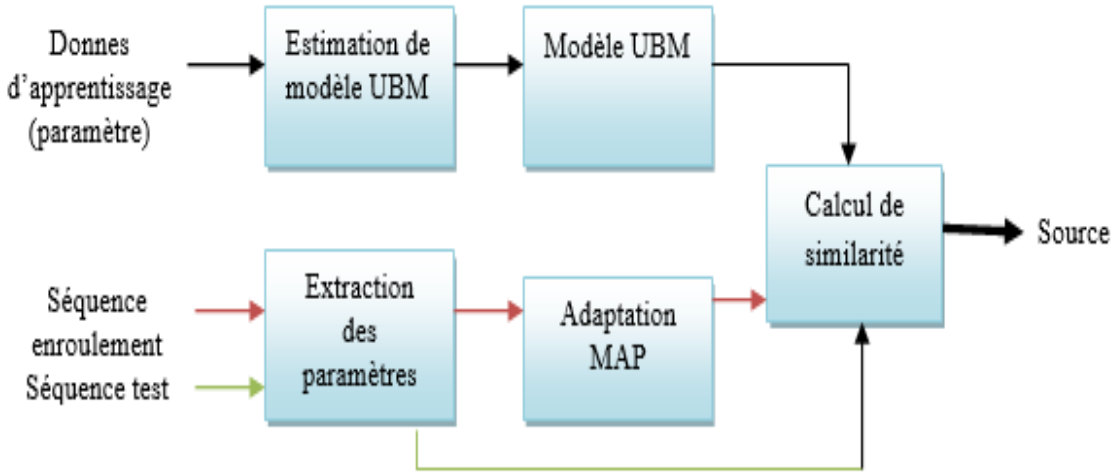

Fig. II.21: Architecture du système RA à base de GMM-UBM. [50]

II.11.5.2 Estimation du modèle du monde UBM

1. Un seul modèle indépendant des téléphoneλ_{UBM}, appelé modèle universel du monde (UBM - Universel Background Model) est employé. L'apprentissage du modèle UBM se réalise en utilisant un vaste ensemble de données vocales obtenues en concaténant une large gamme de téléphones différents.

2. Un algorithme itératif appelé Expectation Maximisation (EM) est utilisé pour estimer la vraisemblance maximale du modèle par rapport au vecteur de paramètres d'apprentissage. La

moyenne et la matrice de covariance de l'ensemble des V vecteurs sont calculées et une valeur de 1 est affectée au poids. . Pour chaque itération j allant de 1à
log_2 (M) (où : M est le nombre de composantes gaussiennes), les étapes de l'algorithme EM sont détaillées comme suite [50] :

3. Etape de l'algorithme EM

- **Estimation :** L'estimation consiste à calculer l'appartenance de chaque vecteur acoustique$x_t$de la matrice X_{UBM}à chacune des gaussiennes i avec (1 ≤ i ≤ M) du modèleΛ_{UBM}.
- **Maximisation** : La maximisation consiste à mettre à jour les poids, les moyennes et les matrices de covariance obtenus lors de l'estimation.

Poids du mélange [50] :

$$\overleftarrow{w_i} = \frac{1}{V}\sum_{t=1}^{V} P_r\,(i|x_t, \Lambda_{UBM}) \qquad \text{II.26}$$

Moyenne

$$\overline{\boldsymbol{\mu}}_{\mathbf{i}} = \frac{\sum_{\mathbf{t=1}}^{\mathbf{V}} \mathbf{P_r(i|x_t,\Lambda_{UBM})x_t}}{\sum_{\mathbf{t=1}}^{\mathbf{V}} \mathbf{P_r(i|x_t,\Lambda_{UBM})}} \qquad \text{II.27}$$

Variance

$$\overline{\boldsymbol{\delta}}_{\mathbf{i}}^{\mathbf{2}} = \frac{\sum_{\mathbf{t=1}}^{\mathbf{V}} \mathbf{P_r(i|x_t,\Lambda_{UBM})x_t^2}}{\sum_{\mathbf{t=1}}^{\mathbf{V}} \mathbf{P_r(i|x_t,\Lambda_{UBM})}} - \overline{\boldsymbol{\mu}}_{\mathbf{i}}^{\mathbf{2}} \qquad \text{II.28}$$

II.11.5.3 Estimation des modèles des locuteurs par l'adaptation MAP

La dérivation des caractéristiques du téléphone est réalisée de manière adaptative dans le système GMM-UBM. Les enregistrements d'apprentissage spécifiques à chaque téléphone sont utilisés pour ajuster les paramètres du modèle universel du monde (UBM) en utilisant l'algorithme d'estimation du maximum a posteriori (MAP), qui se résume à :

$$\boldsymbol{n_i} = \sum_{\boldsymbol{t=1}}^{\boldsymbol{V}} \boldsymbol{P_r(i|x_t, \Lambda_{UBM})} \qquad \text{II.29}$$

$$\boldsymbol{E_i(x)} = \frac{\boldsymbol{1}}{\boldsymbol{n_i}}\sum_{\boldsymbol{t=1}}^{\boldsymbol{V}} \boldsymbol{P_r(i|x_t, \Lambda_{UBM})x_t} \qquad \text{II.30}$$

$$\boldsymbol{E_i(x^2)} = \frac{\boldsymbol{1}}{\boldsymbol{n_i}}\sum_{\boldsymbol{t=1}}^{\boldsymbol{V}} \boldsymbol{P_r(i|x_t, \Lambda_{UBM})x_t^2} \qquad \text{II.31}$$

x 2 est la notation signifiant diag (xx').

Ces nouvelles statistiques suffisantes sont utilisées pour mettre à jour les paramètres de la gaussienne i :

$$\widehat{\mathbf{w}}_{\mathbf{i}} = \left[\frac{\mathbf{a}_{\mathbf{i}}^{\mathbf{w}}\mathbf{n}_{\mathbf{i}}}{\mathbf{V}+(\mathbf{1}-\mathbf{a}_{\mathbf{i}}^{\mathbf{w}})\mathbf{w}_{\mathbf{i}}}\right] \quad \text{II.32}$$

$$\widehat{\boldsymbol{\mu}}_{\mathbf{i}} = \mathbf{a}_{\mathbf{i}}^{\mathbf{m}}\mathbf{E}_{\mathbf{i}}(\mathbf{x}) + (\mathbf{1} - \mathbf{a}_{\mathbf{i}}^{\mathbf{m}})\boldsymbol{\mu}_{\mathbf{i}} \quad \text{II.33}$$

$$\widehat{\boldsymbol{\delta}}_{\mathbf{i}}^{\mathbf{2}} = \mathbf{a}_{\mathbf{i}}^{\mathbf{v}}\mathbf{E}_{\mathbf{i}}(\mathbf{x}^{\mathbf{2}}) + \left(\mathbf{1} - \mathbf{a}_{\mathbf{i}}^{\mathbf{v}}\right)\left(\boldsymbol{\delta}_{\mathbf{i}}^{\mathbf{2}} + \boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{2}}\right) - \widehat{\boldsymbol{\mu}}_{\mathbf{i}}^{\mathbf{2}} \quad \text{II.34}$$

Les coefficients d'adaptation contrôlant l'équilibre entre les anciennes et nouvelles estimations sont $a_{i,}^{p}\{w, m, v\}$ pour le poids, la moyenne et la variance respectivement avec$a_i^p = \frac{n_i}{n_i + r_p}$. Le facteur d'échelle γ.est calculé sur tous les poids du mélange adapté assurer leur somme a l'unité [50]

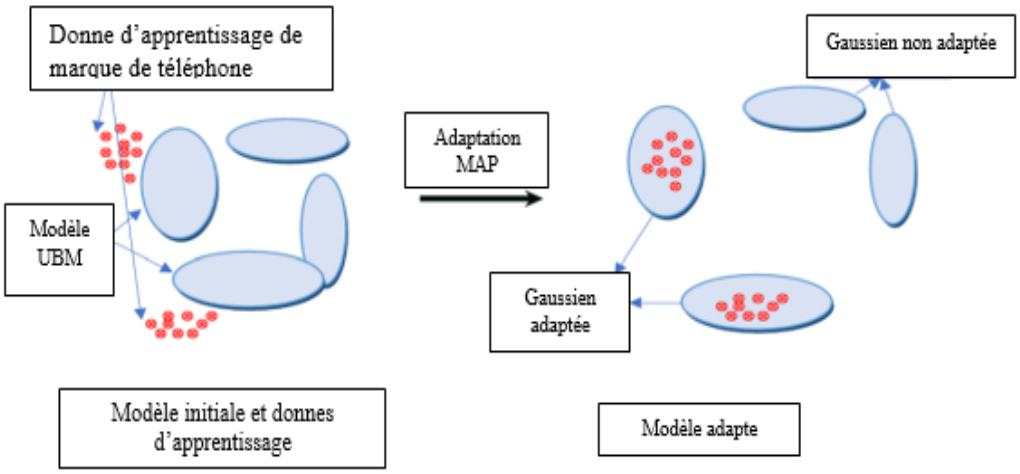

Fig. II.22 : Adaptation MAP d'un modèle GMM-UBM.

II.11.6. Compensation de l'effet de la variabilité

II.11.6.1. Principal Component Analysis OU PCA

L'analyse en composantes principales (ACP) permet de définir un sous-espace à partir d'un ensemble de données d'apprentissage, ce qui permet de préserver les informations distinctives tout en éliminant les informations secondaires (non informatives). Cette méthode consiste à trouver une nouvelle base dans l'espace de données où tous les vecteurs sont orthogonaux entre eux. Le premier de ces vecteurs correspond à la direction de variance maximale des données d'apprentissage, tandis que les autres composantes sont déterminées en respectant les contraintes orthogonales par rapport à la direction de variance maximale. Dans la méthode de l'ACP, la normalisation de l'éclairage reste essentielle.

La procédure de l'ACP est largement employée en reconnaissance de formes en raison de sa rapidité et de sa simplicité. Elle offre la meilleure approche pour reconstruire une base de dimension réduite, grâce à des projections optimales des vecteurs propres de la matrice de covariance composée des diverses images de notre ensemble d'apprentissage, la procédure est comme suit [50] :

Etape1 : Sélectionnez data matrice, X^T moyenne nulle.

Etape2 : Calculer la moyenne.

$$\boldsymbol{\psi} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{X_i} \qquad \text{II.35}$$

Etape3 : Soustraire la moyenne de la distribution à partir de l'ensemble de données.

$$\boldsymbol{X_i} = \boldsymbol{X^T} - \boldsymbol{\psi} \qquad \text{II.36}$$

Etape4 : Calculer la matrice de covariance XX^T

$$\mathbf{C}=\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{X_i X_i^T} \qquad \text{II.37}$$

Etape5 : Calculer les valeurs propres et les vecteurs propres V de la matrice de covariance.

Où = 1 . . . N.

Etape6 : Ordonner les vecteurs propres V_i (i = 1 . . . N) par leurs valeurs propres correspondantesλ_i, par ordre décroissant.

Etape7 : Ne conserver que les vecteurs propres avec les valeurs propres les plus importantes (les composants principaux), k (k ≪ N) $X^k = V^k.X$

Etape8 : Résoudre pour PCA.

$$\boldsymbol{\lambda V_{x^T}} \mathbf{T}=\boldsymbol{C_x V_{x^T}} \qquad \text{II.38}$$

Le fonctionnement de l'ACP peut être considéré comme révélateur de la structure interne des données de manière à mieux expliquer la variance dans les données.

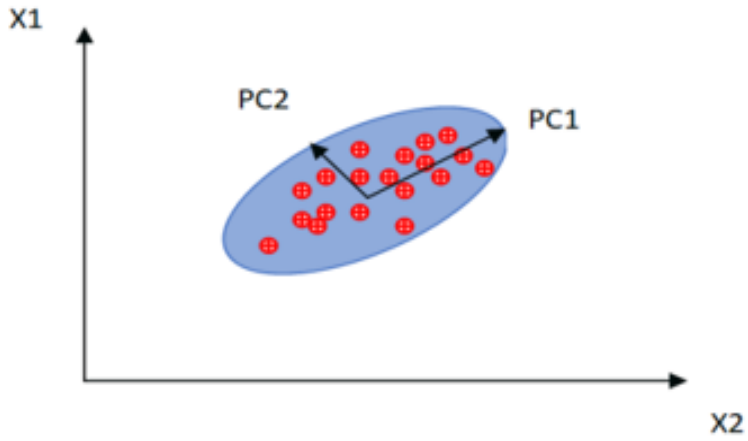

Fig. II.23 : Analyse en Composantes Principales.

II.12. Calculé le scores avec la Méthode CSS (Cosine Similarité Scoring)

À ses débuts, la mesure de similarité en cosinus (Cosine Distance) fut employée dans les articles fondateurs du paradigme de la variabilité totale. Par la suite, elle fut également adoptée dans le domaine de la reconnaissance faciale. Dans cette approche, le score entre les paramètres du client (wclien) et ceux du test (wtest) est évalué par le biais d'un produit scalaire normalisé. :

$$\boldsymbol{Score_{CD}} = (\boldsymbol{w_{client}}, \boldsymbol{w_{test}}) = \frac{\boldsymbol{w_{client}} \cdot \boldsymbol{w_{test}}}{||\boldsymbol{w_{cleint}}|| \, ||\boldsymbol{w_{test}}||} \qquad \text{II.39}$$

Le ”.” Fait référence au produit scalaire entre deux vecteurs.

Malgré sa simplicité structurelle, cette approche a démontré une grande efficacité dans le domaine de la réalité augmentée (RA), grâce à son absence d'informations préalables sur l'apprentissage des classes (en l'occurrence, les téléphones mobiles). Dans la littérature, cette distance est couramment associée à des algorithmes de compensation de variabilité canal/session, tels que la normalisation WCCN et/ou l'analyse discriminante linéaire) [50].

II.13 Conclusion

Dans le présent chapitre, nous avons succinctement examiné les méthodes employées pour l'extraction des paramètres acoustiques qui alimentent un système de reconnaissance. Nous avons survolé ces techniques de calcul des coefficients représentant le signal sonore. Le chapitre suivant abordera de manière plus détaillée leur utilisation dans le cadre expérimental du système de reconnaissance du locuteur.

Chapitre III :
Résultats et discussions

III.1Introduction

A partir de ce qui a été évoqué dans les deux chapitres précédents, nous avons présenté les aspects généraux de la parole (production et audition), les outils nécessaires pour son traitement et sa paramétrisation, ainsi les deux déférentes méthodes existantes des deux blocs : extraction de paramètre et modélisation SRAP. Après avoir expliqué toutes les notions théoriques nécessaires à notre travail, Passons maintenant au côté pratique.

Nous parlerons d'abord du contexte expérimental dans lequel nous présenterons le système sur lequel nous travaillons, la base de données d'enregistrements utilisée dans nos expériences, le logiciel qui sera mis en œuvre dans ce travaille.
Ensuite, nous effectuions un certain nombre d'expériences en soumettant leurs résultats à Choisissez les paramètres optimaux avec lesquels notre système donne les meilleurs résultats. Tout en Mise en évidence de l'effet du changement de canal sur les résultats du système.
Enfin, sur la base de ces paramètres, nous testerons des techniques pour tenter de compenser ce Contraste en affichant les résultats des expériences tout en les analysant.

III.2Description de base des données

Nous avons choisi d'utiliser la base de données acoustique américaine TIMIT dans le développement de notre système de reconnaissance. Cette base de données tirée est composée de 60 locuteurs, dont trente femmes et trente hommes, dans le but de répondre à plusieurs objectifs :

- ✓ Elle permet d'illustrer au mieux la variabilité acoustique de l'anglais américain. De plus, elle est fournie avec une segmentation phonétique de référence, ce qui facilite l'apprentissage initial des modèles phonétiques.
- ✓ TIMIT est largement reconnue comme une base de données de référence. Sa diffusion étendue au sein de la communauté internationale offre la possibilité d'évaluer objectivement les performances des systèmes développés.

Dans la base de données internationale de parole TIMIT, les parties réservés à l'apprentissages et tests ont été effectués à partir de phonèmes parlés extraits manuellement des phrases complètes multi-locuteurs, de la base de données TIMIT qui

contient 61 phonèmes constituant la phonétique de la langue anglaise, tirés d'un total de 6300 phrases, 10 phrases parlées par chacun des 630 orateurs de 8 dialectes principaux de l'anglais américain.

III.3Environnement de travail

Dans cette étude, nous avons employé MATLAB (Matrix Laboratory) comme langage de programmation, en utilisant spécifiquement la version MATLAB 2021a, qui est une interface interactive développée par Math Works. MATLAB nous a offert une approche aisée et efficace pour implémenter les algorithmes nécessaires, ainsi que pour réaliser des tâches exigeant un calcul intensif. Grâce à cette plateforme logicielle, nous avons pu rapidement et simplement mettre en œuvre nos méthodes et profiter de la puissance de calcul requise.

III.4Protocole de travail

Le fonctionnement dans notre système de reconnaissance du locuteur se déroule en deux phases :

a) La phase d'apprentissage.

b) La phase de reconnaissance.

Notre base de données est composée de (60) fichiers audio dont 30 femmes et 30 hommes.

- Les fichiers audio femmes sont numérotés de 1 à 30.
- Les fichiers audio hommes sont numérotés de 31 à 60.

Le tableau ci-dessus représente les 60 fichiers audois, L'apprentissage comte les 40premier fichiers audio et la partie teste compte les 20 derniers fichiers audio.

Fichiers audio	Femmes	Hommes
Apprentissage	1 à 20	31 à 50
Test	21 à 30	51 à 60

Tab. III.1 : représentation des fichiers audio.

III.5 Répartition naïve

On considère une base de données qu'on divise en 2/3 apprentissage et 1/3 test, on aura donc [51] Apprentissage= $100*\frac{2}{3} \approx 60\ Fechieraudio.$

Teste $=100*\frac{1}{3} \approx 40$ Fichier audio.

III.6Méthodologie proposée pour l'extraction des paramètres désirés

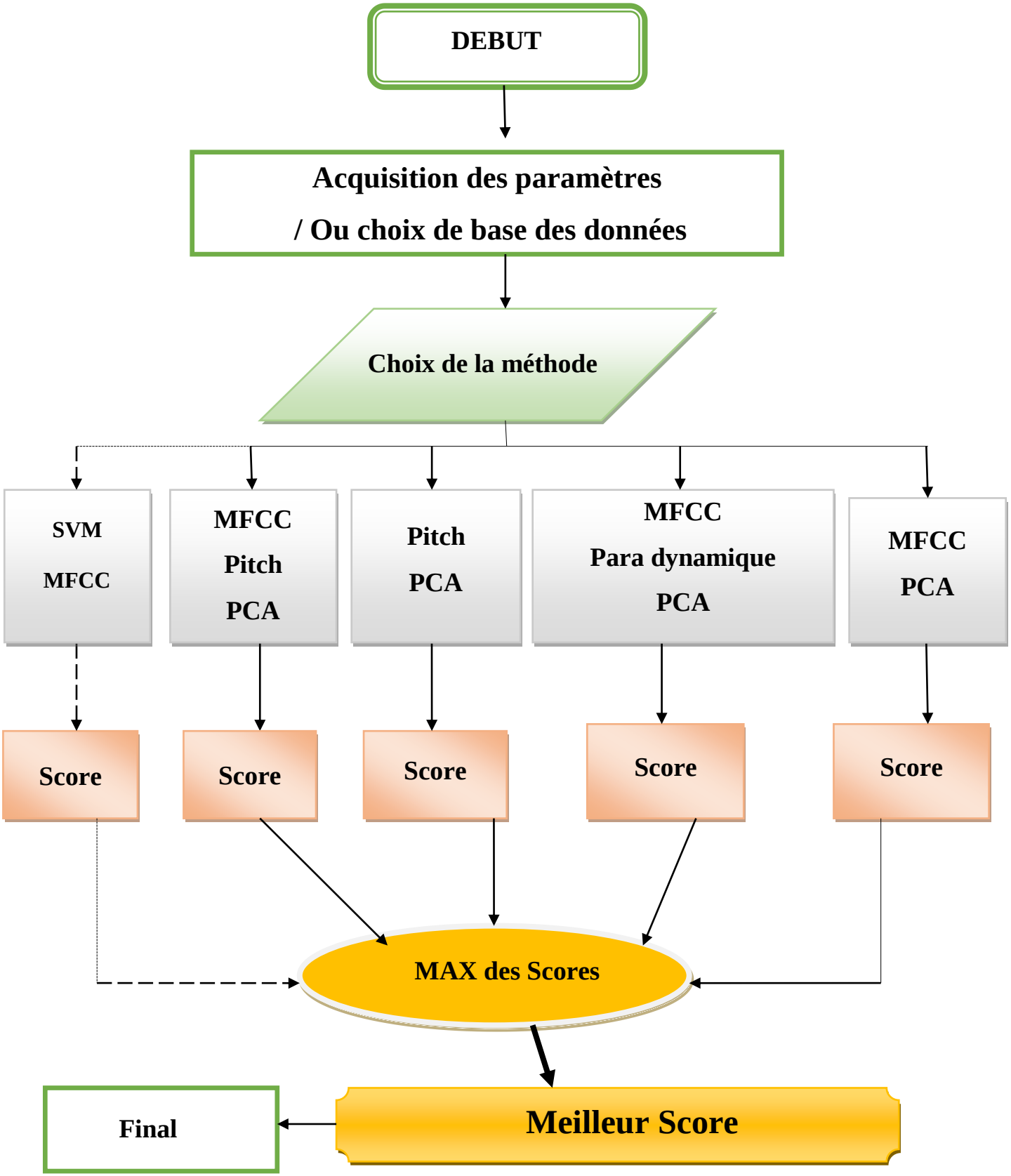

Fig.III.1 : Procédure étape par étape de la méthodologie d'extraction des paramètres.

III.7Expérimentations et résultats

III.7.1Extraction des paramètres acoustiques

En utilisant le logiciel de programmation MATLAB, nous avons commencé à construire notre système de reconnaissance comme expliqué au chapitre 2. Comme tous les systèmes de reconnaissance notre système est divisé sur deux phases

La première étape consiste à extraire les paramètres acoustiques à l'aide des coefficients MFCC. Ces paramètres sont ensuite utilisés pour créer chaque modèle à l'aide de transférer le paramètre pour un vecteur comme l'histogramme de l'image, chaque vecteur prestone l'audio de chaque personne. Puis nous réduire et projeter les vecteurs avec la méthode de PCA pour trouver les principaux composants. Les modèles résultants seront stockés dans la base de données. Ces étapes forment la phase d'apprentissage.

La phase de test implique la récupération des paramètres et leur comparaison. La récupération des paramètres se fait de manière similaire à la phase d'apprentissage. La comparaison est réalisée en calculant une mesure de similarité, utilisant la similarité cosinus, avec les modèles créés. Les résultats obtenus sont évalués en fonction du taux de reconnaissance correct, qui représente le nombre de tests corrects effectués par le système sur le nombre total de tests réalisés.

III.7.2 Modélisation

Dans notre Project nous avons essayé deux différentes méthodes de modélisation pour la reconnaissance automatique de locuteur la premier c'est la modélisation par vecteur histogramme et la deuxième méthode est la SVM.

III.7.2.1Modélisation avec la technique basée sur les vecteurs histogrammes

L'histogramme d'une image est un outil important en traitement d'images. Il représente la répartition de la luminosité des pixels dans une image en utilisant un graphique. L'axe horizontal représente les niveaux de luminosité, allant du noir (0) au blanc (255), alors que l'axe vertical représente le nombre de pixels correspondant à chaque niveau de luminosité. Cette information est très utile pour comprendre la dynamique de l'image et peut être utilisée pour réaliser des modifications efficaces sur l'image. Dans notre travail, nous utilisons cette méthode directement sur les propriétés MFCC, qui sont des matrices à deux dimensions comme l'image. Avant d'utiliser cette méthode nous avons normalisé ce paramètre 'MFCC' entre 0 et 255.

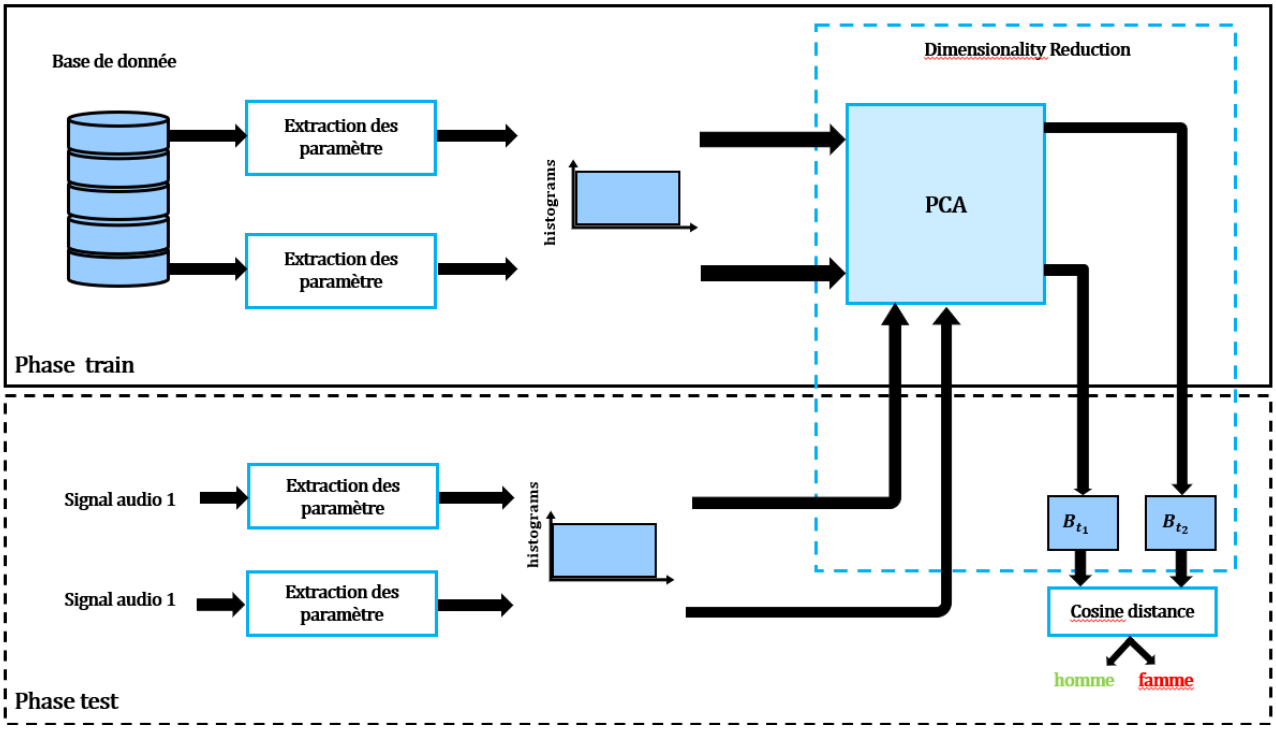

Fig.III.2 : schéma représente la Structure du système de base.

III.7.2.1.1 Influence du nombre de coefficients MFCC

Dans cette partie nous avons teste des déférente nombre de MFCC (de 12 a 22) et on ana obtenue un taux pour chaque nombre de MFCC. Les résultats sont représentés dans le tableau ci-dessus :

Nbr_MFCC	12	14	16	18	20	22
Taux	72.22	77.78	83.33	83.33	83.33	83.33

Tab. III.2 : Représente le taux d'influence de nombre

D'âpres les résultats obtenus dans le tableau III.2 nous observant que le millier score est 83.33 pour un nombre de MFCC supérieure ou égale à 16(Nbre MFCC ≥16)

III.7.2.1.2 Influence des paramètres dynamiques

Dans cette partie de simulation, Nous allons fixerons le nombre de coefficient de MFCC a une valeur fixe de 18. Nous jouerons sur les paramètres dynamiques : énergie 'e', vitesse 'd', accélération 'D' et le pitch. On a obtenu des taux représentés dans le tableau ci-dessous :

Paramètres dynamiques	Taux
E	55.56
d	50.00
D	50,00
Pitch	72.22

Tab. III.3 : Représentation de l'influence des paramètres

Dans ce cas, nous avons étudié différentes situations pour montrer l'efficacité de notre système. D'après les résultats obtenus dans le tableau III.3 5 nous constatons que le millier score est

- 55.56 pour les paramètres dynamique « énergie ».
- 72.22 pour le pitch.

III.7.2.1.3 Influence de nombre de projection PCA

Dans cette partie de simulation, Nous jouerons sur le nombre de PCA (de 1 à 40) et on fixera toujours le nombre de coefficient de MFCC à une valeur fixe de 18. Et de cela On obtient destaux représentés dans le tableau ci-dessus :

Nbr_PCA	5	10	20	25	30	35	40
Taux PITCH	83.33	83.33	77.78	72.22	72.22	72.22	77.78
Taux MFCC	66.67	83.33	83.33	83.33	83.33	83..33	83.33

Tab.III.4 : L'influence de PCA

D'âpres les résultats obtenus nous constatant que le meilleur score aux :

- Taux de Pitch égale à 83.33 représente un meilleur score a nombre de PCA ce varie entre 5 et 10.
- Taux MFCC égale 83.33 représente meilleur score à un nombre de PCA supérieure ou égale à 10 ($\geq$10).

III.7.2.1.4 Influence de nombre MFCC, Pitch de projection PCA :

Dans cette partie de simulation, Nous allons étudier l'influence de ces nombres représentés dans le tableau ci-dessus :

Nombre MFCC	Nombre PCA	Influence MFCC, PCA, Pitch
18	10	50.00

Tab.III.5 : Représente l'influence des MFCC, Pitch, PCA

D'âpres les résultats obtenus nous constatant que cette simulation obtient un score de 50 % de reconnaissance et 50 % d'erreur.

III.7.2.2 Modélisation par SVM Support Vector Machine

SVM est un algorithme de classification pour voir si le système fonctionne e, est ce que le taux de reconnaissance est élevé ou faible on va essayer plusieurs combinaisons tel que SVM simple, SVM linière et SVM poly-kernel les résultats seront représenté par le tableau ci-dessus

SVM	Simple	Linéaire	Poly-kernel
Taux de reconnaissance	15.00	35.00	50.00

Tab. III.6 : représente le taux de reconnaissance par déférente combinaison SVM

D'âpres les résultats obtenus nous constatant que :

- Le taux de reconnaissance celui égaleà50 représentant le meilleur score pour la modélisation SVM et y'est de combinaison poly-Kernel
- Taux égale a15 représente un score de reconnaissance très faible pour la combinaison SVM simple
- le diagramme ci-dessus représente la classification SVM poly-kernel

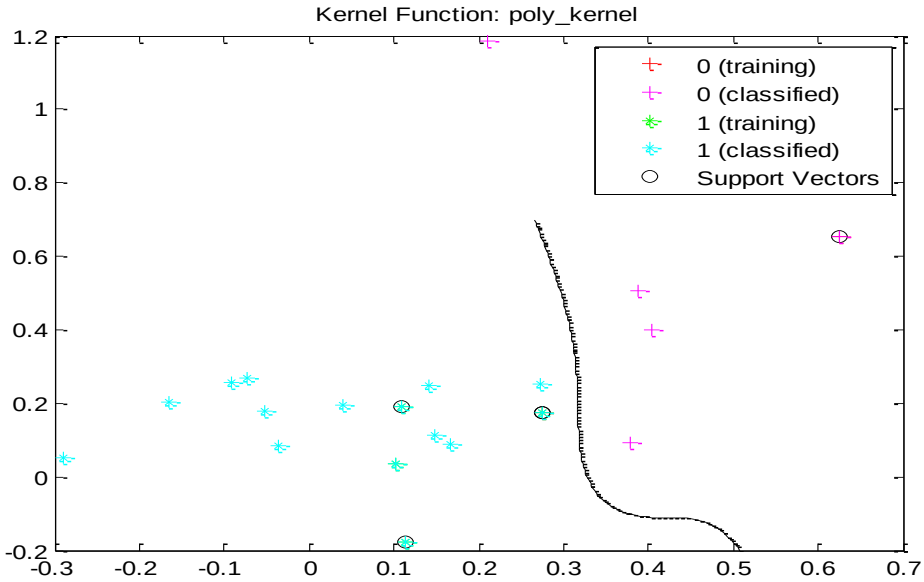

Fig. III.3 : Classification SVM poly-kernel

III.8Conclusion :

Dans ce chapitre, une méthode a été élaborée pour la reconnaissance automatique de locuteur basée sur la technique des vecteurs histogrammes.

Les résultats obtenus montrent l'efficacité de la méthode utilisée pour un meilleur taux de reconnaissance correcte.

La méthode avec le taux de reconnaissance correcte le plus élevé est la simulation de la méthode de MFCC .La méthode avec le taux de reconnaissance correcte le plus diminue est la simulation des paramètres dynamique et l'influence des chaque des pitch, MFCC et Project PCA.et pour bien conclure le millier score obtenue pour un millier taux de reconnaissance correcte est pour la méthode MFCC s'approche a 100%.

Conclusion Générale

L'objectif de notre travail tout au long de ce mémoire était d'aborder le traitement du signal sonore et comment implémenter un système de reconnaissance Automatique de locuteur. Pour cela nous avons commencé par voir de façon globale et générale les différents concepts qui constituent le traitement de la parole. Ensuite Nous avons vu détaillé les différentes étapes qui se trouvent au cœur de ce système. Enfin nous avons essayé pratiquement en utilisant la technique d'extraction de paramètres, de modélisation, de calculer la distance et de calcul de scores.

Les voix utilisent de reconnaissance ont été effectués sur la base de données TIMIT, dans ce genre dans systèmes, l'environnement et les différents types de variabilités influent énormément sur ses performances. Notre system est compose en deux phase (train et teste) :

Dans la Phase Train on a commencé par les coefficients MFCC comme extraction de paramètre puis nous avons créé un vecteur d'histogramme comme un modèle de modélisation et pour finir nous avons utilisé la PCA son rôle est de trouve les informations en vedette (transformée un vecteur d'une valeur infinie en un vecteur de valeur finis).

Dans la Phase teste nous avons fait les mêmes étapes que la phase traine et âpres nous avons calculé la distance (par cosine similarité) entre (femme et femme, femme et homme, homme et homme, homme et femme).Et on a essayé une autre méthode de modélisation est la SVM qui nous a obtenu un taux de reconnaissance très faible.

A partir d'analyse de tous les résultats obtenus par ces expériences, nous avons constaté une meilleure performance pour la reconnaissance est de la méthode MFCC.

Les travaux menés dans le cadre de ce projet nous ouvrent une porte vers de nouvelles technologies et vers de nouvelles méthodes, le monde de la reconnaissance est bien vaste et ne cesse de s'étendre, et quel que soit le niveau d'efficacité atteint par notre application il y'aura toujours un moyen de l'améliorer.

Références bibliographiques

[01] Heidegger, Martin, and Jean Beaufret. "Acheminement vers la parole." (1981).

[02] Ahmed Badis CHENNI, Abderrazak CHARIF."Analyse et synthèse d'un signal de parole par la Matrice de Pencil en vue d'une discrimination de locuteurs. "Master Académique, Université Mohamed Boudiaf MSILA, Septembre2020.

[03] G. Fant, 1960. Acoustic Theory of Speech Production. Hague: Mouton's Co.

[04] Dave, Namrata. "Feature extraction methods LPC, PLP and MFCC in speech recognition." International journal for advance research in engineering and technology 1.6 (2013): 1-4.

[05] Martin, Pierre. "Le système vocalique du français du Québec. De l'acoustique à la phonologie." La linguistique 38.2 (2002) : 71-88.

[06] Hueber, Thomas. Synthese de la parolea partir d'imagerie ultrasonore et optique de l'appareil vocal. Diss. Thèse de doctorat Ecole Supérieure de Chimie Physique Electronique de Lyon.

[07] El Hilah Fatima, Fatiha Ben Akka, et al. "Étude ethnobotanique des plantes médicinales utilisées dans le traitement des infections du système respiratoire dans le plateau central marocain." Journal of Animal &Plant Sciences 25.2 (2015) : 3886-3897.

[08] Granat, Jean, and Evelyne Peyre. "La situation du larynx du genre Homo. Données anatomiques, embryologiques et physiologiques." Biométrie Humaine et Anthropologie-revue de la Société de biométrie humaine 22.3-4 (2004) : 139-161.

[09] Crevier-Buchman, Lise, et al. "Comportements laryngés en voix chuchotée, étude en caméra ultra rapide." *Congres de la Société Française de Phoniatrie et des Pathologies de la Communication*. 2008.

[10] Ghio, Alain, and Bernard Teston. "Caractéristiques de la dynamique d'un Pneumotachographe pour l'étude de la production de la parole : aspects acoustiques et aérodynamique." Journées d'Etude sur la Parole. INRIA et AFCP, 2002.

[11] Chelli, Dalenda, and Badis, Chanoufi. "Audition fœtale. Mythe ou réalité ?" Journal de gynécologie obstétrique et biologie de la reproduction 37.6 (2008) : 554-558.

[12] Schmitt, Arnaud. La région de l'oreille osseuse chez les Proboscidea (Afrotheria, Mammalia) : anatomie, fonction, évolution. Diss. Paris, Muséum national d'histoire naturelle, 2016.

[13] Purves, Dale, et al. Neurosciences. De Boeck Supérieur, 2019.

[14] Jacques, P. "Contribution à la physiologie normale et pathologique de l'oreille moyenne." Acta Oto-Laryngologica 15.2-4 (1931) : 308-311.

[15] Sauvage, Jean-Pierre, et al. "Anatomie de l'oreille interne." EMC-Oto-rhino-laryngologie (1999) : 1-16.

[16] Yasmin AZIZA."Modélisation AR et ARMA de la Parole pour une Vérification Robuste du Locuteur dans un Milieu Bruité en Mode Dépendant du Texte." Magister, Université Ferhat Abbas –Setif1- UFAS (ALGERIE), Octobre 2013.

[17] Amina KABOUT, Yasmine LOUNACI."Comparaison à travers l'analyse temporelle entre deux variantes de signaux sonores – parole et ronflement comme applications."Master02, UniversityAkliMohandOulhadj-Bouira, 2021

[18] Tarik HADJ ALI, Houssam BOUSBAI LAICHE."Speaker recognition." University Mohamed Bougara – Boumerdes- Institut of Electrical and electronic engineering, 2011/2012.

[19] ManelOULMI."Reconnaissance et identification du genre par empreinte acoustique." Maser02, Universite Akli Mohand Oulhadj-Bouira, 2022.

[20] Riad AJGOU."Techniques de détection de la période du pitch par les méthodes temps Fréquence et temps échelle." Magister en électronique, Université de Biskra, Mars 2010.

[21] Boite, René. Traitement de la parole. PPUR presses polytechniques, 2000.

[22] Calliope, La, and G. Fant. La parole et son traitement automatique. Paris : Masson, 1989.

[23] Mariani, Joseph. Reconnaissance de la parole. Hermès science, 2002.

[24] BRAFFORT, ANNE. Reconnaissance et compréhension de gestes, application à la langue des signes. Diss. Paris 11, 1996.

[25] Siwar ZRIBI BOUJELBEN, Dorra BEN AYAD MEZGHANI Et Noureddine ELLOZE."Identification du Locuteur par Système Hybride GMM-SMO. "Département Informatique, FSHST. Département Génie logiciel des systèmes d'information, ISI. Département Génie èlèctrique, ENT. Tunisie, 2009

[26] Maamar Hamadouche. "Technique D'analyse En Vue DE La Reconnaissance Automatique De La Parole." Magister Université Saad Dahlab De Blida, Mai2008.

[27] Riadh ADJOU. "Reconnaissance Automatique du Locuteur à Travers les Canaux Digitaux." Doctorat en Sciences, Université Mohamed Khider – Biskra, Fèvrier2016.

[28] Dessalles, Jean-Louis. "La fonction shannonienne du langage : un indice de son évolution." Langages (2002) : 101-111.

[29] Mounia CHABANE, Kahina BENSAFIA. "Reconnaissance Automatique De Locuteur Par La Méthode Du Taux du Passage Par Zéro." Mémoire De Fin D'étude, Université

Mouloud Mammeri De Tizi-Ouzou Faculté De Génie Electrique Et D'informatique Département Electronique, 2007/2008.

[30] Fields, W. S. "The asymptomatic carotid bruit--operate or not?" Stroke 9.3 (1978) : 269-271.

[31] De Coulon, Frédéric. Théorie et traitement des signaux. Vol. 6. PPUR Presses polytechniques, 1998.

[32] Soliman, Samir S., and S-Z. Hsue. "Signal classification using statistical moments." IEEE Transactions on Communications 40.5 (1992) : 908-916.

[33] Hibatellah DAOUI, Amira SILEM."Reconnaissance Automatique Des émotions Par La Voix." Master02, Université Akli Mohand Oulhadj-Bouira, 2022.

[34] Chaima DEBILOU, Samiha BOUDAOUD."Amélioration Synthétiseur De La Parole Concaténation." Master Académique, Universitè Echahide Hamma Lakhdar El-Oued, Juin2019.

[35] Lyes ALOUCHE, Yanice LOUGGANI." Compensation de la variabilité du canal en Reconnaissance du locuteur."Master02, Université Akli Mohand Oulhadj-Bouira, 2018/2019.

[36] M.A. 1 YUSNITA, M.P2 PAULRAJ, Yaacob 3 SAZALI, Abu Bakar 4 SHAHRIMANE and Saidatul 5 A."Malaysian English Accents Identification using LPC and Formant Analysis." Faculty of Electrical Engineering Universiti Technology MARA Malaysia Shah Alam, Malaysia. School of Mechatronic Engineering University Malaysia Perlis Pauh, Malaysia, November2011.

[37] Farge, Marie, and Gabriel Rabreau. "Transformée en ondelettes pour détecter et analyser les structures cohérentes dans les écoulements turbulents bidimensionnels." CR Acad. Sci. Paris 307 (1988) : 1479-1486.

[38] Farge, Marie, and Gabriel Rabreau. "Transformée en ondelettes pour détecter et analyser les structures cohérentes dans les écoulements turbulents bidimensionnels." CR Acad. Sci. Paris 307 (1988): 1479-1486.

[39] Plessis, R-E., and Q. Cao. "A parametrization study for surface seismic full wave form inversion in an acoustic vertical transversely isotropic medium." Geophysical Journal International 185.1 (2011) : 539-556.

[40] Leleu-Merviel, Sylvie. "Effets de la numérisation et de la mise en réseau sur le concept de document." Revue I3-Information Interaction Intelligence 4.1 (2004).

[41] Savoie-Zajc, Lorraine. "Comment peut-on construire un échantillonnage scientifiquement valide." Recherches qualitatives 5 (2006) : 99-111.

[42] Lopes, R., et al. "La géométrie fractale pour l'analyse de signaux médicaux : état de l'art." IRBM 31.4 (2010) : 189-208.

[43] Point, Sébastien, and Catherine Voynn et Fourboul. "Le codage à visée théorique." Recherche et Applications en Marketing (French Edition) 21.4 (2006) : 61-78.

[44] Sjare, B. L., and T. G. Smith. "The vocal repertoire of white whales, Delphinapterus leucas, summering in Cunningham Inlet, Northwest Territories." Canadian Journal of Zoology 64.2 (1986) : 407-415.

[45] Garnier Hugues. "Signaux Alèatoirs."Universiter De Lorraine, hugues.garnier@univ-lorraine.fr. (consulté le: 03/06/2023)

[46] Merten, Pascaline. "O'Hagan (Minako) et Mangiron (Carmen), Game Localization. Translating for the global digital entertainment industry, Amsterdam-Philadelphia, John Benjamins Publishing Company, 2013." Equivalences 43.1 (2016) : 181-188.

[47] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998) : 18-28.

[48] Houda KADI. "La Reconnaissance Automatique du Locuteur Par La Voix IP." Master02, Université Sidi Mohamed Ben Abdellah, juin 2014.

[49] Ridha. RAMDANI."Commande Vocale D'une Plateforme Mobile." Master02, Université Saad Dahlab De Blida, 2016/2017.

[50] Ayoub BENGHARABI, Elouanas BELABBACI. "Descripteurs Audio-Visuel Pour La Reconnaissance Des Marque De Téléphones Mobiles." Maser02, Université Akli Mohand Oulhadj-Bouira, 2020/2021.

[51] http://www.xavierdupre.fr/app/mlstatpy/helpsphinx/notebooks/split_train_test.html (consulté le: 07/06/2023)

ملخص

يصف مشروعنا عدة طرق للتعرف التلقائي علي السماعات سنقوم بتسجيل قاعدة بيانات تتكون من 60 ملفا صوتيا (نساء ورجال).يتكون هدا النظام من مرحلتين (التعلم والاختبار).

تركز دراستنا علي دراسة تطبيقات بعض الخوارزميات MFCC(معامل تردد تردد مقياس ميل)و PCA(تحليل المكون الرئيسي) كاستخراج معلمة.

نمدجة ناقلات الرسم البياني والخوارزمية بما في دلك تلك من الدكاء الاصطناعي (مثل آلات دعم SVM) من اجل التعرف علي الجنس (رجل وامرأة) من خلال بصمة حديثة (الصوت).

لهدا لاتوجد طريقة بسيطة، واشهرها هو الملعب،MFCC ,طاقة , PCA ,لكنها تعطي نتائج مقبولة اكثر واقل , لكننا وجدنا من خلال تطبيق الأساليب التي يمكن تحسينها معدل التعرف علي الجنس .

بعد تحليل النتائج التي تحصلنا عليها، وجدنا افضل توليفة ممكنة (للتقنيات التي اخترناها) هي التهجين بين MFCC و PCA, مما أدى الي زيادة الدقة.

الكلمات المفتاحية:

التعرف التلقائي، MFCC، الملعب. PCA, الصوت, النمدجة الخوارزمية

Résumé

.Notre projet décrit plusieurs méthodes de reconnaissance automatique de locuteur, nous allons enregistre une base des données qui se compose en 60 fichier sonores (femmes et hommes). Ce système compose de deux phases (apprentissage et teste).

Notre étude porte sur l'étudedes applications des certains algorithmes la MFCC (mel-scale fréquence cepstralcoeffecient) et PCA (Principal component analysis) comme extraction de paramètre. La modélisation vecteur histogramme et l'algorithme y'compris ceuxissus de l'intelligence artificielle (tel que SVM Support Vectors Machines)afin de faire une reconnaissance de genre(homme-femme) à travers l'empreinte de son speech (voix).

Pour cela il y'a pas mèl de méthode simple dont la plus connue est le pitch, MFCC,l'énergie PCA mais ça donne des résultats plus au moins acceptables mai on a trouvé à traversl'application des méthodes qu'on peut améliorer le taux de reconnaissance du genre.

Apres l'analyse des résultats qu'on a eus, on a trouvé que la meilleure combinaisonpossible (pour nous techniques choisis) est l'hybridation entre le MFCC et le PCA, Cequi a permet l'augmentation de la précision.

Mots clés :

Reconnaissance automatique, MFCC, Pitch, PCA, Audio, Algorithme combinée.SVM

Abstract

Our project describes several methods of automatic speaker recognition; we are going to save a database, which consists of 60 sound files (women and men). This system consists of two phases (learning and testing).

Our study focuses on the study of the applications of certain algorithms the MFCC (Mel-scale frequency cepstral coefficient) and PCA (Principal component analysis) as parameter extraction. The histogram vector modeling and the algorithm including those from artificial intelligence (such as SVM Support Vectors Machines) in order to make a gender recognition (man-woman) through the imprint of his speech (voice).

For this, there is no simple method, the best known of which is pitch, MFCC, PCA energy, but it gives more or less acceptable results, but we have found through the application of methods that can be improved. Gender recognition rate.

After analyzing the results, we had, we found that the best possible combination (for our chosen techniques) is the hybridization between the MFCC and the PCA, which has increased the precision.

Key words:

Auto Recognition, MFCC, Pitch, PCA, Audio, Combined Algorithm.SVM