Ordre...../F.S.S. A/UAMOB/2019

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
UNIVERSITE AKLI MOHAND OULHADJ-BOUIRA

جامعة البويرة

Faculté des Sciences et Sciences Appliquées
Département Génie électrique

Mémoire de fin d'étude

Présenté par :

BRAHIMI Mohamed **BOUZOUANE Rachid**

En vue de l'obtention du diplôme de **Master** en :

Filière : ELECTRONIQUE
Option : Electronique des Systèmes Embarqués

Thème :

Effet de la durée des enregistrements sur la biométrie vocale

Devant le jury composé de :

S LAADJOUZI	MAA	UAMOB	Président
A ALIMOHAD	MCB	UAMOB	Encadreur
R KASMI	MCA	UAMOB	Examinateur

Année Universitaire 2019/2020

REMERCIEMENT

*Nous remercions tout d'abord le bon **DIEU** le tout puissant pour la bonne santé, la volonté et la patience qu'il nous a donné durant ces années d'études.*

Nous remercions sincèrement, Dr. Abdennour ALIMOHAD, notre encadreur, pour ses conseils et suggestions avisés, qui nous ont aidés à mener à bien ce travail.

Nous tenons à remercier les membres du jury Dr. Ridha KASMI et M. Samir LAADJOUZI pour leur présence, pour leur lecture attentive de notre mémoire, ainsi que pour les remarques qu'ils nous adresseront lors de cette soutenance afin d'améliorer notre travail.

Nos remerciements et notre gratitude vont aux professeurs et

Enseignants de département génie électriques de l'Université de BOUIRA.

Nos remerciements vont également à nos très chers parents, et familles.

Enfin, nos remerciements à tous nos amis, collègues, et proches, qui nous ont toujours soutenus et encouragés au cours de la préparation de ce mémoire.

« Merci à toutes et tous ».

Dédicace

Je dédie ce modeste travail à la mémoire de mes parents et ma sœur

Amina,

À ma famille BRAHIMI et celle de mon épouse ABDERRAHIM

À mes frères à mon épouse et mes enfants Athman, Amira et Nihale.

M. BRAHIMI. Mohamed

Dédicace

Je dédie ce modeste travail

*Louange à **DIEU**, le seul et l'unique*

A mes très chers parents, mon père et ma mère, pour leurs soutiens tout au long de mes études.

A ma grande mère, A mon frère Ahmed et ma sœur Sabrina. A tous les membres de ma famille BOUZOUANE, à mes proches grands et petits,

A tous mes collègues et amis,

A tous ceux qui ont participés de

Près ou de loin dans la réalisation de ce travail,

A mes très chères.

Rachid

Sommaire

CHAPITRE II : BIOMETRIE VOCALE

CHAPITRE III : METHODOLOGIE ET RESULTATS

Liste des figures

Figures chapitre I

Figures chapitres II

Figures chapitres III

Liste des tableaux

Résumé

Dans ce mémoire de fin d'études, nous avons présenté un système de reconnaissance automatique de locuteur (RAL), basé sur le modèle GMM (Gaussian Mixture Model), et en fonction de la variabilité des courtes durées d'enregistrements de la parole. Nous cherchons à vérifier l'influence des enregistrements des voix de courtes durées sur un système (RAL), pour trois différentes méthodes d'extraction des paramètres acoustiques, ces méthodes sont MFCC, LPCC et PLP, qui donnent des coefficients les plus utilisés pour représenter le signal de parole.

Les résultats des simulations réalisées donnent des performances supérieures du système (RAL) en utilisant les coefficients MFCC par rapport aux coefficients LPCC et PLP, et cela pour une variabilité en courte durée des enregistrements des locuteurs.

Mots clés : Reconnaissance automatique de locuteur (RAL), paramètres acoustiques, courte durée, MFCC, PLP, LPCC.

ملخص

في إطار أطروحة التخرج، سنقدم نظامًا للتعرف التلقائي على المتحدث (ASR) وذلك بشكل استنادًا إلى نموذج GMM (Gaussian Mixture Model) واعتمادًا على تنوع الفترات القصيرة في تسجيلات الكلام. نسعى إلى مقارنة تأثير التسجيلات الصوتية قصيرة المدة على نظام (ASR) بثلاث طرق مختلفة تستخدم في استخراج المعاملات الصوتية. الطرق الثلاث المطبقة هي MFCC وLPCC و PLPوالتي تعطي المعاملات الأكثر استخدامًا لتمثيل إشارة الكلام.

تعطي نتائج عمليات المحاكاة التي تم تجربتها أداءً فائقًا للنظام (RAL) باستخدام معاملات MFCC مقارنةً بمعاملات LPCC وPLP، من أجل تباين قصير المدى في تسجيلات المتحدث.

الكلمات الرئيسية: التعرف التلقائي على المتحدث، المعاملات الصوتية، المدى القصير، MFCC، PLP، LPCC.

Abstract

In this graduation thesis, we presented an automatic speaker recognition (ASR) system based on the GMM model (Gaussian Mixture Model), and depending on the variability of short durations of speech recordings. We seek to compare the influence of short duration voice recordings on an ASR system for three different methods used in the extraction of acoustic parameters. The three methods are MFCC, LPCC and PLP, which give the most used coefficients representing the speech signal in an automatic speaker recognition task.

The simulations results carried out give superior performance of the ASR system by using the MFCC coefficients compared to the LPCC and PLP coefficients, for a short-term variability of speaker's recordings.

Keywords: Automatic speaker recognition (ASR), acoustic parameters, short term, MFCC, PLP, LPCC.

Liste des abréviations

ADN	:	Acide Désoxyribo-Nucléique.
CDTA	:	Centre de Développement des Technologies Avancées.
DCT	:	Discrete Cosin Transform.
DFT	:	Discrete Fourier Transform.
EER	:	Equal Error Rate.
EM	:	Expectation Maximisation.
FAR	:	False Acceptance Rate.
FRR	:	False Reject Rate.
FFT	:	Fast Fourier Transform.
HMM	:	Hidden Markov Models.
GMM	:	Gaussian Mixture Models.
LPCC	:	Linear Predictive Cepstral Coefficients.
MFCC	:	Mel Frequency Cepstral Coefficients.
MSSO	:	Méthodes Statistiques du Second Ordre.
PLP	:	Perceptual Linear Prediction.
PIN	:	Personal Identification Number.
RAL	:	Reconnaissance Automatique de locuteur.
RASTA-PLP	:	Relative Spectral-Perceptual Linear Prediction.
TER	:	Total Error Rate.
WAV	:	Waveform audio format

Introduction générale

La biométrie fait référence à l'identification automatique d'une personne en fonction de ses caractéristiques physiologiques ou comportementales. Cette méthode d'identification est préférée aux méthodes traditionnelles impliquant des mots de passe pour diverses raisons.

- La personne à identifier doit être présente lors de l'identification
- L'identification basée sur des techniques biométriques évite d'avoir à se souvenir d'un mot de passe.

Différents types de systèmes biométriques sont utilisés pour l'identification. Les plus populaires sont basés sur la reconnaissance faciale et la correspondance d'empreintes digitales. D'autres systèmes biométriques utilisent le balayage de l'iris et de la rétine, la parole, la reconnaissance facial et la géométrie de la main.

Particulièrement, la biométrie vocale consiste à développer des applications permettant de reconnaître l'identité d'un individuel utilisant sa voix.

Comme tout système de reconnaissance de formes, le système de reconnaissance du locuteur comporte deux phases, la phase apprentissage dans laquelle on crée une base de donnée contenant les modèles des individus et la phase test dans laquelle on peut vérifier l'identité proclamée d'une personne.

Dans ce travail, nous nous intéressons à l'étude de l'effet de la courte durée des enregistrements des locuteurs sur les performances d'un système RAL, en utilisant trois techniques d'extraction de paramètres acoustiques MFCC, PLP, et LPCC. Cette variabilité a été expérimentée dans les deux phases, apprentissage et test pour évaluer les trois types de techniques d'extraction de paramètres acoustiques.

Ce mémoire est organisé comme suit :

- Dans le premier chapitre, nous introduisons quelques définitions de la biométrie et différentes modalités biométriques.
- Pour le deuxième chapitre nous présentons un système de reconnaissance automatique du locuteur, qui se résume en trois étapes principales : l'analyse acoustique du signal parole, la modélisation du locuteur et la prise de décision.
- Le troisième chapitre est consacré à l'implémentation sous Matlab d'un système de reconnaissance de locuteur pour une variabilité de la courte durée des signaux acoustiques. Nous terminons ce travail par une conclusion générale et quelques perspectives.

Chapitre I

Généralités sur la biométrie

I.1. Introduction :

Nos voix ne sont pas seulement un moyen de communication. Elles offrent également un moyen fiable pour nous reconnaître, et font partie intégrante de notre identité. C'est la raison pour laquelle les banques et d'autres grandes entreprises se tournent aujourd'hui vers l'authentification vocale.

La voix humaine est unique. Elle est avec nous tout le temps contrairement à nos clés de voitures, et aux mots de passes ou codes PIN qu'on peut très souvent oublier. C'est à la fois cette sécurité et cette simplicité d'usage offerte par l'authentification biométrique vocale qui poussent les banques, les opérateurs de télécommunications et autres grandes organisations à choisir ce mode d'authentification [1].

La biométrie vocale, tout comme la reconnaissance et la synthèse vocale, s'est d'abord propagée dans les serveurs vocaux automatiques des centres d'appels. Mais aujourd'hui, elle est également utilisée dans des domaines aussi variés que l'authentification mobile et le paiement par cartes de crédit. Aussi dans les domaines :

- Sécurisation des applications mobiles
- Réinitialisation de mots de passes employés
- Sécurisation des transactions à risque par carte de crédit
- Paiement en ligne.

Ce ne sont là que quelques exemples où la biométrie vocale est utilisée pour faire de l'authentification forte avec un « outil » aussi simple et sûr que sa voix. Aussi dans ce chapitre nous avons abordé les différentes modalités biométriques, le concept général de la biométrie et nous le terminons par une introduction à la production vocale.

I.2. Définition :

Le mot **biométrie** signifie littéralement « mesure du vivant » et désigne dans un sens très large l'étude quantitative des êtres vivants. La biométrie est "toutes caractéristiques physiques ou traits personnels automatiquement mesurables, robustes et distinctives qui peuvent être utilisées pour identifier un individu ou pour vérifier l'identité prétendue d'un individu" [2].

La biométrie est une science qui porte sur l'analyse mathématique des caractéristiques physiques ou comportementales propres à chaque individu et permettent l'authentification ou l'identification de son identité.

I.3. Structure générale d'un système biométrique :

Un système biométrique est un système de reconnaissance des formes qui procède en premier par l'acquisition des données, puis extrait un ensemble de caractéristiques à partir de celles-ci, enfin il compare ces caractéristiques avec les modèles de la base de données. Selon le contexte de l'application, un système biométrique peut fonctionner soit en mode vérification ou d'identification. Tout système biométrique comporte deux processus qui se chargent de réaliser les opérations d'enregistrement et de tests [3] :

- **Processus d'enregistrement :** Ce processus a pour but d'enregistrer les caractéristiques des utilisateurs dans la base de données.
- **Processus de tests :** Ce processus réalise l'identification ou la vérification d'une personne.

Dans chacun des deux processus précédents le système exécute quatre opérations fondamentales, à savoir :

- **L'acquisition :**

On utilise un système d'acquisition pourvu d'un capteur pour acquérir une caractéristique spécifique de l'individu, par exemple : un microphone dans le cas de la voix.

- **L'extraction :**

Après avoir fait l'acquisition d'une image ou d'une voix, on réalise l'extraction de la caractéristique dont le processus d'authentification a besoin. Par exemple : extraire le visage du fond d'une image dans le cas de l'identification de visage.

- **La classification :**

En examinant les modèles stockés dans la base de données, le système collecte un certain nombre de modèles qui ressemblent le plus à celui de la personne à identifier, et constitue une liste limitée de candidats. Cette classification intervient uniquement dans le cas d'identification car l'authentification ne retient qu'un seul modèle.

- **La décision :**

En ce qui concerne l'authentification, la stratégie de décision nous permet de choisir entre les deux alternatives suivantes : soit que l'identité de l'utilisateur correspond à l'identité proclamée ou recherchée soit qu'elle ne correspond pas. Elle est basée sur un seuil prédéfini. L'estimation du seuil de la décision constitue la plus grande difficulté de ces techniques, et elle peut engendrer deux types d'erreurs, souvent prises comme mesures de

performances pour ces techniques d'authentification : faux rejet (FR) qui correspond à rejeter un vrai utilisateur ou une identité valable, et fausse acceptation (FA) qui donne accès à un imposteur.

I.4. Caractéristiques biométriques :

Le choix des caractéristiques physiques est important. Il faut qu'elles soient toutes à la fois [2].

- **Universelles :** les caractéristiques biométriques existent chez tous les individus.
- **Uniques (unicité) :** permettre de différencier un individu par rapport à un autre.
- **Permanentes :** (La robustesse) autoriser l'évolution dans le temps, et tout au long de la vie d'une personne.
- **Enregistrables :** collecter les caractéristiques d'un individu avec son accord.
- **Mesurables :** (La quantifiabilité) autoriser une comparaison future.

I.5. Les différentes techniques de la biométrie :

Les techniques biométriques se divisent en deux groupes selon la coopération ou non de l'individu [4] :

1. **Techniques intrusives :** Il y a un contact physique avec l'individu (l'iris, la rétine et les empreintes digitales, etc.)
2. **Techniques non intrusives :** Il n'y a pas un contact direct avec l'utilisateur (façon de marcher, signature, frappe de clavier, etc.).

I.6. Les catégories (modalités) de la technologie biométrique :

Dans les systèmes biométriques il y a trois catégories :

- **Analyses biologiques** : basé sur les caractéristiques biologiques des individus (Odeur, sang, salive, urine, ADN).
- **Analyses comportementale**s : elle se base sur l'analyse de certains traits personnels du comportement de l'individu comme la frappe sur le clavier, la voix, la manière de marcher (démarche)...
- **Analyses morphologiques** : elle est basé sur l'identification des traits physiques particulier, comme l'empreintes digitales, forme de la main, visage.

I.7. Modalités biométriques :

La biométrie est basée sur les caractéristiques biométriques de l'individu, ces caractéristiques peuvent être classées dans une des catégories citées précédemment (paragraphe I.6) :

a. **Empreintes digitales :** C'est la méthode biométrique la plus connue et la plus utilisée au monde [05]. Une empreinte digitale se compose principalement de trois caractéristiques : les arcs, les verticilles et les boucles. Il existe deux types empreintes :
   L'empreinte directe (qui laisse une marque visible) et l'empreinte latente (saleté, sueur ou autre résidu déposé sur un objet).
   La technologie de l'empreinte digitale fait le traitement rapide et fiable mais nécessite un contact physique. Cette technique est appliquée dans plusieurs structures comme les hôpitaux, les écoles, l'identification des criminelles, les aéroports, les cartes d'identité, les passeports...etc.

b. **Visage** : cette technique s'appuie sur les caractéristiques principales du visage. Un système de reconnaissance faciale est une application pour détecter et identifier un individu à travers des images ou des vidéos en se basant sur les caractéristiques de son visage. En général, les caractéristiques utilisées sont la position du nez, des yeux et de la bouche et de la distance entre ces différents organes [6]. Cette technologie n'est pas intrusive, ces taux de reconnaissance sont très élevés, et elle est peu coûteuse. Elle est utilisée dans de nombreuses applications liées à la sécurité (aéroports, identifier des criminels, et commerciales...).

c. **L'iris :** Cette technique a été identifiée au milieu des années 1980 comme une bonne méthode, L'iris est le muscle coloré à l'intérieur de l'œil (la zone colorée située entre le blanc de l'œil et la pupille), visible à travers la cornée, placé devant le cristallin et percé en son centre. Une caméra parcourt l'œil à l'aide d'une lumière infrarouge et capture une image [7]. L'iris est stable dans la vie mais la fiabilité du système diminue en fonction de la distance entre l'œil et la camera.

d. **La voix** : La reconnaissance de la voix n'est pas intrusive pour la personne et n'exige aucun contact physique avec le lecteur du système (Exemple Microphone). Le logiciel de reconnaissance peut être centralisé et la voix transmise par le réseau, d'où un impact de réduction des coûts. Le dispositif nécessite un microphone en source de capture. Les systèmes d'identification de la voix sont basés sur les caractéristiques de voix, uniques pour chaque individu. Ces caractéristiques de la parole sont constituées par une combinaison des

facteurs comportementaux (vitesse, rythme, etc…) et physiologiques (âge, sexe, fréquence, accent, harmoniques, …).

La voix présente des inconvénients qui peuvent influer sur la fiabilité et la robustesse des systèmes.

e. **Signature dynamique** : Chaque personne a un style d'écriture unique. On peut donc identifier une personne à partir de sa signature. Cette technique est utilisée dans beaucoup de pays comme élément juridique ou administratif. Elle analyse les caractéristiques dynamiques du processus de signature. Elle est basée sur des critères précis comme la pression, l'accélération, la souplesse, les courbes, et plusieurs autres paramètres [8]. On retrouve ce type de système dans différents secteurs comme moyen de vérification d'identité, aussi comme pour les services postaux et les banques.

f. **L'ADN :** Une empreinte génétique (L'acide désoxyribonucléique (ADN)), ou profil génétique, est le résultat d'une analyse génétique, rendant possible l'identification d'une personne à partir d'une petite quantité de ses tissus biologiques (bulbe de cheveux, sang, salive, sécrétion vaginale, sperme). C'est la technique la plus fiable mais elle est coûteuse.

g. **Veines de la main** : Cette technique sonde par infrarouge le dessin du réseau de veines soit du doigt ou de la main, et fonctionne en émettant une lumière infrarouge par des diodes, cette lumière est ensuite absorbée par les tissus de la peau et les vaisseaux sanguins [4]. Les caractéristiques des veines sont lues par une caméra infrarouge qui tire l'image en deux dimensions, cette image est enregistrée pour une comparaison future.

I.8. Les modes de fonctionnement de la biométrie :

L'enchaînement du fonctionnement d'un processus biométrique est détaillé comme suit :

- **L'apprentissage :** C'est la première étape de tout système biométrique, un utilisateur est enregistré dans le système pour la première fois et où une ou plusieurs modalités biométriques sont capturées et enregistrées dans une base de données.
- **Authentification (ou vérification) :** elle consiste à confirmer l'identité revendiquée par un utilisateur. C'est une comparaison ***« un pour un »*** dans laquelle le modèle biométrique saisi est comparé au modèle de référence [9].

- **Identification** : Permet de vérifier que l'identité d'un individu qui se présente existe bien dans la base de référence. C'est une comparaison ***« un pour plusieurs »*** où le modèle saisi est comparé à tous les modèles stockés dans la base [9].

I.9. Avantages et inconvénients d'un système Biométrique :

I.9.1. Avantages du système biométrique :

- **Facilité d'utilisation :** une analyse d'empreinte digitale, d'iris, ou de frappe sur le clavier est plus facile à utiliser qu'un mot de passe.
- **Sécurité améliorée :** Augmentation du niveau de sécurité pour permettre l'accès à un visiteur.
- **Ne peut être oublié :** comme cela dépend des caractéristiques physiques, le risque d'oubli est éliminé, ce qui peut arriver dans le cas du mot de passe.
- **Coût opérationnel moins élevé :** la biométrie nécessite un coût opérationnel très faible.

I.9.2. Inconvénients du système biométrique :

- **La biométrie dure toute la vie** : les caractéristiques physiques ne peuvent pas être modifiées.

  Les mots de passe peuvent être changés, mais la biométrie ne peut pas être réinitialisée une fois que le traitement a été modifié.
- **L'environnement et l'utilisation peuvent affecter les mesures :** un dommage dans les attributs physiques peut changer le modèle qui permet l'accès.
- **Nécessite une intégration** : pour enregistrer les données, une intégration matérielle supplémentaire est obligatoire.

I.10. Production du signal de parole :

Les mécanismes de production de la voix sont longtemps restés très mystérieux. Depuis deux siècles environ, l'évolution très rapide de la médecine et des techniques d'investigation a permis de mieux comprendre comment le son était généré avant d'être rayonné dans le milieu extérieur. L'ensemble de l'appareil vocal est schématisé (**figure I.1)**. Sa physiologie globale est décrite par trois étages distincts : la soufflerie, le larynx et les résonateurs [10].

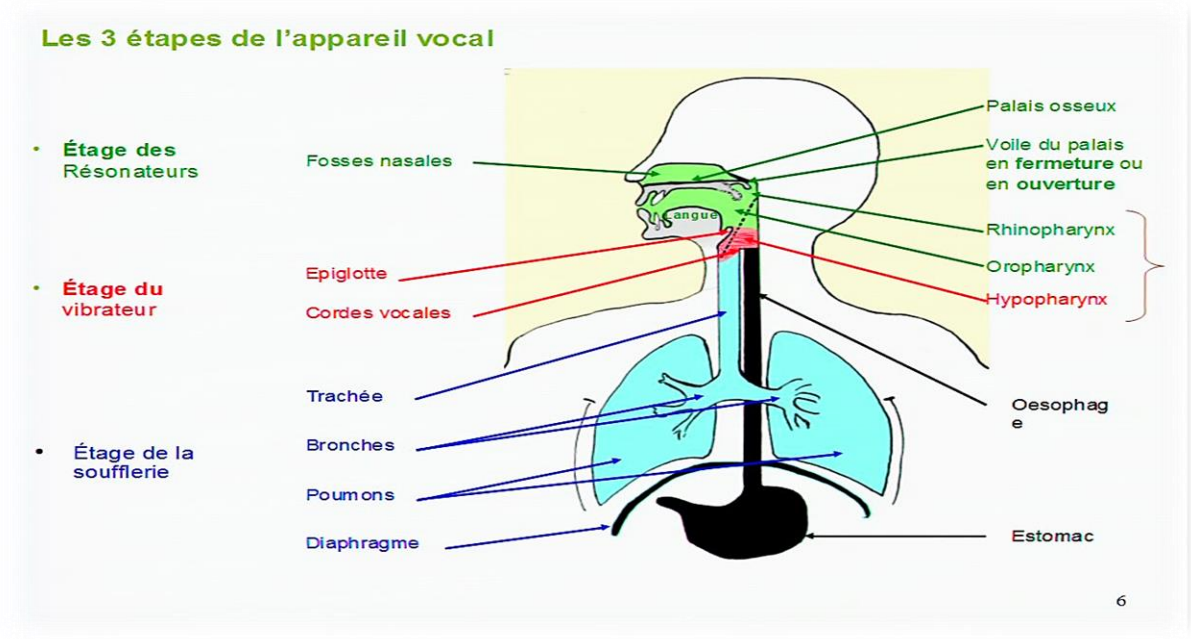

***Figure I.1**: Appareil vocal.*

I.10.1. Soufflerie :

Correspond à l'ensemble de l'appareil respiratoire (les poumons et les muscles respiratoires).

I.10.2. Larynx :

C'est un organe complexe situé dans le cou, au niveau de l'extrémité supérieure la trachée. Sa fonction principale est d'obturer le conduit respiratoire lors de la déglutition.

Les principaux muscles et cartilages laryngés sont présentés dans la figure I.2**.** Le cartilage thyroïde est situé à l'avant du larynx, l'une de ses proéminences correspond à la pomme d'Adam. Les cordes vocales correspondent à deux replis pouvant, lorsqu'elles sont en contact, venir obstruer totalement le conduit respiratoire. On appelle alors glotte l'espace entre les cordes vocales [10].

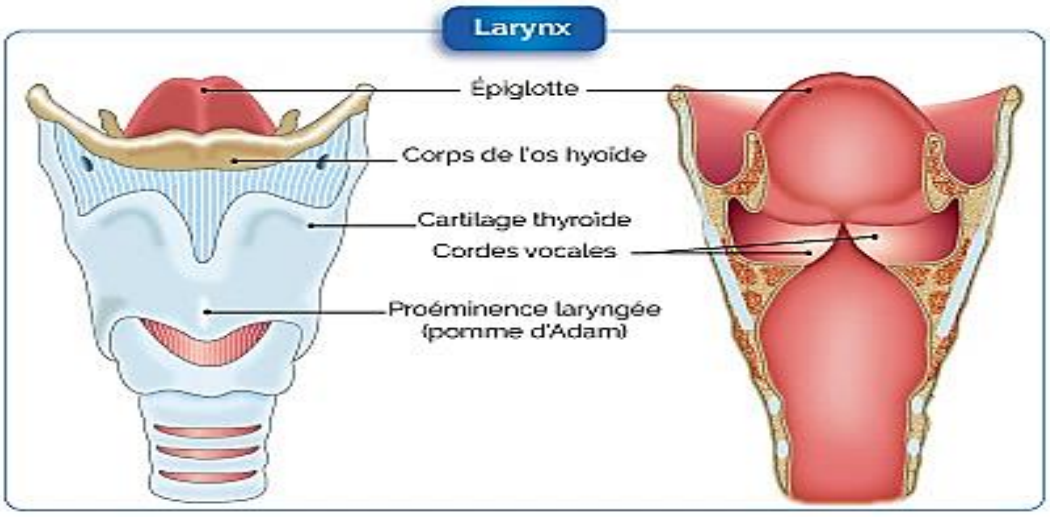

Figure I.2: *Schéma des muscles du larynx [11]..*

I.10.3. Résonateurs :

Le souffle produit la voix. La voix est alimentée par une colonne d'air dont la profondeur et la régularité déterminent la qualité vocale. Le diaphragme sert de base à cette colonne d'air, qu'il contrôle, alors qu'elle se soulève pour rejoindre les organes de la phonation. A mesure que l'air monte en poussant contre les plis vocaux, ces derniers se séparent momentanément pour lui laisser le passage. La poussée d'air ainsi que l'élasticité des plis vocaux les ramènent à leur position initiale. Cette production de vibrations est appelée « phonation ».

L'air sous pression s'infiltre dans la gorge, la bouche et le nez, entraînant des perturbations continues dans l'air ambiant appelées « ondes sonores ». Les voix diffèrent selon la dimension des plis vocaux et selon l'impact des résonateurs (gorge ou pharynx, bouche, fosses nasales ou nasopharynx) sur la tonalité de la voix. En renforçant le son produit par le passage de l'air dans le larynx, les plis vocaux donnent à la voix son timbre propre, qualifié de voix de poitrine. L'articulation résulte surtout des mouvements des lèvres, de la langue, de la mâchoire inférieure et du voile du palais, dont l'abaissement fait intervenir une cavité supplémentaire, les fosses nasales.

I.11. Les performances des systèmes biométriques :

En biométrie, chaque système est en face de deux populations [12] ; les clients appartenant au système, ceux qui sont autorisés à pénétrer dans la zone protégée, et les imposteurs n'appartenant pas au système, mais généralement qui essayent de rentrer.

Pour évaluer les performances d'un système biométrique plusieurs mesures sont employées :

- **FAR « *False Acceptance Rate* »** : c'est le taux de Fausses Acceptations : défini comme le nombre de Fausses Acceptations (*FA*) divisé par le nombre d'imposteurs dans la base N_i. FAR est calculé selon l'équation **(1.1)** :

$$FAR = \frac{FA}{N_i} \quad \text{..... (I.1)}$$

- **FRR *« False Reject Rate »*** c'est le taux de Faux Rejetés indique le nombre de Faux Rejets (*FR*) divisé par le nombre de clients dans la base *Nc*. *FRR* est calculé par L'équation **(1.2)** :

$$FRR = \frac{FR}{N_c} \quad \text{..... (I.2)}$$

- **TER *« Total Error Rate » :*** c'est le taux d'erreur totale d'un système biométrique. Cette mesure est calculée par la relation suivante :

$$TER = FAR + FRR \quad \text{.....(I.3)}$$

- **EER : *« Equal Error Rate »*** c'est le taux d'égale erreur, correspond à :

$$FRR = FAR \quad \text{..... (I.4)}$$

C'est-à-dire le meilleur compromis entre les faux rejets et les fausses acceptations. La relation entre FAR et FRR et le seuil T sont montrés, où on constate que si on choisit le seuil T faible, le système laissera passer tous les utilisateurs authentiques (clients), mais il laissera passer aussi les imposteurs facilement ce qui donne un système de faible sécurité. Si on choisit le seuil T fort, le système bloquera les imposteurs mais malheureusement bloquera aussi quelques clients. Par conséquence FRR augmente avec le seuil contrairement au FAR qui diminue. La figure I.3 illustre le FRR et le FAR à partir de distributions des scores et authentiques et imposteurs.

.

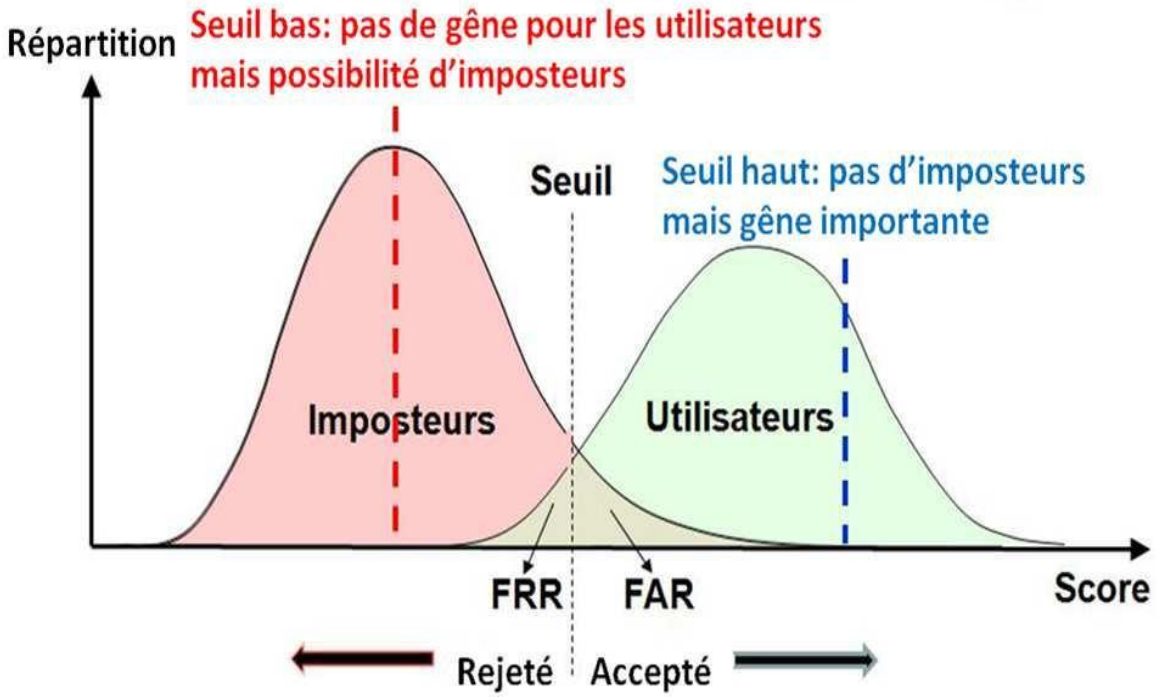

Figure I.3: *Illustration du FRR et du FAR.*

I.12. Conclusion

Dans ce chapitre nous avons décrit les technologies utilisées dans les systèmes biométriques pour l'identification ou l'authentification des personnes, et leurs caractéristiques, ainsi nous avons montré les différentes modalités biométriques et différentes applications, et aussi comment la production de signal vocal est faite.

Dans le chapitre suivant, nous allons étudier le système de la reconnaissance de locuteur.

Chapitre II

Biométrie vocale

II.1. Introduction

La forme du conduit vocal est caractéristique de la voix d'un locuteur donné et dépend de l'emplacement exact de chaque organe tout au long de la cavité du conduit vocal.
Cependant, la nature non-stationnaire du signal de la parole et sa grande variabilité fait en sorte que les motifs acoustiques générés pour un message donné varient avec le temps et le contexte (contenu linguistique, condition psychologique, maladie, âge, ...). De plus, il convient de noter que l'information d'identité d'un locuteur est une information non-linguistique incorporée dans le signal de parole, par conséquent, il est peu probable qu'une mesure de paramètres simples caractérise de façon unique un locuteur en tout temps.

Des études réalisées dans ce contexte mettent l'accent sur le caractère non-linguistique de l'information du locuteur et sur la non-existence de caractéristiques de parole simples à extraire qui contiennent exclusivement les informations discriminantes correspondant au locuteur. Cependant, certains outils d'analyse spectrale, à savoir les spectrogrammes, se sont avérés utiles pour l'analyse phonétique et ont aussi été utilisés avec succès pour la différenciation des locuteurs.

Cette approche a été validée par plusieurs travaux qui ont montré la pertinence des paramètres spectraux pour les tâches de reconnaissance automatique du locuteur. Motivés par les observations mentionnées ci-dessus, les systèmes de RAL utilisent principalement les paramètres spectraux de la forme d'onde du signal de la parole. Cette analyse est faite sur des segments de courte durée (typiquement entre 10 et 30ms) où le signal devient quasi stationnaire et les propriétés acoustiques, potentiellement uniques au locuteur, sont saisies [13].

On discute dans la section suivante le processus de paramétrisation du signal de parole et on explore les caractéristiques acoustiques à court terme ainsi que leurs avantages et points faibles.

II.2. Paramétrisation du signal de parole :

Le signal de parole est par nature, complexe et redondant et possède une grande variabilité ce qui le rend difficile à utiliser d'une manière directe par les systèmes de RAL. Cette complexité provient de la combinaison de plusieurs facteurs ; la grande variabilité inter-locuteur et intra -locuteur, les effets de la coarticulation en parole continue, les conditions d'enregistrement, etc. De ce fait, il devient nécessaire de procéder à une étape de paramétrisation qui a pour but d'extraire une représentation plus compacte de cette information acoustique qui réduit les redondances et permet d'accentuer les propriétés spécifiques au locuteur.

II.2.1. Caractéristiques des paramètres acoustiques :

Dans le cas idéal, les paramètres acoustiques utilisés en reconnaissance du locuteur doivent vérifier les conditions suivantes :

1. Ils se produisent naturellement et fréquemment dans la parole normale.
2. Ils sont facilement mesurables.
3. Ils varient autant que possible entre les différents locuteurs, mais sont aussi consistants que possible pour chaque locuteur.
4. Ils ne changent pas avec le temps et ne sont pas affectés par la santé du locuteur.
5. Ils ne sont pas affectés par un bruit de fond d'intensité raisonnable et ne dépendent pas du moyen de transmission.
6. Ils ne sont pas modifiables par l'effort conscient du locuteur, ou, au moins, peu susceptibles d'être affectés par des tentatives de déguisement de la voix.

Il est clair que satisfaire simultanément toutes ces conditions est difficile à accomplir en pratique. Cependant, ces critères peuvent être considérés comme des objectifs de conception idéalistes pour la caractérisation de la parole en RAL.

II.2.2. Propriétés acoustiques du conduit vocal :

Dans la littérature de traitement de la parole, le terme conduit vocal (ou tractus vocal) fait référence à la totalité de la cavité remplie d'air qui se trouve entre la glotte (entrée du larynx) et les lèvres. Cette cavité est plastique et dynamique, capable de prendre un nombre considérable de configurations et de changer très rapidement de forme. Ces modifications sont faites par le mouvement d'articulateurs (comme la langue).

En essence, la parole produite correspond à une variation de la pression d'air suite à la modulation de l'air sortant du larynx par l'activité des cordes vocales et des cavités dans le conduit vocal, produisant différents sons phonétiques. Le contenu en fréquence du signal acoustique est modifié par les propriétés de résonance des différentes cavités le long du trajet. Ces fréquences de résonance sont connues sous le nom de formants et sont généralement numérotées en allant des basses fréquences vers les hautes fréquences (F1, F2, F3, ...). La fréquence **Fo** est appelée fréquence fondamentale et correspond à la fréquence de vibration des cordes vocales.

En se basant sur ces propriétés acoustiques ainsi que sur la réponse fréquentielle de l'appareil phonatoire, un modèle dit source-filtre a été établi pour décrire les mécanismes et les propriétés de production des sons et modéliser le couple {cordes vocales, cavités supra-glottiques} en {source, filtre}. Ce genre de modèles permet d'assimiler la réponse de l'appareil

phonatoire (plus précisément celle des cavités supra-glottiques) à celle d'un filtre qu'il est possible de modéliser et comparer entre différents locuteurs.

Dans ce modèle (détaillé dans la Figure II.1), le spectre de la source glottique est harmonique du fait de sa périodicité (le son contient de l'énergie à la fréquence fondamentale de vibration des cordes vocales F_0 ainsi qu'aux fréquences $2 \times F_0$, $3 \times F_0$, ...$n \times F_0$) [12]. L'énergie diminue toutefois avec la fréquence (Figure II.1 (A)). L'effet de filtrage des cavités buccale et nasale sur la source glottique est représenté par une fonction de transfert. (La Figure II.1 (B) donne un exemple qui contient trois fréquences de résonance). Suite au passage de l'air par le conduit vocal, il en résulte un spectre (toujours harmonique) dont l'énergie varie avec la fréquence (Figure II.1 (C)) [13].

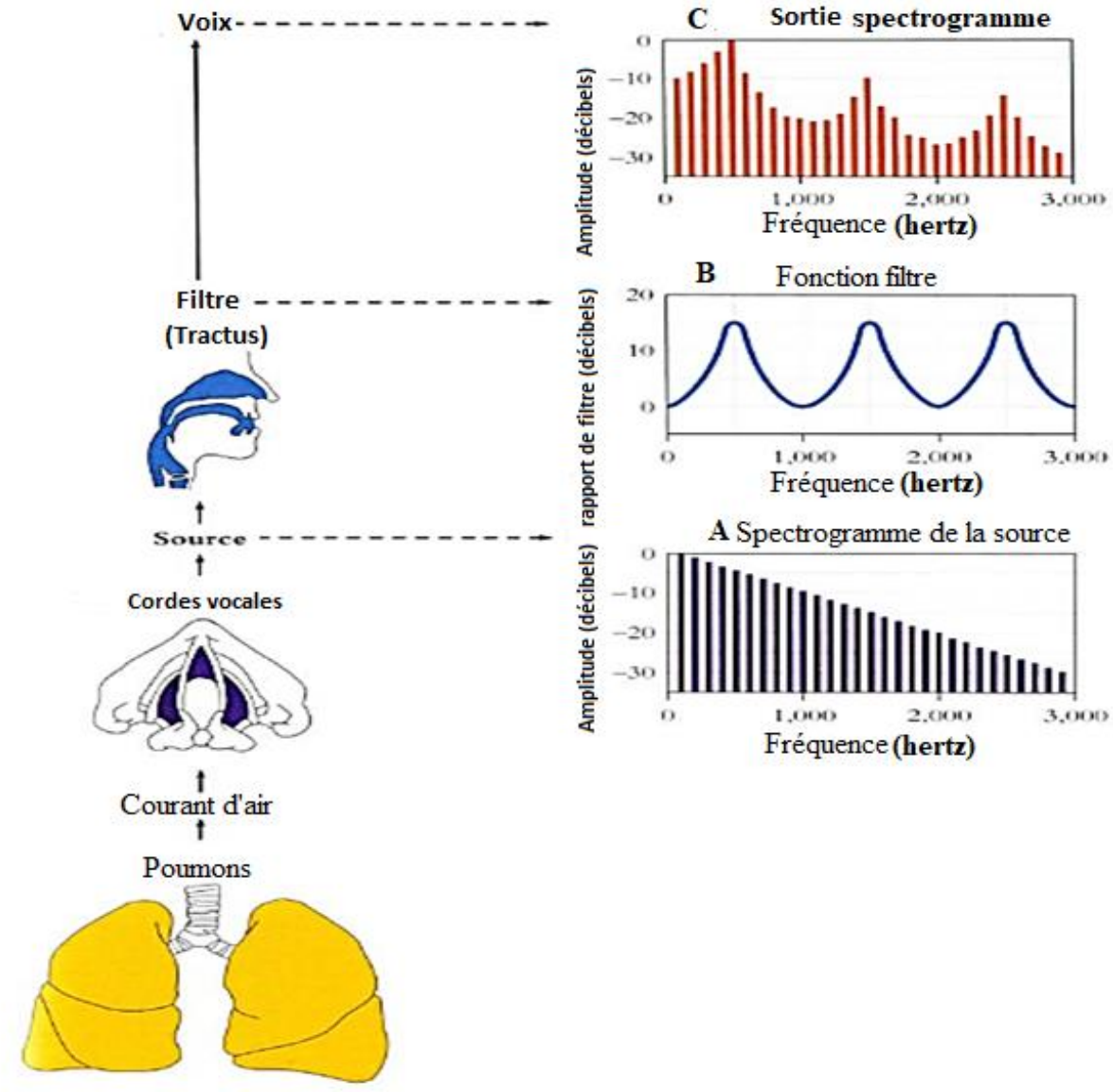

Figure II.1 : *Aperçu détaillé du modèle source-filtre [13].*

Cette modélisation permet de réduire la forme complexe du signal de la parole à un vecteur de paramètres (les coefficients du filtre). Ces paramètres permettent de décrire la réponse fréquentielle du conduit vocal et peuvent être utilisées pour la caractérisation du locuteur.

II.3. Modèle source filtre de la parole :

Le modèle source filtre de la parole est au cœur de plusieurs analyses de la parole. L'idée de ce modèle est que les sons sont produits par l'action d'un filtre, représentant le conduit vocal, sur la source sonore qui peut être au niveau de la glotte ou une autre constriction dans le conduit vocal **(Figure II.2)** [14].

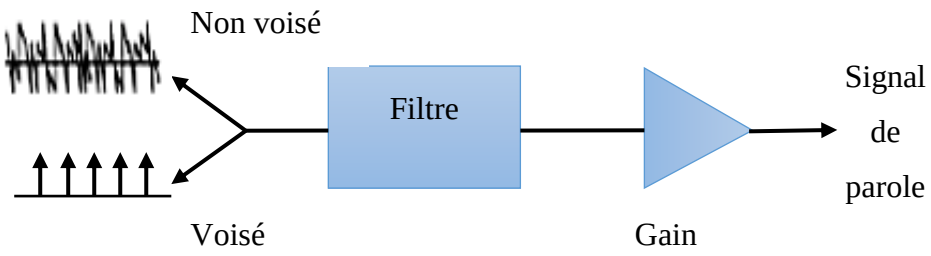

Figure II.2 : *Modèle source - filtre de la parole.*

Ce modèle se base sur la supposition que la source et le filtre sont indépendants. Cette supposition implique que la modification des propriétés des filtres ne change pas les propriétés de la source et vice versa. Quoique ce ne soit pas strictement vrai dans tous les cas, en pratique, cette supposition donne un modèle d'une grande utilité et largement précis de la production de la parole.

II.3.1. Source :

D'un point de vue acoustique, les sources peuvent correspondre aux sons voisés et non voisés de parole.

La source de la parole voisée est la vibration des cordes vocales en réponse à un courant d'air provenant des poumons. Cette vibration est périodique. Son examen montre qu'elle est constituée d'une série de larges pointes **(Figure II.3).**

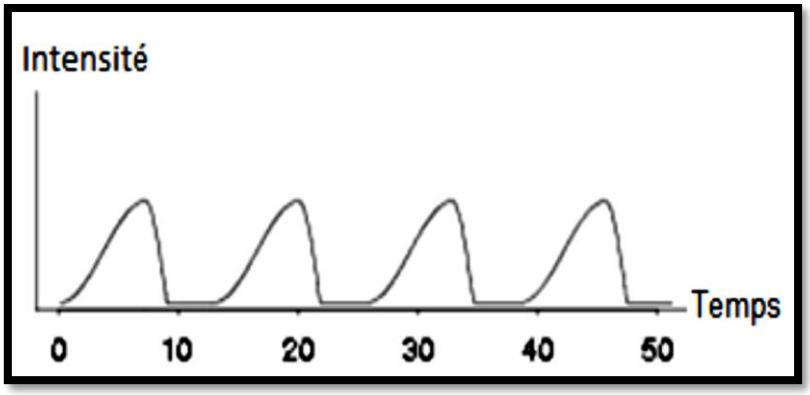

Figure II.3 : *Forme périodique de la source de parole.*

Le spectre de la source glottique est constitué de pics de fréquences correspondant aux harmoniques de la fréquence fondamentale de vibration des cordes vocales **(FigureII.4).**

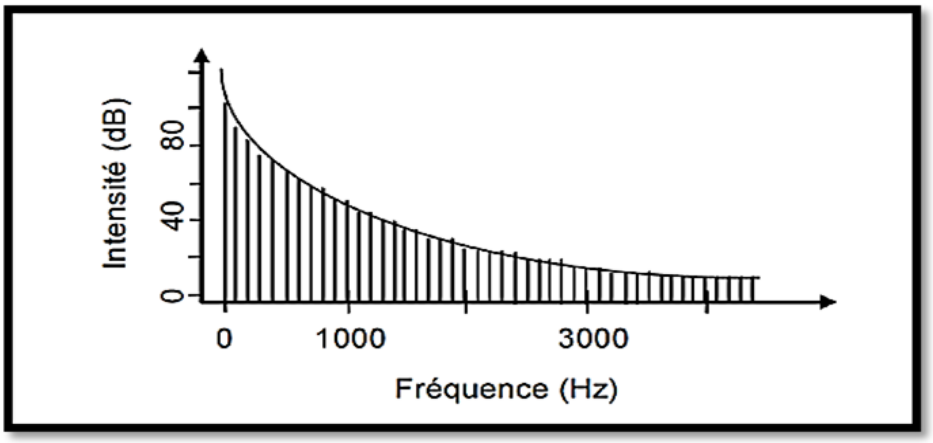

Figure II.4 : *Spectre de la source glottique.*

L'amplitude du spectre décroit avec l'augmentation de la fréquence. La fréquence fondamentale de vibration des cordes vocales dépend de la masse et la tension de ceux-ci. Elle est de l'ordre de 100 Hz, 200 Hz, et 300 Hz pour les hommes, les femmes, et les enfants, respectivement.

La source de parole non voisée est créée par une vibration non régulière des cordes vocales. Ces vibrations sont causées par un flux d'air turbulent.

II.3.2. Filtre :

En général, un filtre est un système qui altère la composition fréquentielle d'un signal d'entrée. En production de la parole, le filtre est le conduit vocal. Ce dernier est considéré comme un tube acoustique constitué d'un assemblage de sous tubes de sections variables qui est fermé d'un côté (par la glotte) et mesure, en moyenne, environ 17.5 cm pour l'homme (Figure II.5) [14].

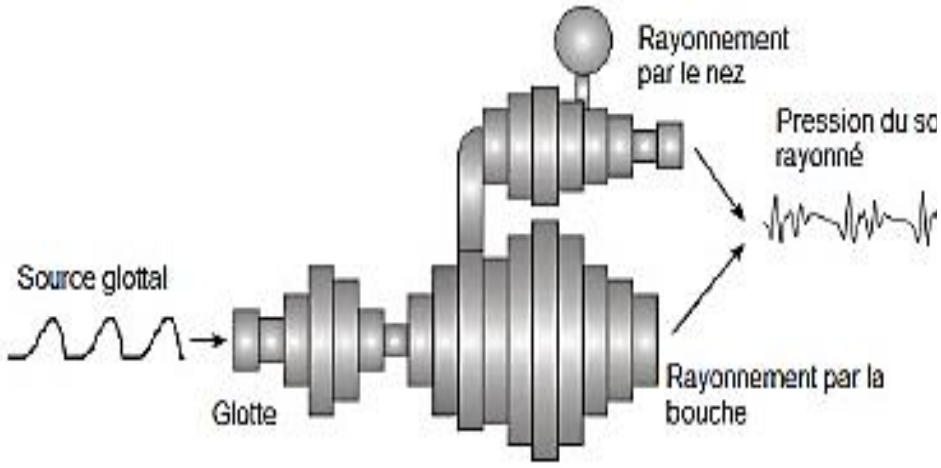

Figure II.5 : *Conduit vocal à tubes.*

Comme pour tout filtre, ce tube est caractérisé par un spectre. Essentiellement, ce spectre change quand la forme du conduit vocal change durant la production de la parole. Par conséquent, différents sons sont produits par le changement de la forme du conduit vocal, qui donne un ensemble particulier de caractéristiques du filtre. L'espace du conduit vocal, composé de cavités orale et nasale, peut être vu comme un filtre acoustique à temps variable qui amplifie certaines fréquences du spectre et attenue d'autres. Les fréquences de résonance du conduit vocal sont appelées formants. Ces formants dépendent des phonèmes, mais aussi, de la forme générale, de la longueur, et du tissu du conduit vocal. Les sons voisés sont composés d'environ 3 à 5 formants [14].

II.4. Analyse du signal de parole :

L'analyse des signaux de parole prend en considération les façons par lesquelles les sons de parole sont produits. Les signaux de parole résultent de l'interaction de deux facteurs ; la source glottique et le conduit vocal. Par conséquent, le signal de parole est obtenu par une source passant dans un filtre linéaire à temps variable, la source représente le flux d'air au niveau des cordes vocales et le filtre représente les résonnances du conduit vocal qui change dans le temps. La pression du flux d'air rayonné par la bouche et / ou le nez est convertie en un courant électrique à travers un microphone.

Ce signal de parole électrique utilise essentiellement des fréquences allant de 100 Hz à quelques 8000 Hz et des amplitudes variant de 30 à 90 dB [14].

L'analyse des signaux de parole est le processus d'estimation des paramètres d'un modèle, variant dans le temps, dans le but d'extraire des informations sur la production de ce signal. Les méthodes modernes d'analyse des signaux sont basées sur le traitement numérique des signaux.

La première raison de la numérisation est la facilitation des techniques de traitement sophistiquées qui sont très difficiles, voire impossible à réaliser en analogique.

La numérisation implique l'échantillonnage pour discrétiser le temps et la quantification pour discrétiser l'amplitude. Le taux avec lequel le signal analogique est échantillonné, est appelé fréquence d'échantillonnage. Dans les réseaux de télécommunications les signaux de parole analogiques sont limités entre 300 et 3400 Hz, la fréquence d'échantillonnage est dans ce cas de 8000 Hz, conformément au théorème de Nyquist. Pour une qualité supérieure, la bande passante fréquentielle de la parole est limitée entre 0 et 7000 Hz, dans ce cas la fréquence d'échantillonnage est choisie égale à 16 KHz.

Le signal échantillonné est ensuite quantifié en amplitude en utilisant un convertisseur analogique - digital permettant de représenter chaque échantillon réel en un nombre limité de bits. 12 bits sont nécessaires en pratique pour assurer un rapport signal sur bruit supérieur à 35 dB [14].

II.4.1. Analyse Temporelle :

Le signal vocalique n'est pas strictement périodique, mais beaucoup d'analyses font une hypothèse de stationnarité du signal sur un temps d'environ 10ms à 30ms. Cette hypothèse permet l'analyse des signaux vocaux en les considérant comme périodiques. L'intervalle de temps minimum nécessaire à la reproduction de la forme du signal est appelé période du signal, noté T_0. L'inverse de cette période $F_0 = \frac{1}{T_0}$ est la fréquence fondamentale du signal, exprimée en Hz.

II.4.2. Analyse fréquentielle :

En traitement de la parole, la plupart des paramètres se trouvent dans le domaine spectral. Le signal de parole est plus facile à analyser et plus systématique en fréquence qu'en temps. Aussi, le spectre de sortie du modèle de production de la parole n'est rien d'autre que le produit de la réponse fréquentielle du conduit vocal et le spectre de l'excitation. Donc, il est évident que cette sortie spectrale reflète les propriétés des réponses fréquentielles de l'excitation et du conduit vocal.

Vu la nature non stationnaire des signaux de parole, due au changement du système phonatoire dans le temps, il est impératif de considérer des séquences courtes en temps tel que les caractéristiques du conduit vocal restent inchangées, on dira que le signal est quasi stationnaire.

II.4.3. Analyse spectrographique de la parole :

Dans la recherche phonétique acoustique, deux représentations utiles qui sont la forme d'onde et un espace-temps-fréquence appelée le spectrogramme. La forme d'onde montre les variations de pression d'air, tandis que le spectrogramme montre l'importance des différentes fréquences en fonction de temps. Des exemples de spectrogrammes sont montrés sur les **figures II.6** et **II.7** montrent l'importance de la fréquence (f) au temps (t) de sorte que des régions plus foncées correspondent à des grandeurs plus élevées.

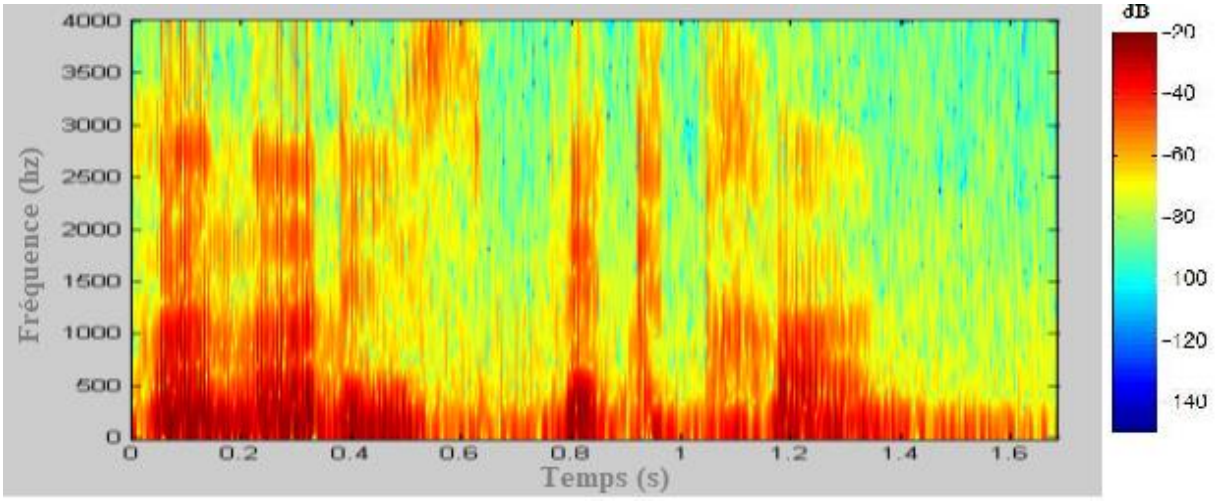

Figure II.6 : *Spectrogramme à large Bande[15].*

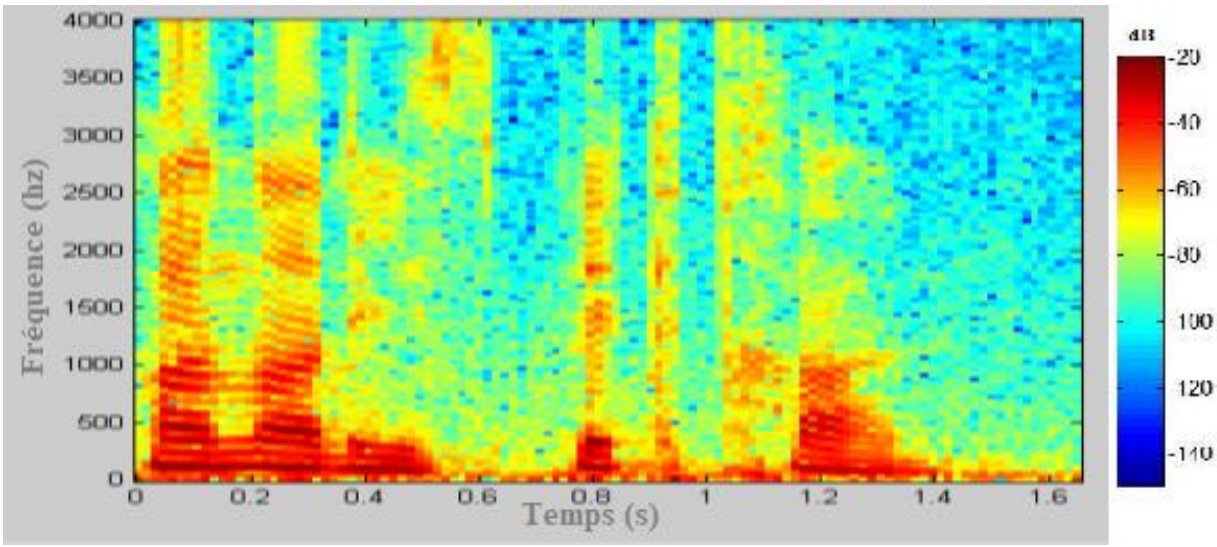

Figure II.7 : *Spectrogramme à Bande étroite[15].*

Il y a deux types de spectrogrammes : les spectrogrammes à large bande et ceux à bande étroite. Dans les spectrogrammes à large bande, la largeur de bande du filtre d'analyse est autour de 300 hertz et l'espacement temps est ainsi approximativement de 1/300 s = 0.0033s pour l'analyse à bande étroite, la largeur de bande est autour 50 hertz et l'espacement temps est ainsi

autour de 1/50 s = 0.020 s. Les spectrogrammes à large bande conviennent au cheminement des formants des voyelles tandis que les spectrogrammes à bande étroite peuvent être employés dans l'évaluation de F_0 [15].

II.4.4. Analyse cepstrale :

L'étude de la production de la parole nous a conduit à une supposition fondamentale qui représente le signal de parole comme sortie d'un système linéaire invariant dans le temps. Par conséquent, l'excitation et le filtre du système sont reliés par le produit de convolution. L'analyse cepstrale est une transformation qui vise à séparer la source du filtre d'un système [14].

II.5. Système de Reconnaissance de Locuteur :

La reconnaissance du locuteur est divisée en deux grandes classes : l'identification et la vérification **(Figure II.8**).

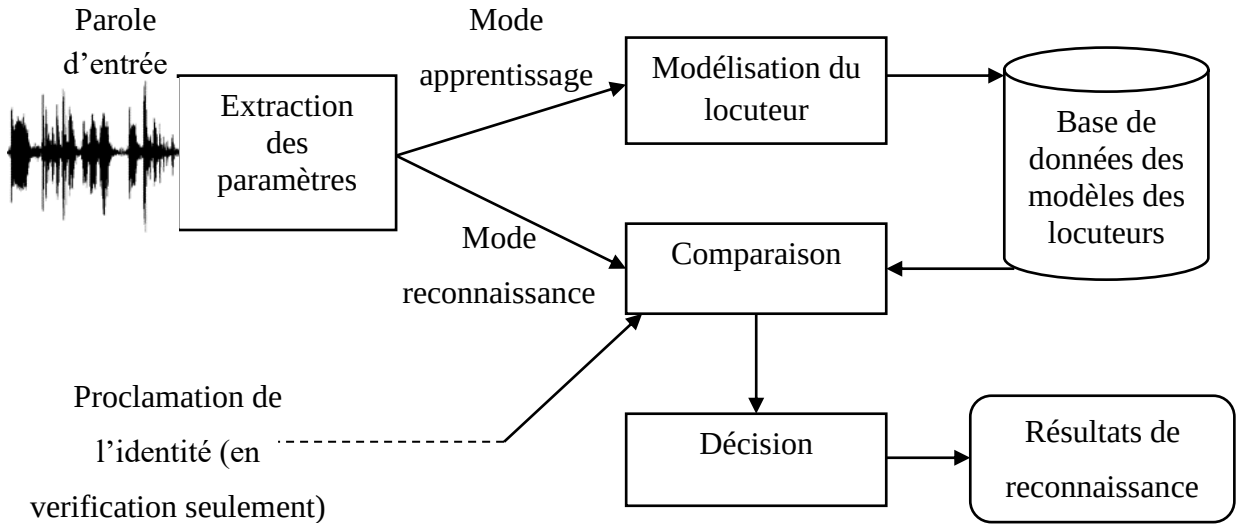

Figure II.8 : *Système de reconnaissance du locuteur [14].*

- En identification, aucune identité n'est prétendue par le locuteur. Le système doit déterminer, automatiquement, celui qui parle. Si le locuteur appartient à un ensemble prédéfini de locuteurs connus, cette approche est dite identification du locuteur dans un ensemble fermé. En général, le système doit traiter des cas où les locuteurs peuvent ne pas être modélisés dans la base de données. Dans ce cas l'approche est dite identification du locuteur en ensemble ouverte.
- En vérification de locuteur, le système doit déterminer si oui ou non la personne est celle qui prétend l'être. Cela implique que l'utilisateur doit fournir une identité, le système

accepte ou rejette cette personne selon que la vérification a réussi ou échoué. Une autre classification des systèmes de reconnaissance de locuteur divise ceux-ci en systèmes dépendants ou indépendants du texte. Les systèmes de reconnaissance dépendants du texte sont dits "avec contrainte".

Dans ce cas, on demande à l'utilisateur de prononcer soit un mot fixe (comme un mot de passe) ou une phrase. Cette information peut améliorer les performances du système de reconnaissance. Le système de reconnaissance indépendant du texte est sans contrainte. Dans ce cas, le locuteur doit être reconnu quel que soit le texte prononcé, ce qui constitue un problème plus difficile à résoudre [14].

De ce qui précède on dira qu'un système RAL comprend trois modules fondamentaux, une unité d'extraction de caractéristiques, qui transforme le signal de parole sous une forme compacte, une unité de modélisation statistique pour caractériser les caractéristiques extraites, et enfin un module de classification pour classifier une parole de test [16].

II.5.1. Module d'extraction des caractéristiques de la parole :

Les systèmes RAL utilisent trois grands types de techniques d'extraction de caractéristiques : l'analyse sous-segmentaire, segmentaire et supra-segmentaire. Les signaux vocaux, analysés à l'aide de la taille de l'image avec décalage dans la plage de 3 à 5 ms, sont appelés analyse sous-segmentaire.

Analyse segmentaire, la parole est fenêtrée avec une taille d'image et un décalage de l'ordre de 10 à 30 ms pour extraire les informations du locuteur caractérisant principalement le tractus vocal. Les informations sur les voies vocales spécifiques au locuteur peuvent être supposées stationnaires pour les analyses et le traitement pratiques lorsque les trames de taille et de décalage sont maintenues dans la plage de 10 à 30 ms.

Pour l'analyse supra-segmentaire, extraction de caractéristiques, la parole est tronquée en utilisant la taille de l'image et le décalage dans la plage de 100 à 300 ms.

Principalement, cette technique est utilisée pour analyser et extraire les caractéristiques des traits comportementaux du locuteur. Celles-ci intègrent les informations sur la durée des mots, l'intonation, la vitesse de parole, l'accent.

De nombreux paramètres à court terme ont été développés au fil des années pour des applications de reconnaissance de la parole puis utilisés en reconnaissance du locuteur. Ces modules utilisent généralement les techniques classiques d'extraction de caractéristiques acoustiques telles que Perceptual Linear Prediction (PLP) (Hermansky 1990), Linear Prediction Coding (LPC) (Atal et Hanauer 1971), Linear Prediction Cepstral Coefficients (LPCC) (Atal 1974) et Mel Frequency Coefficients Cepstral (MFCC) (Davis et Mermelstein 1980) [17].

- **Analyse prédictive linéaire cepstrale (LPCC)** : *Linear Predictive Cepstral Coefficients*) : C'est une méthode d'analyse cepstrale paramétrique qui repose sur un modèle source-filtre [12]. Dans cette approche, la réponse fréquentielle du conduit vocal est représentée par un filtre et chaque échantillon de parole est exprimé comme une combinaison linéaire d'échantillons précédents (d'où le nom analyse prédictive linéaire). Une fois estimés, les paramètres du filtre sont convertis en coefficients cepstraux. Les coefficients cepstraux résultants sont utilisés comme paramétrisation à court terme du signal de parole.

Le processus de l'extraction des paramètres LPCC est illustré par la Figure II.9 :

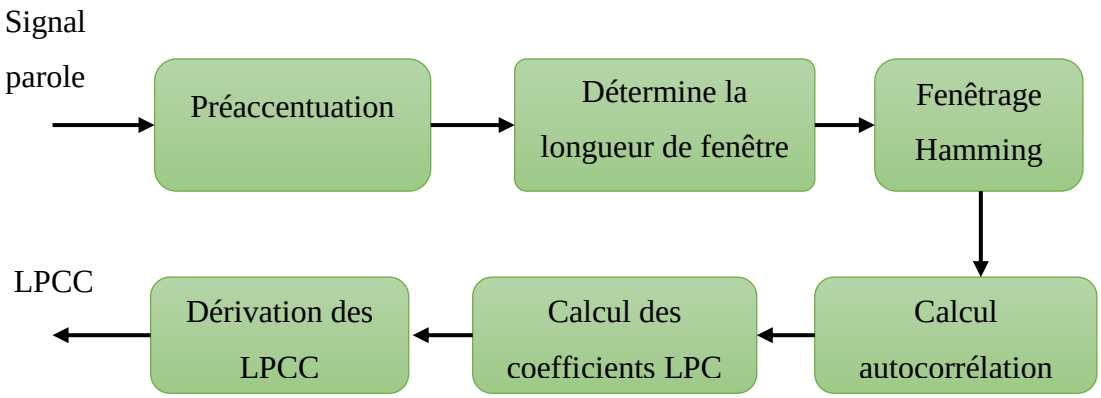

Figure II.9 : *Méthode de calcul des coefficients LPCC.*

Les étapes de calcul des coefficients LPCC sont montre comme suit [18] :

Le signal d'entrée est d'abord pré-accentué en utilisant un filtre passe-haut de premier ordre, puisque l'énergie est contenue dans la parole. Le signal est plus distribué dans les basses fréquences que dans le fréquences plus élevées. Afin de rehausser les énergies contenues dans les hautes fréquences, la préaccentuation du signal est effectuée.

La fonction de transfert de ce filtre dans le domaine z est exprimée comme :

$$H_p(z) = 1 - az^{-1} \quad \text{(II.1)}$$

Où, (coefficient de filtre) est une constante avec une valeur typique de 0,97. Le signal pré-accentué est bloqué dans les trames. Afin de réduire les discontinuités de signal aux bords de chaque trame, le fenêtrage du signal est effectué. La fenêtre couramment est utilisée Hamming et est décrit dans l'équation ci-dessous :

$$W(n) = 0.54 - 0.46\cos\left(2\pi\frac{n}{N}\right); \quad 0 \le n \le N \quad \text{(II.2)}$$

Où N est la longueur de la fonction de fenêtrage. Linéaire l'analyse prédictive est fondée sur l'hypothèse que le la forme du tractus vocal décide du caractère du son produit. Un filtre numérique multipolaire est utilisé pour modéliser la voie vocale est a une fonction de transfert représentée dans le domaine z donnée comme suit :

$$V(Z) = \frac{G}{1-\sum_{k=1}^{p} a_k z^{-k}} \quad \text{.....(II.3)}$$

Où V(z) est la fonction de transfert de la voie vocale. G est le gain du filtre,$a_k$est l'ensemble des coefficients d'auto-régression coefficients de prédiction linéaire connus (LPC), p est l'ordre de filtre omnipolaire. Une des méthodes efficaces d'estimation des coefficients LPC et le gain du filtre. La dernière étape de cet algorithme est l'analyse cepstrale qui se réfère au processus de découverte du cepstre de séquence de la parole. Fondamentalement, il existe deux types de Cepstral approches FFT cepstrum et LPC cepstrum. Dans l'ancien cas le cepstre réel est défini comme la transformée FFT inverse du logarithme du spectre d'amplitude de la parole défini par équation suivante :

$$\hat{S}[n] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} ln[S(\omega)]\, e^{j\omega n} d\omega \quad \text{..... (II.4)}$$

- **Analyse perceptive linéaire** (**PLP :** *Perceptual Linear Prediction*) : Cette méthode paramétrique se base aussi sur un modèle source-filtre. La paramétrisation PLP est identique à l'analyse LPC, sauf que les caractéristiques spectrales sont transformées pour correspondre aux caractéristiques du système auditif humain. La figure II.10 résume le principe de la méthode dont une analyse spectrale est effectuée au signal parole afin d'obtenir un spectre suivant une échelle d'audition.

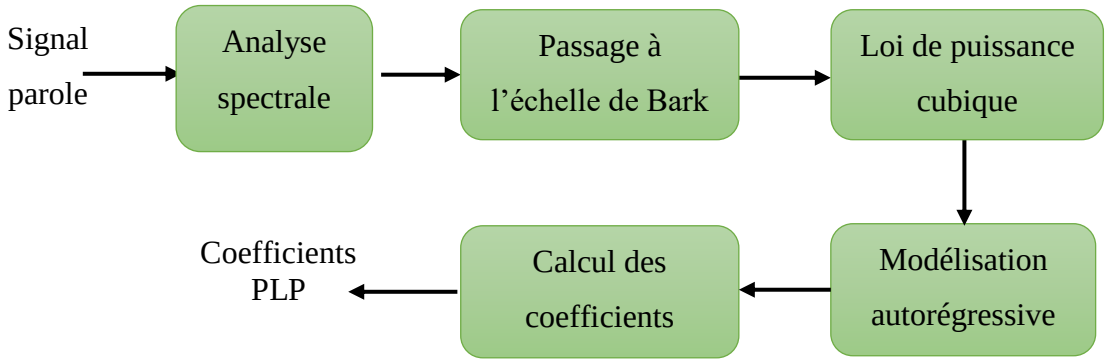

Figure II.10: *Méthode de calcul des coefficients PLP.*

Une analyse par un banc de filtres, appelées "bandes critiques", est effectuée. Ainsi, ce banc de filtres est appliqué dans l'échelle Bark qui s'exprime en fonction de fréquence par[14] :

$$B(f) = 6ln(\frac{f}{600} + (\frac{f}{600} + 1)^{1/2} \qquad (II.5)$$

f Étant la fréquence en Hz.

- **Analyse en banc de filtres Mel (MFCC :** *Mel Frequency Cepstral Coefficients*) : C'est une technique d'analyse cepstrale non-paramétrique qui utilise l'échelle Mel pour refléter la perception non-linéaire (linéaire jusqu'à 1000Hz et logarithmique au-delà de 1000 Hz) des fréquences par l'oreille humaine. C'est une échelle qui consiste en la définition de bandes critiques de perception (à l'aide d'un banc de filtres). Elle correspond à la distribution fréquentielle de l'oreille humaine. Le calcul des coefficients MFCC est montré dans la Figue II.11.

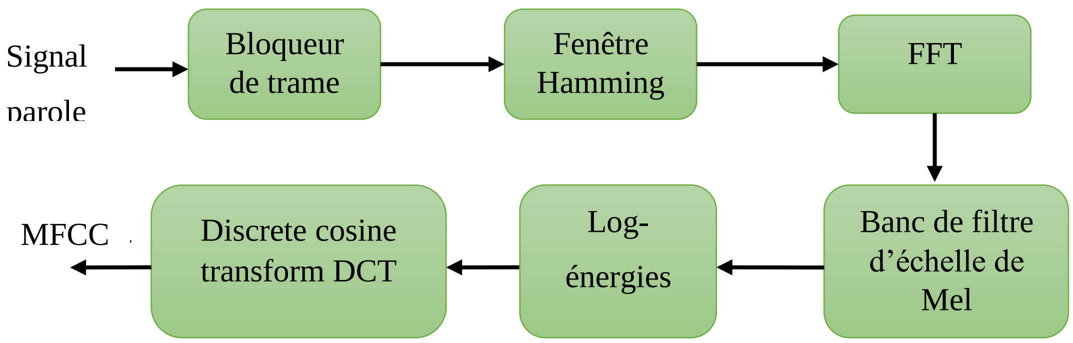

Figure II.11 : *Calcul des coefficients MFCC avec une échelle Mel.*

Les étapes de processus d'extraction des paramètres sont détaillées comme suit [19] :

- ✓ **Préaccentuation :**

C'est l'étape première, consiste à découper le signal en trames chevauchées de faible durée où le signal est considéré comme quasi stationnaire, et avec des trames de 10 à 30ms.

Le signal représente à partir une famille S(n) avec n [1, N] où N est le nombre d'échantillons dans le signal, la formule de préaccentuation d'un signal est :

$$S_1(n) = S(n) - \alpha . S(n-1) \qquad \text{et } n = 1, \dots, N \qquad (II.6)$$

- ✓ **Fenêtrage Hamming :**

Applique sur chaque trame une pondération de Hamming, est par la formule suivant :

$$S_2(n) = S_1(n).(0.54 - 0.46.\cos(2\pi.n/(N-1) - \pi)) \quad \text{avec } n = 0, \dots, N-1 \qquad (II.7)$$

- ✓ **Transformée de Fourier rapide (Fast Fourier Transform, FFT) :**

Applique la transformer de fourrier(FFT), on convertir ainsi chaque trame de domaine temporel au domaine fréquentiel. La DFT sur $i^{éme}$ trame est définie par la formule suivante :

$X_k = \sum_{n=0}^{N-1} x_n . e^{-2\pi jkn/N}$ avec $k = 0, \dots, N-1$ ……(II.8)

où j=$\sqrt{-1}$

avec X_k le spectre du signal numériquex_n.

- ✓ **Banc de filtres Mels :**

Utilise Banc de filtre pour transformer fréquentiel vers l'échelle de MELL, Le passage de l'échelle fréquentielle à l'échelle de Mel est régi par l'équation suivante :

$Mel(f) = xlog(1 + \frac{f}{y})$…… (II.9)

On trouve différentes valeurs pour *x* et *y* :

$$x = \frac{1000}{log2}, \qquad y = 1000$$

$$x = 2595, \qquad y = 700$$

- ✓ **Logarithme énergie :**

Utilisé pour mesurer cette quantité est le calcul du logarithme de l'énergie ($log\ E_i$) des données d'une trame de la parole dans le domaine temporel. Telle que :

$log\ E_i = log\ \sum_{n=1}^{N} s_n^2$…… (II.10)

Où s_n et N sont le $n^{ème}$ échantillon et le nombre d'échantillons de la trame i.

- ✓ **Transformation en cosinus décrète inverse (DCT) :**

C'est l'étape finale, on transforme les données dans l'échelle des Mels vers l'échelle des temps, est pour assurer le retour de domaine temporel.

Malgré les efforts qui ont été investis dans la conception de paramètres acoustiques plus pertinents pour la reconnaissance automatique du locuteur et plus robustes aux distorsions acoustiques, les paramètres MFCC et PLP restent des techniques d'extraction de

caractéristiques largement utilisées en raison de leurs performances considérables et de leur moindre complexité de calcul [13].

II.5.2. Modules de modélisation :

La tâche de reconnaissance du locuteur implique la comparaison d'un locuteur inconnu avec des locuteurs connus dans une base de données. Sur la base de cette comparaison le locuteur correspondant sera choisi. L'utilisation directe de vecteurs de paramètres représentant chaque locuteur n'est pas très pratique lorsque les vecteurs d'apprentissage sont larges. La modélisation est un moyen efficace de compression des données permettant d'obtenir un petit ensemble de points et qui doit bien représenter le locuteur. Essentiellement, les techniques de modélisation peuvent appartenir à :

L'approche vectorielle, l'approche statistique, l'approche connexionniste, ou bien l'approche prédictive. Une autre partition divise la modélisation en approches paramétriques et approches non paramétriques.

Les modèles paramétriques supposent que la distribution des données suit une forme connue a priori comme le modèle Gaussien. Dans le cas des modèles non paramétriques, aucune hypothèse sur la distribution des données n'est faite, comme pour la quantification vectorielle. Dans ce qui suit nous décrirons les approches de modélisation les plus répandues.

II.5.2.1. Approche vectorielle :

Dans l'approche vectorielle, un modèle de signal parole est un ensemble de vecteurs de paramètres représentatifs de l'espace acoustique construit lors de la phase de paramétrisation à partir des signaux d'apprentissage. Lors de la reconnaissance, une distance entre cet ensemble de vecteurs et les vecteurs de paramètres (MFCC, PLP, etc.) issus des signaux de test est calculée.

II.5.2.2. Approche statistique :

L'approche statistique repose sur la modélisation de la distribution des vecteurs de paramètres correspondant à un signal de parole.

- **Méthodes statistiques du second ordre :** Le principe des méthodes statistiques du second Ordre (**MSSO**) est de représenter une séquence de vecteurs acoustiques par une distribution Gaussienne multidimensionnelle. Le modèle d'un signal parole se résume alors par le triplet **{μ, ∑, X}** où **μ** est un vecteur moyen, ∑ est une matrice de covariance, qui sont tous les deux estimés à partir de la séquence de **X** vecteurs acoustiques.

- **Modèles de Markov cachés :** Les modèles de Markov cachés (ou **HMM**, Hidden Markov Models) s'appliquent parfaitement à la reconnaissance automatique de la parole. Dans cette approche, on ne s'intéresse pas à mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme acoustique ait été engendrée par le modèle. Le modèle d'un signal est constitué de l'association d'une chaîne de Markov, une succession d'états avec des probabilités de transition d'un état à l'autre, et de lois de probabilités (probabilités d'observation d'un vecteur acoustique dans un état).

- **Mélanges de Gaussiennes :** La reconnaissance de la parole par mélanges de gaussiennes (ou **GMM** pour Gaussian Mixture Models) consiste à modéliser un signal par une somme pondérée de composantes Gaussiennes. Ainsi une large gamme de distributions peut être parfaitement représentée.

  Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques. Les mélanges de gaussiennes est considéré comme un cas particulier des **HMM** et une extension de la quantification vectorielle.

  Un modèle de Gaussiennes est une somme pondérée de M densités Gaussiennes. Soit un locuteur s et un vecteur acoustique x de dimension D, le mélange de gaussiennes est défini comme suit :

$$P\ (x/\lambda_s) = \sum_{m=1}^{M} \pi_m^s\ b_m^s(x) \quad \text{.....(II.11)}$$

  Où $b_m^s(x)$ représentent des densités Gaussiennes, paramétrées par un vecteur de moyenne μ_m^set une matrice de covariance Σ_m^s :

$$b_m^S(x) = \frac{1}{(2\pi)^{D/2}.|\Sigma_m^S|^{1/2}}\, exp\left[-\frac{1}{2}(x-\mu_m^S)^1(\Sigma_m^S)^{-1}(x-\mu_m^S)\right] \text{.....(II.12)}$$

  Et les π_m^s représentent les poids du mélange, avec$\sum_{m=1}^{M} \pi_m^s$=1.

  Un locuteur est donc modélisé par un ensemble de paramètres noté λs :

$$\lambda s = (\pi_m^s, \mu_m^S\ , \Sigma_m^s) \quad \text{.....(II.13)}$$

- Un exemple de mélange de trois gaussiennes est montre dans la figure II.12.

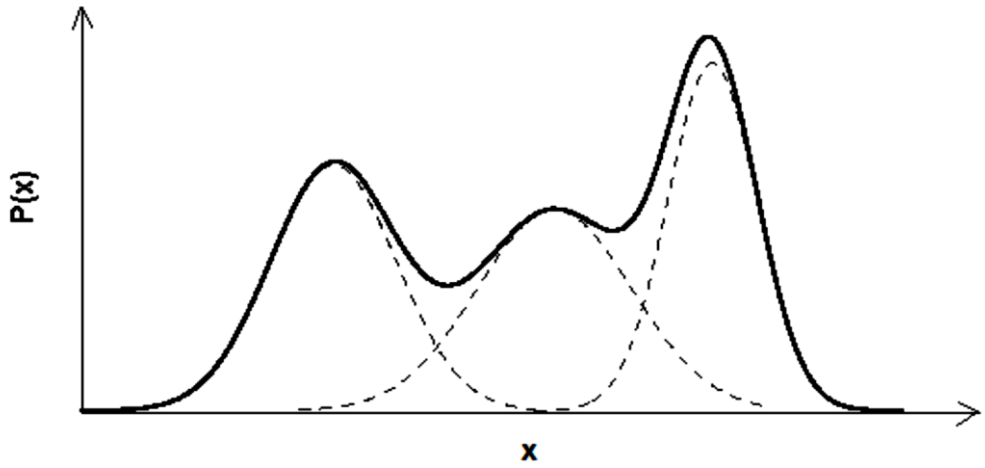

Figure II.12 : *Exemple de mélange de trois Gaussiennes.*

II.5.2.3. Approche connexionniste :

Les réseaux de neurones ont été assez largement utilisés en reconnaissance de la parole. Ces outils de classification permettent de séparer des classes, dans un espace de représentation donné, de façon non linéaire. On peut aussi utiliser les réseaux de neurones en les couplant à d'autres techniques, comme par exemple les modèles de Markov cachés. On parle alors de méthodes hybrides.

II.5.2.4. Approche prédictive :

L'approche prédictive repose sur le principe qu'une trame de signal peut être prédite par la seule observation des trames précédentes. De par ce concept, cette approche est considérée dans la littérature comme une approche dynamique. Une approche tenant compte des informations dynamiques véhiculées par le signal de parole.

Elle s'appuie principalement sur l'estimation d'une fonction de prédiction, propre à chaque signal et apprise sur les signaux d'apprentissage. Lors de la reconnaissance, une erreur de prédiction peut être calculée entre une trame prédite (par la fonction de prédiction) et la trame réellement observée dans la séquence de test. L'erreur de prédiction moyenne constitue alors la mesure de similarité entre le signal de test et le modèle (fonction de prédiction). Une autre solution envisageable est d'estimer une fonction de prédiction sur la séquence de test et de la comparer, à l'aide d'une distance, à la fonction de prédiction estimée lors de l'apprentissage [16].

II.5.3. Module de décision et mesure des performances :

On distingue deux tâches principales en reconnaissance du locuteur : la vérification du locuteur et l'identification du locuteur. Cependant, un système de RAL peut aussi servir à

identifier les segments de chaque locuteur dans un document audio, à la poursuite du locuteur ou à faire l'indexation des documents audio :

- **L'identification du locuteur** : Consiste à reconnaître un locuteur parmi un ensemble de Locuteurs en comparant son identité vocale à des références connues. Les performances du système d'identification sont données en termes de taux d'identification correcte $\mathbf{I_c}$ ou incorrecte $\mathbf{I_i}$, soit :

$$\mathbf{I_c} = \frac{Nombredetestescorrectementidentifiés}{Nombretotaledetentatives} \quad \text{(II.14)}$$

$$\mathbf{I_i} = \frac{Nombre\ de\ tests\ mal\ identifiés}{Nombre\ total\ de\ tentatives} \quad \text{(II.15)}$$

Avec : $\mathbf{I_c + I_i = 100\%}$

- **La vérification du locuteur :** consiste après que le locuteur a décliné son identité, à vérifier l'adéquation du message vocal avec la référence acoustique du locuteur qu'il prétend être. C'est une décision en tout ou rien. Les performances de vérification de locuteur sont données en termes des faux rejets FR et de fausses acceptations FA :

$$\mathbf{FR} = \frac{Nombre\ de\ tentatives\ d'abonnés\ rejetées}{Nombre\ total\ de\ \ tentatives} \quad \text{(II.16)}$$

$$\mathbf{FA} = \frac{Nombre\ de\ tentatives\ d'imposteurs\ acceptées}{Nombre\ total\ de\ tentatives\ d'imposteurs} \quad \text{(II.17)}$$

II.6. Conclusion :

Dans ce chapitre, nous avons décrit en premier lieu les caractéristiques des paramètres acoustiques de la voix dans le domaine de reconnaissance, puis expliqué le modèle source filtre de la parole et développées les différentes méthodes utilisées dans un système de reconnaissance de locuteur.

Chapitre III

Méthodologie et résultats

III.1. Introduction :

Ce chapitre présente les résultats de la reconnaissance automatique du locuteur à travers des tests effectués avec trois techniques d'extraction de paramètres (MFCC, LPCC, PLP), et sur la base de variabilité de la courte durée du signal enregistré.

Rappelons que notre travail consiste à concevoir un système de reconnaissance de locuteur et d'évaluer les performances de chaque technique. Cette évaluation permet, éventuellement, de déterminer la technique la plus adaptée dans notre cas. Plusieurs étapes sont nécessaires. L'étape d'extraction des caractéristiques est la plus importante car les performances du système en dépendent (résultats et robustesse, un temps de latence acceptable). Aussi nous évaluerons dans ce chapitre les résultats obtenus par rapport à plusieurs valeurs de courte durée des enregistrements de voix de locuteurs.

III.2. Logiciel de Simulation MATLAB :

Matlab (Matrix LABoratory) est un logiciel de calcul puissant, qui permet de réaliser des programmes structurés, à travailler sur des matrices de grande dimension, faire des calculs sur des nombres complexes et tracer des graphes en deux et trois dimensions (2D et 3D). Ce logiciel est considéré comme un langage de programmation similaire aux langages C et Pascal. En revanche, sa particularité est qu'il s'agit d'un langage interprété (c'est-`a-dire, les instructions sont exécutées immédiatement après avoir été tapées) [20].

Matlab présente comme avantage la disponibilité d'un nombreux algorithmes et fonctions de calcul numérique et traitement du signal, de plus, il est constitué d'un noyau de base extensible à l'aide de nombreuses boites à outils (TOOLBOXS). La versions de MATLAB utilisée pour la simulation du programme réalisé dans ce PFE est MATLAB R2014b/64Bits.

III.3. Présentation de la base de données :

La base de données utilisée dans ce projet est constituée d'un ensemble d'enregistrements des voix mono locuteur, provenant du Centre de Développement des Technologies Avancées [CDTA]. Chaque enregistrement est au format Wav, avec une durée supérieure à deux minutes, échantillonné à une Fréquence de 8000Hz en mono et une résolution de 16 Bits. Le nombre total des enregistrements composant la base de données voix est de 39 locuteurs. Il est à noter que deux fichiers enregistrés avec une durée de moins de deux minutes ont été exclus. Le tableau III.1contient la liste des fichiers des enregistrements de voix numérisées.

Tableau III.1 : Caractéristiques de la base de données utilisée.

N°	Nom du fichier (mono)	Nombre de Byte	Durée enregistrement (s)
1	'SP_10_Micro_Ismail_INI_OK.wav'	2160044	2Mn et 15s
2	'SP_11_Micro_Abd_INI_OK.wav'	2440046	2Mn et 32s
3	'SP_12_Micro_Hamza_RSI_ok.wav'	2160044	2Mn et 15s
4	'SP_13_Micro_Abdou_RSI_OK.wav'	2544046	2Mn et 39s
5	'SP_14_Micro_Etudiant_Abdelhak_INI.wav'	2485414	2Mn et 35s
6	'SP_15_Micro_Traich_OK_1.wav'	2279110	2Mn et 22s
7	'SP_16_Micro_IDIR.wav'	2432044	2Mn et 32s
8	'SP_17_Micro_Nouredine_Robotic_1.wav'	2224044	2Mn et 19s
9	'SP_18_Micro_PFE_Toufik_à refaire.wav'	2928044	3Mn et 3s
10	'SP_20_Micro_Fatah.wav'	2448044	2Mn et 33s
11	'SP_21_Micro_Rabah_OK.wav'	2752044	2Mn et 52s
12	'SP_22_Micro_bencherif_OK.wav'	2152046	2Mn et 14s
13	'SP_23_Micro_Yassin_AS_OK.wav'	2528044	2Mn et 38s
14	'SP_24_Micro_beggar_OK.wav'	2448044	2Mn et 33s
15	'SP_25_Micro_Mohamed Loucif_ess1.wav'	3000046	3Mn et 7s
16	'SP_26_Micro_Mihoubi.wav'	2728046	2Mn et 50s
17	'SP_27_Micro_hendaoui_OK.wav'	2262792	2Mn et 21s
18	'SP_28_Micro_Tounsi_Billal.wav'	2536046	2Mn et 38s
19	'SP_29_Micro_Oualid_OK.wav'	2776046	2Mn et 53s
20	'SP_2_Micro_tarek.wav'	2224044	2Mn et 19s
21	'SP_30_Micro_Merriche_OK.wav'	2544044	2Mn et 39s
22	'SP_31_Micro_PFE_DRIOUECHE.wav'	3072044	3Mn et 12s
23	'SP_32_Micro_Aimeur_AS_OK.wav'	2544044	2Mn et 39s
24	'SP_33_Micro_Bey_OK.wav'	2448044	2Mn et 33s

25	'SP_34_Micro_Oussalah_OK.wav'	2904046	3Mn et 1s
26	'SP_35_Micro_guenda.wav'	2432044	2Mn et 32s
27	'SP_36_Micro_Chiker_Rob_OK.wav'	2672044	2Mn et 47s
28	'SP_37_Micro_Kadri_Rob_OK.wav'	2632046	2Mn et 44s
29	'SP_38_Micro_Hamid_AS_OK1.wav'	2832044	2Mn et 57s
30	'SP_39_Micro_Benselama_OK.wav'	2424046	2Mn et 31s
31	'SP_3_Micro_Harizi.wav'	3112046	3Mn et 14s
32	'SP_40_Micro_Sofian_Lyas_OK.wav'	2840046	2Mn et 57s
33	'SP_5_Micro_Mounes.wav'	2390128	2Mn et 29s
34	'SP_6_Micro_Benghaouia_OK.wav'	2848044	2Mn et 58s
35	'SP_7_Micro_PFE_SofianUSTHB_OK.wav'	2280348	2Mn et 22s
36	'SP_8_Micro_PFE_Kamel.wav'	2896044	3Mn et 1s
37	'SP_9_Micro_Etudiant_Djamel_INI.wav'	3256046	3Mn et 23s

III.4. Protocole d'implémentation du programme sur Matlab :

Le schéma synoptique simplifié du programme de reconnaissance du locuteur a été implémenté sur Matlab, sa structure de base est donnée en **figure III.1**. Le système RAL se décompose en deux phases distinctes. La première phase est nécessaire à la construction des références ou modèles de chaque locuteur, les modèles d'apprentissage des locuteurs sont calculés et sauvegardés pendant cette phase d'apprentissage. La deuxième phase est celle de la reconnaissance réalisée pour tester les locuteurs.

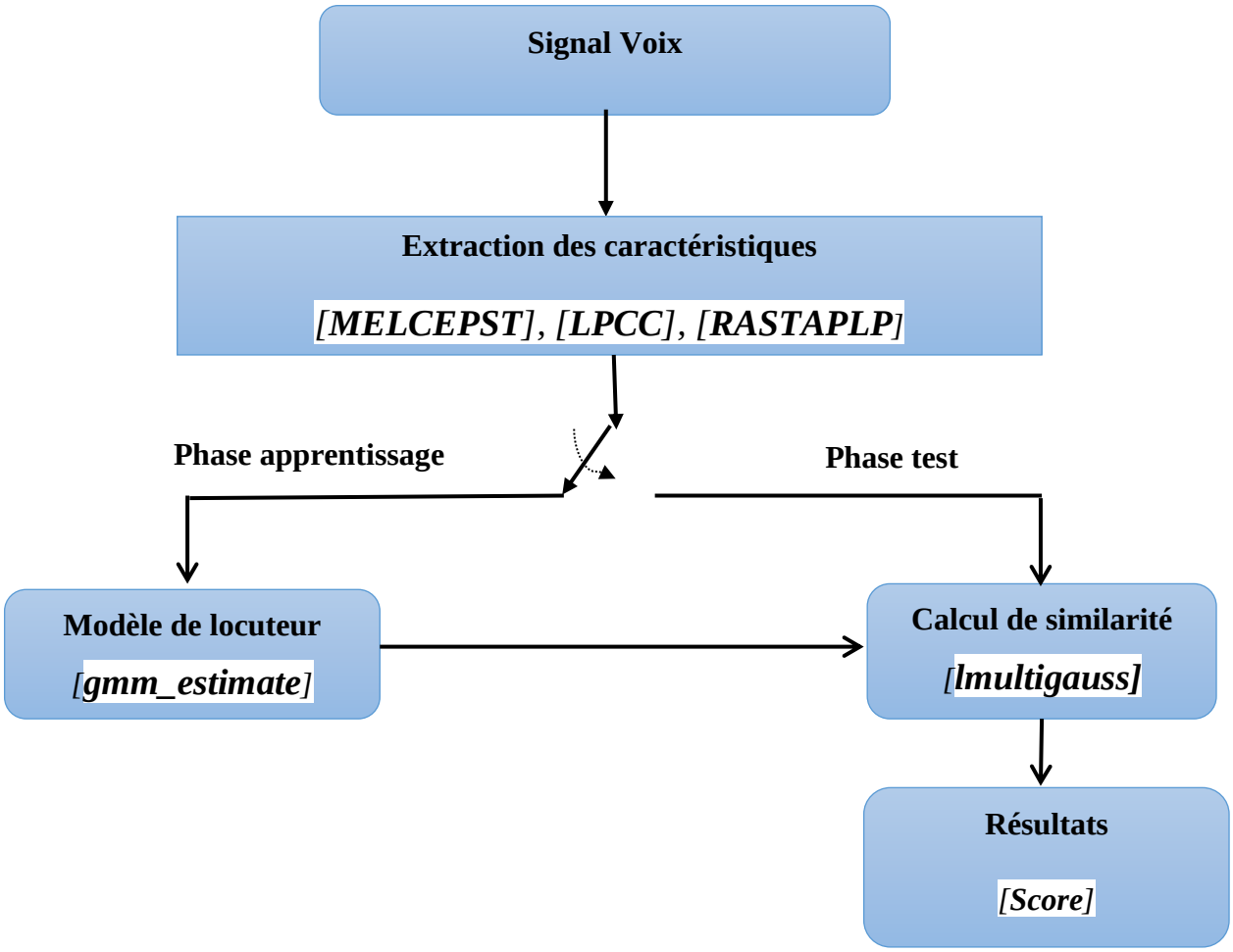

Figure III.1 : *Schéma synoptique du système d'identification du locuteur avec les principales.*

III.4.1. Extraction des caractéristiques acoustiques :

Les signaux en format Wav enregistrés dans la base de données CDTA sont lus par le programme et compte tenu que notre objet principal de PFE est aussi d'étudier la variabilité des enregistrements de locuteurs pour des courtes durées. Nous avons extrait des morceaux d'enregistrements voix pour 6 valeurs de différentes durées, variant de 20s, 15s, 10s, 6s, 3s, et 1s. Cela dans le but d'étudier l'influence de ce facteur sur le système de reconnaissance de locuteur pour les techniques d'extraction utilisées qui sont : MFCC, LPCC, et PLP.

L'extraction des caractéristiques acoustiques spécifiques de chaque locuteur passe par des étapes standard. La division du signal acoustique en segments de durée variant entre 10 et 30 ms [14]. Dans cet intervalle le signal de parole est considéré comme étant un signal quasi stationnaire. Dans notre cas, une durée de 20 ms est choisie pour segmenter les signaux des enregistrements de locuteurs de la base de données. Après segmentation en trames, nous multiplions chaque trame par une fonction de fenêtre (typiquement la fenêtre de Hamming),

suivi du calcul de la DFT. Enfin nous obtenons les coefficients Cepstraux en fonction des techniques utilisées. Ainsi, les fonctions utilisées pour l'extraction des paramètres acoustiques dans l'implémentation du programme sont données ci-après :

- ✓ **Technique MFCC :** fonction Matlab utilisée [**MELCEPST**], nombre de paramètres est de l'ordre de12.
- ✓ **Technique LPCC :** fonction Matlab utilisée [**LPCC**] nombre de paramètres est de l'ordre de12.
- ✓ **Technique PLP :** fonction Matlab utilisée [**RASTAPLP**] nombre de paramètres est de l'ordre de12.

L'ensemble des Fonctions utilisées appartiennent aux toolkits Matlab [21].

III.4.2. Apprentissage et Test :

Après avoir extrait les paramètres acoustiques, le programme entame successivement les deux phases d'apprentissage et de test. Dans la première phase, il s'agit de construire une base de données englobant les modèles des locuteurs. Suivi par la phase de test dans laquelle on compare les locuteurs aux modèles créés dans la première phase. Pour une évaluation rigoureuse, les données servant à l'apprentissage ne doivent pas être reprises lors du test.

La modélisation de locuteur est faite par la technique de modélisation par mélange de gaussiennes (GMM), celle-ci est actuellement l'une des techniques les plus fréquemment utilisées dans les systèmes de reconnaissance automatique de locuteur [14]. Comme dans le cas de MFCC, LPCC et PLP, la voix du locuteur est modélisée avec des coefficients cepstraux extraits du signal parole, ainsi, les vecteurs de caractéristiques déterminés sont utilisés pour l'apprentissage du modèle GMM. Ce processus est réalisé par l'algorithme EM (Expectation-Maximisation).

La classification des locuteurs est fournie avec le calcul de la vraisemblance conditionnelle [22]. Notre réalisation de la reconnaissance de locuteur basée sur le GMM pour la modélisation statistique par la fonction [**gmm_estimate**] qui a été paramétrée pour 16 Gaussiennes, cette fonction permet de déterminer les moyennes, les covariances et les poids des modèles créés pendant la phase d'apprentissage. La fonction [**lmultigauss**] calcule le log-vraisemblance pendant la phase de test et qui représente la similarité maximale entre les modèles et l'enregistrement de locuteur.

Pour ce faire, le programme principal qu'on a établi charge la première moitié du signal parole de la base de données CDTA, qui sera dédiée au calcul de la phase d'apprentissage, ainsi les 37 fichiers de la base de données sont utilisés pour entraîner le système et créer une base de données de modèles de locuteurs. En ce qui concerne la phase de test, les mêmes fichiers de la base de données sont utilisés, mais avec la deuxième moitié des échantillons de chaque signal et cela pour écarter la partie des échantillons déjà utilisés lors de la phase apprentissage. Ainsi, chaque locuteur parmi les 37 fichiers de la base d'origine est modélisé et sauvegardé dans une base de données à part. Ensuite pour chaque identification de locuteur, les 37 fichiers de la base de donnés CDTA sont comparés à chaque modèle sauvegardé, les résultats obtenus sont sauvegardés sur une matrice de l'ordre 37x37.Durant la tâche de reconnaissance, chaque locuteur li est représenté par un modèle GMM λi. En identification, l'objective est de trouver le modèle qui possède la probabilité a posteriori maximale ayant la séquence de parole test X. Ceci est formulé comme :

$$L = \arg \max_{1 \le i \le l} \log P(\frac{X}{\lambda i}) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \text{(III.1)}$$

Pour$L$ locuteurs.

Des résultats ont été obtenus pour différentes techniques d'extraction de paramètres acoustiques en dépendance à la variabilité de la durée d'enregistrements des locuteurs.

III.5. Résultats obtenus et discussion :

Les résultats obtenus pour chaque technique utilisée MFCC, PLP et LPCC sont discutés en matière de temps d'exécution des simulations et les performances obtenues dans les paragraphes qui suivent.

III.5.1. Comparaison des temps exécutions pour les techniques MFCC, PLP et LPCC :

Nous avons comparé le temps d'exécution de notre programme d'identification de locuteur par rapport aux techniques utilisées MFCC, PLP et LPCC. Les résultats obtenus nous ont permis de juger que les techniques PLP et LPCC sont de coût largement supérieur par rapport à celui de la technique d'extraction des paramètres acoustique MFCC, les résultats de la comparaison sont donnés dans le tableau III.2.

***Tableau III.2** : Durées d'exécutions des simulations du système RAL.*

Technique utilisée	Temps d'exécution	Observation	Configuration PC utilisée
PLP	12h, 02mn et 20s	04 fois /MFCC	- Intel cor™ i5 CPU@2,40GHz - RAM 4.00GO - Windows 7 / 64bits
LPCC	12h,49mn et 40s	04 fois /MFCC	
MFCC	2h, 59mn et 56s		

III.5.2. Performances enregistrées pour les Technique MFCC, PLP et LPCC :

Les résultats obtenus pour les trois techniques d'extraction des paramètres acoustiques sont donnés comme suit :

- **Technique MFCC :**

Dans cette expérience, nous avons étudié l'influence de la variation des durées des enregistrements sur les performances du système d'identification de locuteur par rapport à la technique d'extraction des paramètres. Le tableau III.3 fait ressortir les performances du système RAL obtenues en utilisant la technique MFCC.

***Tableau III.3** : Performance de MFCC avec différentes valeurs de courtes durées.*

		Durée d'apprentissage							
		20s	15s	10s	6s	3s	1s	Moyenne%	écart type
Durée Test	20s	56,76	55,86	39,94	41,44	19,22	5,41	36,44	20,43
	15s	62,16	56,76	40,84	41,74	20,42	5,41	37,89	21,58
	10s	59,76	57,36	40,04	39,34	21,32	7,21	37,50	20,39
	6s	58,56	55,26	38,84	39,44	21,62	5,41	36,52	20,20
	3s	54,95	53,45	37,74	37,94	21,92	5,41	35,24	18,98
	1s	43,24	48,65	34,23	28,23	17,42	7,21	29,83	15,63
	Moyenne %	55,91	54,55	38,61	38,02	20,32	6,01	35,57	19,54
	Ecart type	6,68	3,19	2,40	5,00	1,73	0,93	2,96	2,08

- **Technique PLP :**

Les résultats obtenus de cette technique, sont donnés dans le tableau III.4 :

Tableau III.4 *: Performance de PLP avec différentes valeurs de courtes durées.*

		Durée d'apprentissage							
		20s	15s	10s	6s	3s	1s	Moyenne%	écart type
Durée Test	20s	45,95	53,15	45,35	30,93	14,11	6,31	32,63	18,97
	15s	48,65	56,76	45,65	31,83	14,41	7,21	34,08	19,87
	10s	47,15	53,75	45,05	32,03	15,52	9,91	33,90	17,95
	6s	44,74	50,75	42,34	29,63	14,61	5,41	31,25	18,07
	3s	39,94	49,55	40,74	29,83	14,51	5,41	30,00	16,98
	1s	36,04	32,43	28,83	21,02	11,71	6,31	22,72	11,85
	Moyenne %	43,74	49,40	41,32	29,21	14,15	6,76	30,76	17,28
	Ecart type	4,80	8,68	6,42	4,13	1,28	1,69	4,24	2,83

- **Technique LPCC :**

Les résultats obtenus de cette technique, sont donnés dans le tableau III.5 :

Tableau III.5 : *Performance de LPCC avec différentes valeurs de courtes durées.*

		Durée d'apprentissage							
		20s	15s	10s	6s	3s	1s	Moyenne%	écart type
Durée Test	20s	37,84	28,83	26,73	20,12	11,41	4,50	21,57	12,17
	15s	36,04	26,13	26,13	21,62	11,11	6,31	21,22	10,89
	10s	36,04	31,23	27,43	21,12	10,41	5,41	21,94	12,02
	6s	33,33	27,33	26,73	21,12	10,11	3,60	20,37	11,35
	3s	30,63	26,73	26,83	21,42	10,61	5,41	20,27	10,08
	1s	24,32	27,93	22,52	21,32	8,71	4,50	18,22	9,36
	Moyenne %	33,03	28,03	26,06	21,12	10,39	4,95	20,60	10,98
	Ecart type	4,96	1,83	1,78	0,53	0,95	0,94	1,34	1,10

III.5.3. Discussion des résultats :

Les tableaux précédentsmontrent l'influence des variations des durées d'enregistrements des signaux de parole utilisées pour la phase apprentissage et qui sont segmentés comme suit :

- [20s à 10s] : les performances de l'identification du locuteur sont proches et les valeurs des moyennes se situent entre [**55.91%** à **38.61%**] pour le MFCC, [**43.74%** à **41,32%**] pour le PLP et [**33.03%** à **26.06%**] pour le LPCC.
- [06s à 3s] : les performances de l'identification du locuteur se dégradent et atteint le seuil des moyennes qui se situe entre [**38.02%** e et **20.32%**] pour le MFCC, [**29.21%** à **14.15%**] pour le PLP et [**21.12%** à **10.39%**] pour le LPCC.
- [03s à 1s] : les performances se dégradent fortement pour les trois techniques.

Par contre la variation des durées d'enregistrements des signaux de parole utilisées pour la phase de test sont proches pour les durées entre [20s à 3s] pour les trois techniques, par exemple en MFCC : entre 20s et 3s est [**36.34%** à **35,24%**].

Les représentations par histogrammes des performances du système d'identification sont données, par l'Histogramme III.2 et l'Histogramme III.3 respectivement en variant les durées pour la phase apprentissage et phase test.

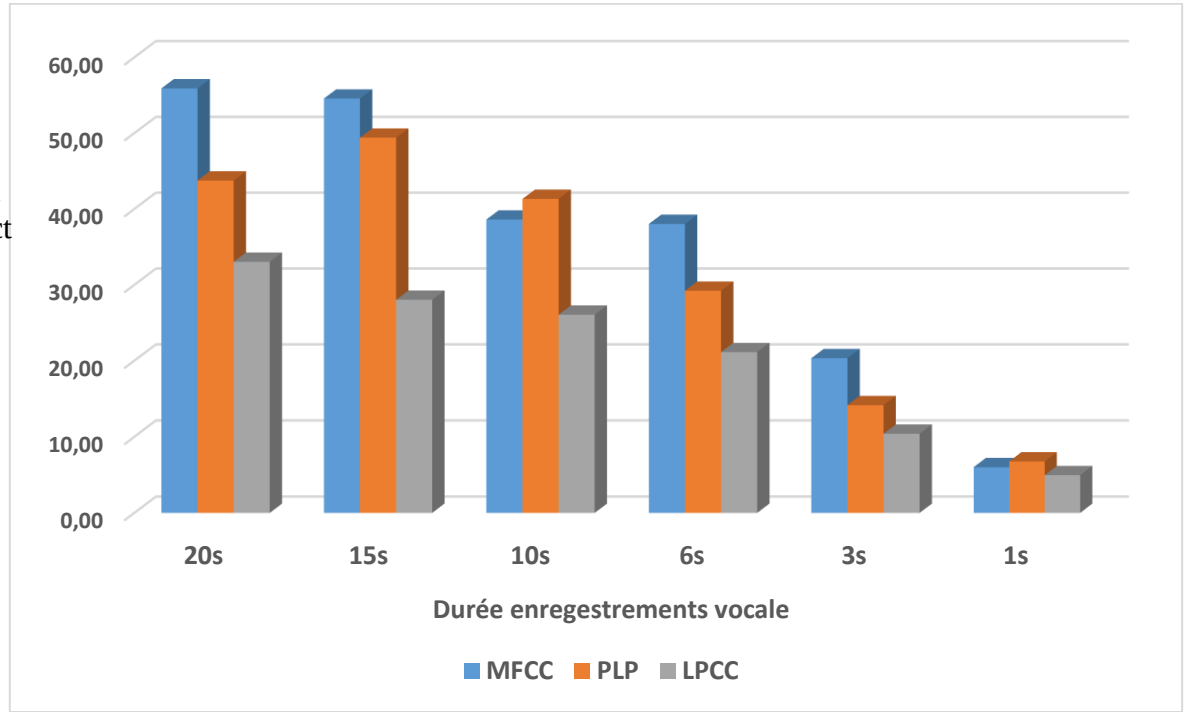

Figure III.2 : *Performance du RAL en variant la durée du signal vocal pour la phase d'apprentissage.*

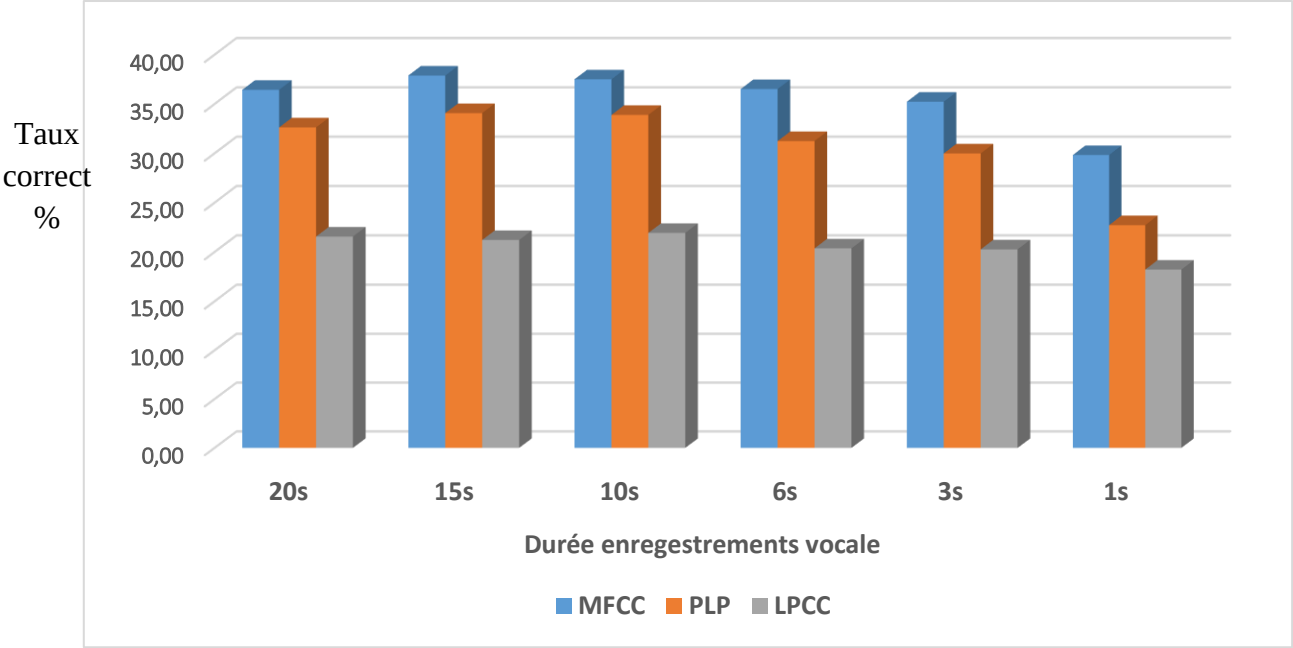

Figure III.3 : *Performance du RAL en variant la durée du signal vocal pour la phase TEST.*

Les performances du système RAL diminuent en fonction de la diminution des durées de la phase apprentissage.

III.5.4. Amélioration des résultats :

Le taux d'amélioration pour la technique MFCC par rapport aux autres techniques se calcule par la formule suivante (par rapport à la technique LPCC) :

$$\text{TAMFCC/LPCC} = \frac{Taux\ MFCC - Taux\ LPCC}{Taux\ LPCC}\%$$

De même par rapport à la technique PLP :

$$\text{TAMFCC/PLP} = \frac{Taux\ MFCC - Taux\ PLP}{Taux\ PLP}\%$$

Les résultats de calcul des améliorations des taux de reconnaissance par rapport aux techniques LPCC et PLP sont donnés dans le tableau III.6.

Tableau III.6 *: Taux d'amélioration par rapport aux techniques LPCC et PLP.*

Technique	20s	15s	10s	6s	3s	1s
LPCC	**33,03**	**28,03**	**26,06**	**21,12**	**10,39**	**4,95**
PLP	**43,74**	**49,4**	**41,32**	**29,21**	**14,15**	**6,76**
TAMFCC/PLP	**28%**	**10%**	**-7%**	**30%**	**44%**	**-11%**
TAMFCC /LPCC	**69%**	**95%**	**48%**	**80%**	**96%**	**21%**

Les taux d'amélioration obtenus montrent que les performances de la technique MFCC sont supérieures que ceux des deux autres techniques PLP et LPCC.

III.6. Conclusion :

Dans ce chapitre, nous avons mis en simulation les différentes étapes du système de reconnaissance du locuteur. Les paramètres caractéristiques du locuteur sont calculés à partir d'échantillons vocaux obtenus à partir de la base de données CDTA. Ces paramètres sont comparés à chaque modèle construit lors de la simulation du programme et stocké dans une base de données. L'enregistrement produisant la similarité maximale est attribué comme identité de locuteur.

Nous avons étudié les différentes techniques d'extractions de paramètres pour une variabilité de temps de courte durée afin de et déterminer les performances et mettre en avant la technique la plus appropriée parmi les trois techniques d'extraction des paramètre acoustiques analysées. Nous avons prouvé dans notre travail de reconnaissance de locuteur dépendant d'enregistrements à courte durée, que la technique MFCC dépasse largement les techniques LPCC et PLP en performance de reconnaissance et en terme de temps d'exécution de son algorithme.

Conclusion générale

Le travail présenté dans ce projet de fin d'étude s'inscrit dans un cadre de reconnaissance automatique de locuteur à partir de la voix. Il est axé sur l'étude de la variabilité du signal parole en terme de courte durée et pour différentes techniques d'extraction de paramètres acoustiques.

Les techniques de modélisation et d'extraction des paramètres sont les parties principales d'un système de reconnaissance du locuteur. Les deux premiers chapitres ont été consacrés à la théorie de la biométrie générale et particulière au système vocal. Après avoir étudié les différents composants d'un système RAL. Nous avons exploré le mode de fonctionnement de trois techniques d'extraction des paramètres (MFCC, LPCC, et PLP) et en implémentant un programme RAL sous Matlab, nous avons comparé ces trois techniques pour déterminer les meilleurs paramètres qui peuvent caractériser une voix. Les résultats de simulation dans un contexte de courte durée, nous a permis de classer la technique MFCC par rapport au technique LPCC et PLP comme étant la plus performante. Cette performance ressort en terme de rapidité d'exécution de son algorithme et son taux de reconnaissance qui atteint une amélioration de 96% par rapport à la technique LPCC et 44% pour la technique PLP.

Comme principale perspective dans le domaine de la biométrie vocale : la première observation que nous pouvons faire est la construction d'une base de donnée de locuteurs propre à l'université de BOUIRA. Afin d'améliorer les performances de l'identification de locuteur présentées dans ce travail, le développement d'autres techniques de rehaussement notamment l'utilisation de prétraitement est une solution pouvant rendre le système RAL plus robuste pour l'indentification de locuteur en terme de courte durée.

Références bibliothèques

[1] https://www.biometrie-online.net/technologies/voix/voix-applications: consulté le 11/10/2020.

[2] S. BOUDJELAL, « Détection et identification de personne par méthode biométrique » Mémoire de magister en électronique, Université Mouloud Mammeri de Tizi Ouzou, 2014.

[3] S. AKROUF, « Une Approche Multimodale pour l'Identification du Locuteur », thèse de Doctorat, Université Ferhat Abbas- Sétif, 2011.

[4] I. BENCHENNANE, « Etude et mise au point d'un procédé biométrique multimodale pour la reconnaissance des individus », thèse de Doctorat, Université Mohamed Boudiaf, d'Oran, 2016.

[5] https://fr.wikipedia.org : consulté le 20/01/2021.

[6] N. K. KOUMADI, « authentification automatique du propriétaire d'un téléphone mobile », Mémoire pour l'obtention du grand de maitre en informatique, Université Québec, Canada 2018.

[7] J. M. GAUTHIER, « Cadre juridique de l'utilisation de la biométrie au Québec : Sécurité et vie privée », Mémoire pour l'obtention du grande de maitrise, Université de Montréal. Avril 2014.

[8] K. AIZI, « Identification Biométrique Multimodale à Distance », Thèse de Doctorat, Université Mohamed Boudiaf, d'Oran ,29/06/2016

[9] F. DEBBECHE-GUERID, « Conception d'un système acoustique-anatomique pour l'identification du locuteur », Diplôme de magistère, Université Badji Mokhtar de Annaba, 2008.

[10] S. LAMECH, « Mécanismes laryngés et voyelles en voix chantée dynamique vocale, phonétogrammes de paramètres glottiques et spectraux, transitions de mécanismes », Thèse de Doctorat, Université Pierre et Marie Curie,2010.

[11] https://infos-diabete.com/anatomie-larynx/:consulté le 19/10/2020.

[12] M. CHAA. « Système de la reconnaissance de personnes par des techniques biométriques », thèse de Doctorat, Université Sétif-1,2017.

[13] W. BEN KHEDAR, « Reconnaissance du locuteur en milieux difficiles, informatique et langage [cs.CL] », Thèse de Doctorat, Université d'Avignon, 2017.

[14] A. ALIMOHAD, « Contribution à l'inférence d'identité en utilisant un Système de reconnaissance du locuteur GMM-UBM », Thèse de Doctorat, Université Blida-1, septembre 2015.

[15] CH. HADRI, « la reconnaissance des paramètres de la trace acoustique et son application dans la reconnaissance de la parole », Diplôme de Magister, Université Annaba année 2007/2008.

[16] A PODDAR, M SAHIDULLAH, G SAHA, « Sepeaker verification with Short Ultranences », Article IET Biometrics, 2015.

[17] Z. YOUSSEF, O. KAIS, "A bio-inspired feature extraction for robust speech recognition», Springer Plus, Springer Open Journal, 2014.

[18] T. GULZAR, A. SINGH AND S. SHARMA, « Comparative analysis of LPCC, MFCC and BFCC for the recognition of hindi words using artificial neural networks », International Journal of Computer Applications (0975 – 8887) Volume 101– No.12, Septembre 2014.

[19] N. YALA, « Reconnaissance automatique du locuteur », Mémoire de magister en électronique, Université Houari Boumediene, Alger 20/12/2010.

[20]https://fr.mathworks.com/ : consulté le 5/11/2010.

[21]https://www.mathworks.com/matlabcentral/fileexchange/27059-speaker-recognition-system: consulté le 09/10/2020.

[22] T. MARCINIAK, R. WEYCHAN, S. DRGAS, A. DĄBROWSKI AND A. KRZYKOWSKA, « Speaker recognition based on short polish sequences », Signal Processing Algorithms, Architectures, Arrangements, and Applications SPA 2010, Poznan, 2010, pp. 95-98.