REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET

DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE BLIDA 1

Institut d'Aéronautique et des études spatiales

PROJET DE FIN D'ETUDES

En vue de l'obtention du diplôme de master en Aéronautique

Spécialité: Navigation Aérienne

Option : CNS/ATM

Simulation de la Méthode de Mel pour la Reconnaissance Automatique d'un locuteur

Réalisé par :
Mlle Moulai Amel

Dirigé par :
Mme Azine Houria
Mr Djabri Mounir

PROMOTION 2018

Remerciements

En préambule à ce mémoire je remercie le Bon Dieu qui m'a aidé et m'a donné la patience et le courage durant ces longues années d'étude.
Je souhaite adresser mes remerciements les plus sincères à mes parents ainsi qu'à mes frangins et frangines de m'avoir encouragée, supportée, épaulée et avoir cru en moi tout au long de ces années. Sans eux, je ne serai pas là.

Mes plus grands remerciements vont tout d'abord à **Mme Azine Houria** qui a accepté de diriger ce travail de Master, ainsi que pour sa disponibilité et pour ses précieux conseils et précieuses orientations.

Je remercie sincèrement l'équipe de la **« DTNA »** à leur tête **Mr Djabri Mounir**, pour leur accueil, leur disponibilité, collaboration et amabilité qui m'a permis d'avoir toutes les informations et données nécessaires pour réussir ce travail.

Mes vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils porteront à mon travail en acceptant d'examiner ce mémoire et de l'enrichir par leurs propositions.

Je remercie enfin tous ceux qui m'ont apporté leur aide et qui ont contribué à l'élaboration de ce mémoire ainsi qu'à la réussite de cette formidable année universitaire et qui n'ont pas pu être cités ici.

Merci à tous.

Résumé

Ce mémoire s'inscrit dans le domaine de la Reconnaissance Automatique du Locuteur, un domaine riche d'applications potentielles allant de la sécurisation d'accès à l'indexation de documents audio, dont l'objectif est de reconnaître une personne par l'analyse de sa voix. Cette authentification peut être réalisée au moyen d'une application intégrée dans les systèmes d'enregistrement des communications **ATC/ATM** dans le but d'atteindre un niveau de sécurité acceptable.

Pour mettre en œuvre un tel système il faut passer par plusieurs étapes. On coupe le signal vocal en trames avec chevauchement. Le résultat obtenu est une matrice, où chaque colonne est une trame de N échantillons du signal de la parole originale. En appliquant ensuite le fenêtrage et la **FFT** pour transformer le signal dans le domaine fréquentiel, et enfin la dernière étape, qui est la conversion du spectre de puissance en coefficients cepstraux de la fréquence Mel **(MFCC)**. Nous appliquerons la technique de reconnaissance de formes basée sur la quantification vectorielle **QV** pour construire des modèles de référence du locuteur.

Abstract

This thesis is part of the field of Automatic Speaker Recognition, a rich domain of potential applications ranging from securing access to the indexing of audio documents, the purpose of which is to recognize a person through analysis in his voice. This authentication can be achieved by means of an application integrated into the **ATC/ATM** communication recording systems in order to achieve an acceptable level of security.

To implement such system, it is necessary to go through several steps. The voice signal is cut into overlapping frames. The result is a matrix where each column is a frame of N samples of the original speech signal. Then applying windowing and **FFT** to transform the signal into the frequency domain, the final stage of speech processing is the conversion of the power spectrum into cepstral coefficients of the Mel frequency **(MFCC)**. We will apply the **QV** vector quantization based pattern recognition technique to build speaker reference models.

ملخص

هذه الأطروحة هي جزء من مجال التعرف التلقائي على المتحدث ، وهو مجال غني من التطبيقات المحتملة التي تسعى الى تأمين الوصول إلى فهرسة المستندات الصوتية ، والغرض منها هو التعرف على شخص من خلال التحليل في صوته. ويمكن تحقيق هذه المصادقة من خلال تطبيق مدمج في أنظمة تسجيل الاتصالات **ATC/ATM** من أجل تحقيق مستوى مقبول من الأمن.

لتنفيذ مثل هذا النظام ، من الضروري المرور بعدة خطوات. يتم قطع الإشارة الصوتية إلى إطارات متداخلة والنتيجة هي مصفوفة حيث يكون كل عمود عبارة عن إطار من عينات N لإشارة الكلام الأصلية. ثم تطبيق النوافذ و **FFT** لتحويل الإشارة إلى مجال التردد ؛ المرحلة النهائية لمعالجة الكلام هي تحويل طيف القدرة إلى معاملات (MFCC). كما سنقوم بتطبيق تقنية التعرف على الأنماط القائمة على تكوير ناقلات **QV** لبناء نماذج مرجعية للسماعات.

Abréviations/Acronymes :

OACI : Organisation de l'aviation civile internationale.

ENNA : Établissement national de la navigation aérienne.

EPIC : Établissement Public à Caractère Industriel et Commercial.

OGSA : Organisation de Gestion et de Sécurité Aéronautique.

ONAM : Office de la Navigation Aérienne et de la Météorologie.

ENEMA : Etablissement National pour l'exploitation Météorologique et Aéronautique.

ENESA : Entreprise Nationale de l'Exploitation et de la Sécurité Aéronautique.

DDNA : Direction du Développement de la Navigation Aérienne.

DENA : Direction de l'Exploitation de la Navigation Aérienne.

DTNA : Direction Technique de la Navigation Aérienne.

DRFC : Direction des Ressources, des Finances et de la Comptabilité.

DJRH : Direction Juridique et des Ressources Humaines.

CQRENA : Centre de Qualification, de Recyclage et d'Expérimentation de la Navigation Aérienne.

DL : Direction de la Logistique.

CCV : Centre de Calibration en Vol.

DSA : Directions de Sécurité Aéronautique.

DAF : Direction Administrative et Financière.

DEB : Département d'Energie et de Balise.

DSSLI : Département des Services de Sauvetage et de Lutte Contre l'Incendie.

PDGEA : Projet de Développement de la Gestion de l'Espace Aérien.

ILS : Instrument landing system, système d'atterrissage aux instrument.

RADAR : RAdio Detection and Ranging.

TCAS : Traffic Collision Avoidance System, Système d'Alerte de Traffic et d'Evitement de Collision.

EGPWS : Enhanced Ground Proximity Warning System, Système Avertisseur de Proximité du Sol Amélioré.

FDS : Filet De Sauvegarde.

STCA : Short Term Conflict Alert, Alerte de Conflit à Court Terme.

RIMCAS : Runway Incursion Monitoring and Collision Avoidance System, Système de Surveillance des Incursions Sur Piste et d'Evitement des Collisions.

APW : Area Proximity Warning, Avertissement de Proximité de Zone.

CCR : Centre de Contrôle Régional.

APP : Centre de Contrôle d'Approche.

TWR : Tour de Contrôle d'Aérodrome.

FIR : Flight Information Region, Région d'Information de Vol.

V-SAT: Very Small Aperture Terminal, Terminal à Très Petite Ouverture.

HF: High Frequency, Haute Fréquence.

VHF: Very High Frequency, Très Haute Fréquence.

AM : Amplitude Modulation, Modulation d'Amplitude.

NM : Nautical Mile, Mile Nautique.

VCCS : Voice Communication Control System, Système de Contrôle de Communication Vocale.

ATC: Air Traffic Control, le Contrôle du Traffic Aérien.

USB : Universal Serial Bus, Bus Universel En Série.

IP : Internet Protocol, Protocole Internet.

RAL : Reconnaissance Automatique du Locuteur.

IAL: Identification Automatique du Locuteur.

VAL: Vérification Automatique du Locuteur.

DTW: Dynamic Time Warping, Alignement Temporel Dynamique.

VQ: Vector Quantization, Quantification Vectorielle.

HMM : Hidden Markov Model, Modèle de Markov Caché.

GMM : Gaussian Mixture Model, Modèle de Mélange Gaussien.

MSSO : Méthodes Statistiques du Second Ordre.

RMP : Regression-Based Model Prediction.

RSW : Reference Speaker Weighting, Réfèrence de Ponderation du Locuteur.

RII: Infinite Impulse Response Filter, Filtre à Réponse Impulsionnelle Infinie**.**

RIF : Finite Impulse Response Filter, Filtre à Réponse Impulsionnelle Finie.

DCT: Discret Cosine Transform, Transformée en Cosinus Discrète.

LBG : Linde,Buzo et Gray.

LPC : Linear Prediction Coding, Codage de Prédiction Linéaire.

MFCC : Mel Frequency Cepstral Coefficient, Coefficient Cepstraux de la Fréquence Mel

FFT : Fast Fourier Transform, Transformée de Fourier Rapide.

DFT : Discret Fourier Transform , Transformée de Fourier Discrète.

WFT : Windowed Fourier Transform, Transformée de Fourier Fenêtrée.

STFT : Short Term Fourier Transform, Transformée de Fourier à Court Terme.

SSLI : Service de Sauvetage et de Lutte Contre l'Incendie.

VSAT : Very Small Aperture Terminal, Très Petite Borne d'Ouverture.

Tables Des Matières

Liste des figures :

Liste des tableaux :

Introduction générale

La parole est depuis tout temps le moyen de communication privilégié de l'Homme. Elle véhicule, en plus du message linguistique prononcé, plusieurs types d'informations.

Ces informations servent en particulier à déterminer l'identité du Locuteur ; elles sont exploitées par les humains pour l'identification des personnes qu'ils connaissent en particulier à distance.

Le segment **ATM / ATC** exige que les systèmes d'enregistrement offrent une stabilité et une sécurité maximales avec un niveau de qualité de service suffisant. On peut dire maintenant que l'utilisation de la biométrie vocale dans les systèmes d'enregistrement sera donc nécessaire parce qu'elle fournit une plus grande sécurité et gain de temps lors d'une authentification des locuteurs et une vérification indépendante du contexte de l'appel enregistré ainsi la détection d'une fraude possible.

Les systèmes d'enregistrement sont connectés avec les systèmes de communication vocale et utilisés pour les communications radio et téléphoniques dans les centres de contrôle de la circulation aérienne. L'enregistreur est spécifiquement conçu pour enregistrer à partir de différentes technologies et canaux de communication (voix, données, écrans et vidéo) avec intégration dans un seul appareil.

Un système de reconnaissance automatique du locuteur est l'analyse acoustique et le traitement du signal, qui transforment le signal parole en une séquence de vecteurs acoustiques, Cette représentation doit être adaptée pour la reconnaissance, on conserve dans les vecteurs acoustiques l'information lexicale et on supprime toutes autres informations, telles que la variabilité intra et interlocuteur, les bruits ambiants etc.

La représentation utilisée généralement en reconnaissance est basée sur des coefficients cepstraux **(LPC, MFCC, PLP)**. Ces coefficients soient utilisés en raison de leurs propriétés de représentation, notamment la décorrélation des coefficients.

Dans le premier chapitre, nous commencerons par un rapport sur **l'ENNA**, plus précisément la **DTNA**. L'établissement où je pouvais accomplir mon stage pratique. Nous décrirons aussi le système de production et de perception de la parole ainsi que les systèmes biométriques couramment utilisés.

Dans le deuxième chapitre, nous présenterons également les différentes étapes classiques utilisées pour la réalisation d'un système de reconnaissance automatique du locuteur et les approches utilisées.

Dans le troisième chapitre, nous présenterons la méthode **MFCC** et son analyse cepstrale. Une méthode qui utilise les coefficients cepstraux à l'échelle de Mel en vecteur de paramètres ainsi que leurs classifications en régions par la quantification vectorielle.

Le dernier chapitre, consiste à présenter l'approche choisie pour la reconnaissance du locuteur où les différents blocs qui le constituent seront développés. Nous présenterons ensuite les différents résultats de simulation obtenus à l'aide du logiciel MATLAB®.

Chapitre I : La Reconnaissance Automatique du Locuteur

1.1. Introduction :

Le transport aérien est un système complexe qui offre de nombreuses interactions entre les différents acteurs (compagnies aériennes, contrôle aérien, aéroports,...) et met en jeu de multiples interventions humaines dans un environnement incertain et fluctuant.
La sécurité aérienne procède de l'ensemble de mesures visant à réduire le risque aérien.

La « sécurité » aérienne ne doit pas être confondue avec la « sûreté » aérienne qui comprend l'ensemble des mesures prises pour lutter contre les malveillances intentionnelles comme les actes de terrorisme.

L'Organisation de l'aviation civile internationale édicte des normes et des recommandations applicables dans les pays signataires de la convention de Chicago.

Selon **l'OACI**, « la Sécurité est une situation dans laquelle les risques de lésion corporelle ou de dommages matériels sont limités à un niveau acceptable. Et maintenus à ce niveau par un processus continu d'identification des dangers et de gestion des risques ».
L'annexe 10 de **l'OACI** définit les normes et recommandations applicables aux radiocommunications aéronautiques.

Toutes les compagnies aériennes, qu'elles effectuent des vols réguliers ou des vols charters, sont assujetties aux règles techniques de leur état de rattachement. Ces règles doivent être conformes aux normes internationales de sécurité édictée par **l'OACI**.

1.2. Présentation de l'Etablissement National de la Navigation Aérienne :

L'établissement national de la navigation aérienne dans sa nature juridique, est un établissement public à caractère industriel et commercial, sous tutelle du ministère des transports, il est dirigé par un directeur général et administré par un conseil d'administration.

1.2.1 Historique :

Depuis l'indépendance, de 1962 à 1991, cinq organismes ont été chargé de la gestion, de l'exploitation et du développement de la navigation aérienne en Algérie : **OGSA**, **ONAM**, **ENEMA**, **ENESA**, **ENNA**.

1.2.2. Mission de l'ENNA :

- Assure le service public de la sécurité de la navigation aérienne au nom et pour le compte de l'état.
- Chargé de la mise en œuvre de la politique nationale en matière de navigation aérienne en coordination avec les autorités concernées et les institutions intéressées.
- Le contrôle de la circulation aérienne pour l'ensemble des aéronefs évoluant dans l'espace aérien algérien.
- L'acquisition, l'installation et la maintenance des moyens de surveillance, de radionavigation et de télécommunications aéronautiques ainsi que leur calibration, (au moyen de son avion laboratoire).
- La fourniture de l'énergie à l'ensemble des aérodromes.
- La concentration, diffusion ou retransmission au plan national et international des messages d'intérêts aéronautique ou météorologique.
- Le service d'alerte au profit des aéronefs évoluant dans l'espace aérienne algérien et son concours aux services des recherches et de sauvetage.
- Le service de sauvetage et de lutté contre l'incendie sur les plates-formes aéroportuaire.
- La Participation à l'élaboration et à la mise en œuvre pour ce qui le concerne:

  ✓Des plans d'urgence d'aérodromes.

  ✓Des plans de servitudes aéronautiques et radioélectriques.

  ✓Des plans et programmes des recherches et de sauvetages.

- Gère le domaine aéronautique constitué pour l'espace aérien, les terrains, bâtiments et installations nécessaires à l'accomplissement de sa mission.

1.2.3. L'Organisation de l'ENNA :

La structure générale de l'ENNA est la suivante :

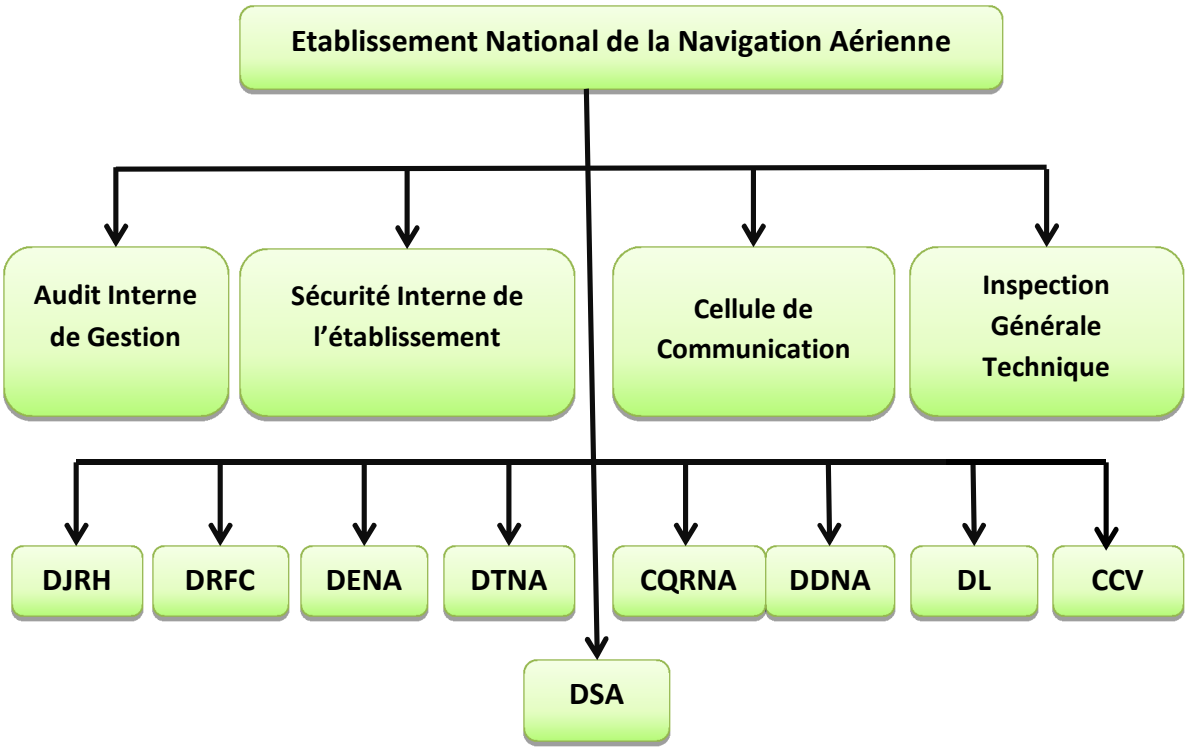


Figure 1.1 : Organigramme de l'établissement national de la navigation aérienne.

1.3. Présentation de la Direction Technique de la Navigation Aérienne :

La Direction Technique de la Navigation Aérienne est une direction à vocation technique, c'est l'une des directions les plus importantes de l'établissement national de la navigation aérienne vue les taches qui lui sont confiées.

1.3.1. Mission de la DTNA :

Parmi les tâches que doit accomplir la **DTNA** nous citons :

- L'acquisition, l'installation et la maintenance des moyens de télécommunications, surveillance aéronautiques et de radionavigation.
- La fourniture de l'énergie à l'ensemble des équipements installés au niveau des plates-formes aéroportuaires
- Chargé de l'inspection des équipements d'aides à la navigation aérienne et cela à l'aide de l'avion labo dont dispose **l'ENNA**.

1.3.2. L'Organisation de la DTNA :

La structure générale de la **DTNA** est la suivante :

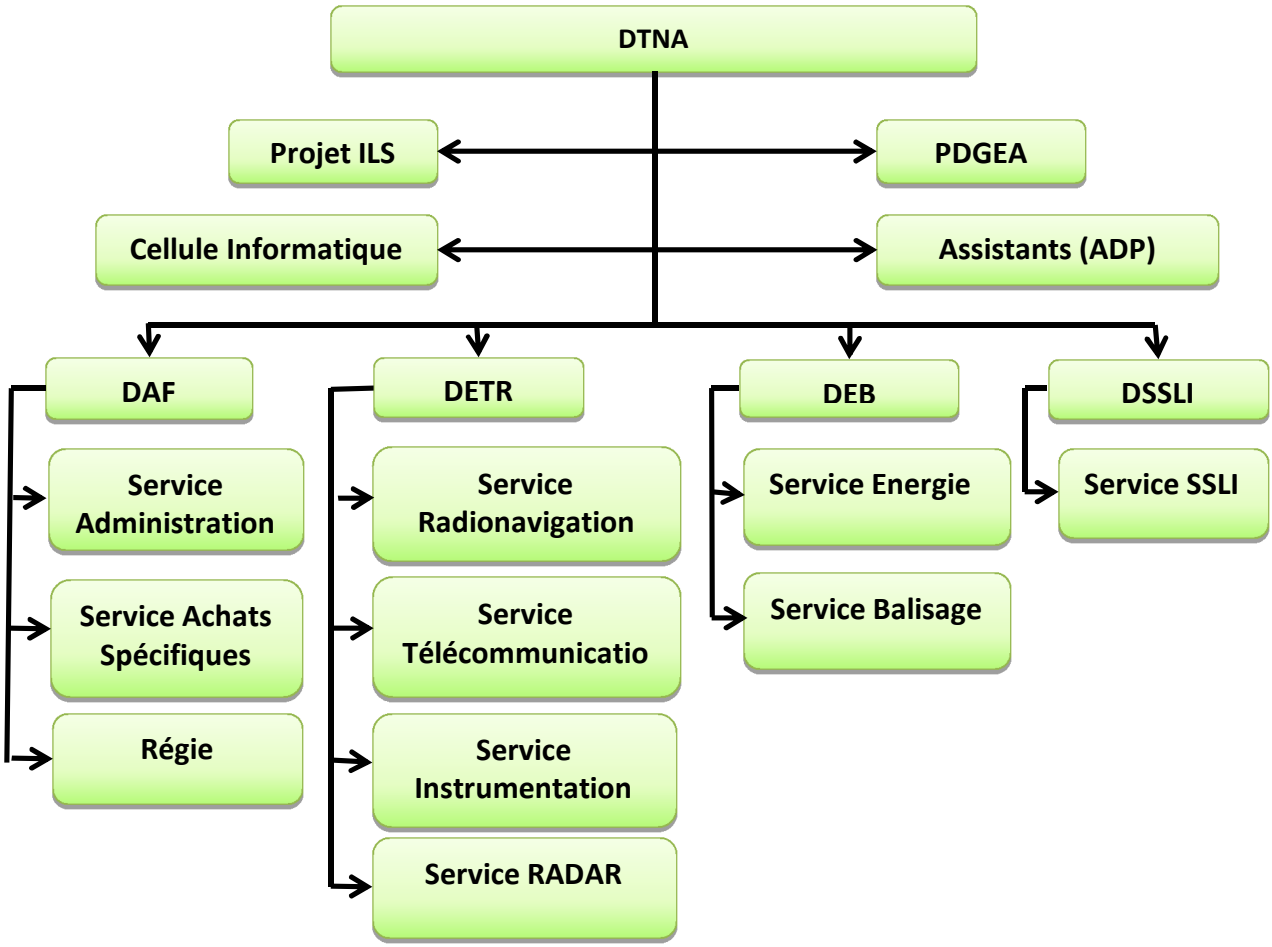


Figure 1.2 : Organigramme de la direction technique de la navigation aérienne.

Afin de se familiariser avec cette organisation, le stage a été effectué au sein de la **DETR** avec la visite des différents services.

1.3.3. Les Services du DETR :

Le schéma synoptique ci-dessous nous donne les quatre principaux services du Département des Equipements de Télécommunications et de Radionavigation, qui est chargé du suivi de l'ensemble des moyens de mesure, d'installation et de maintenance des équipements qu'il possède et qui sont nécessaires à la sécurité aérienne.

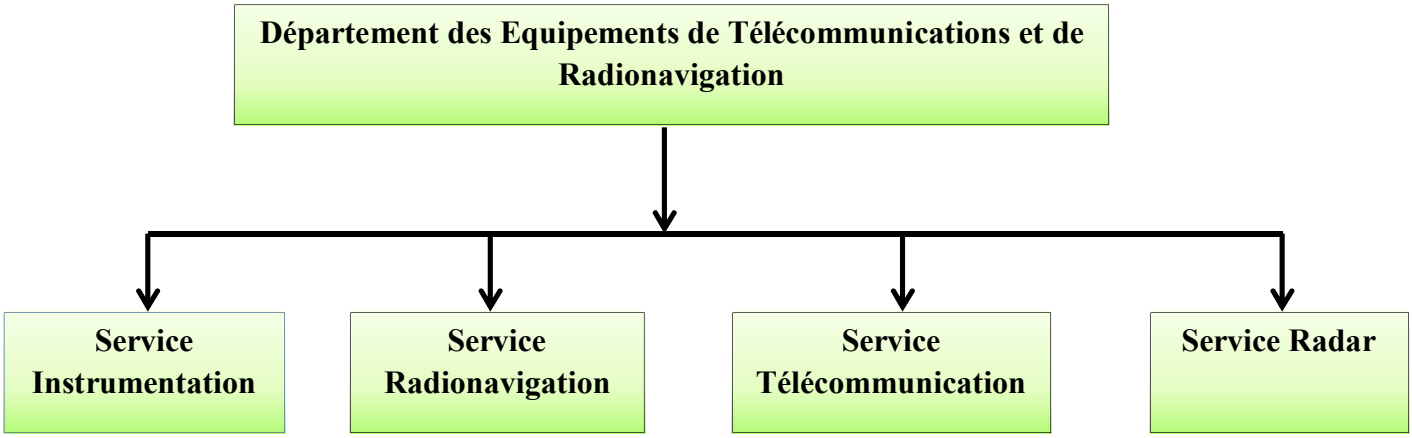


Figure 1.3 : Schéma Synoptique du Département des Equipements de Télécommunications et de Radionavigation.

1.4. Contrôle de la circulation aérienne :

La mission prioritaire du service du contrôle aérien est de prévenir les collisions entre aéronefs par un écoulement sûr du trafic aérien.

Quand la situation est dégradée ou sur le point de se dégrader, il existe des outils qui assistent les pilotes (outils embarqués) et les contrôleurs aériens (outils **ATC**) pour leur permettre de corriger la situation et de prévenir l'accident.

- Les outils embarqués :
  - Le **TCAS**
  - **L'EGPWS**
- Les outils **ATC** :
  - Le **FDS** ou **STCA** : c'est une alarme détectant le non-respect des normes de séparation entre les aéronefs.
  - l'alarme **RIMCAS** : elle vise à prévenir les incursions de piste.
- l'**APW** : alarme pour prévenir la pénétration dans un espace aérien non autorisé.

Pour assurer ces services, un organisme de contrôle est mis en place.
Suivant le type de trafic et sa position, différents organismes assurent les services de contrôle, informations et alertes :

- **Les centres de contrôle régionaux :** Ils sont chargés d'assurer les services de la circulation aérienne au bénéfice des aéronefs en croisière (en dehors de la proximité d'aérodrome).

- **Les centres de contrôle d'approche :** Ils sont chargés d'assurer les services de la circulation aérienne aux abords d'un aérodrome, dans une zone de contrôle dont la taille est variable. Les contrôleurs aériens sont situés soit dans la vigie d'une tour de contrôle, soit dans une salle radar dédiée.
- **Les tours de contrôle d'aérodrome :** Elles sont chargées d'assurer les services de la circulation aérienne dans la circulation d'aérodrome, c'est-à-dire dans une zone restreinte (de l'ordre d'une dizaine de kilomètres) autour d'un aérodrome. Le service est rendu depuis la vigie d'une tour de contrôle**.**

1.5. La radiocommunication aéronautique :

Les radiocommunications aéronautiques sont dans des bandes de fréquences du spectre radioélectrique, réservée à l'aéronautique par des traités internationaux.
Elles sont utilisées pour les communications entre les pilotes et le personnel des stations au sol.

Elle permet de transmettre des clairances et des informations importantes pour la sécurité de la circulation aérienne et l'efficacité de la gestion du trafic aérien.

1.5.1. La communication sol/sol sol/air :

La gestion du trafic aérien nécessite des moyens et des installations qui doivent répondre aux normes très exigeantes de l'aviation civile afin d'assurer la sécurité et la fluidité de la circulation aérienne.

Les communications vocales constituent un moyen à la fois puissant et indispensable à la sécurité. Deux types de communications vocales sont utilisés :

1.5.1.1. Communications Sol – Sol :

Principalement les communications téléphoniques avec les services locaux (Bureau des informations aéronautiques, **SSLI**, Météo, Technique… etc.), les services régionaux (le Centre de Contrôle Régional, ou d'Approche) et 'international (espaces aériens ou **FIR** voisins).

Des communications radios sol – sol sont également utilisées pour les opérations au sol (entre la tour de contrôle et les camions **SSLI** par exemple).

1.5.1.2. Communications Air – Sol :

Elles permettent aux aéronefs de rester tout le temps en contact avec les contrôleurs au sol (**TWR**, **APP** et **CCR**).

Ces communications doivent être précises, concises et conventionnelles, on prend l'exemple des clairances, qui sont des instructions données aux aéronefs. Les messages échangés obéissent à l'ordre de priorité suivant :

1- Messages de détresse (Mayday).
2- Messages d'urgence (Panne Panne).
3- Messages de radiogoniométrie.
4- Messages intéressant la sécurité.
5- Messages météorologiques.
6- Messages intéressant la régularité des vols.

1.5.2. Supports de communication :

Différents types de supports de communication sont utilisés :
-L'atmosphère ou l'air libre.
-Câbles téléphoniques (paires torsadées).
-Câbles coaxiaux.
-Fibres optiques.
-Faisceaux hertziens.

- **V-SAT**.

1.5.3. Les Moyens des télécommunications :

Vu l'importance de la communication dans le domaine de l'aviation, un certain nombre de moyens est utilisé aussi bien à bord des aéronefs qu'au sol. Les moyens installés au sol, se trouvent dans les tours de contrôle, dans les centres de contrôle régionaux, ou même dans des Shelters, qui sont des abris spéciaux positionnés dans certains sites géographiques.

1.5.3.1. La HF (3 – 30 MHz) :

La bande de fréquence 2.85 – 24.89 MHz est dédiée à l'aviation civile. Ce moyen est surtout utilisé dans les régions désertiques non couverte par la **VHF**, et aussi comme moyen d secours. La **HF** assure des communications vocales directes entre le centre de contrôle et les aéronefs à des distances lointaines, pouvant atteindre des milliers de kilomètres. En revanche, la qualité de la communication est très affectée par les perturbations atmosphériques.

Actuellement, on utilise des stations numériques qui permettent un traitement plus sophistiqué au signal, afin d'améliorer la qualité de la communication

1.5.3.2. La VHF (30 – 300 MHz) :

La gamme de fréquence 118 – 137 MHz est dédiée à l'aviation civile. On utilise la modulation d'amplitude pour permettre un espacement entre les canaux allant jusqu'à 8.33 kHz.

Contrairement à la **HF**, la **VHF** permet d'avoir une qualité très nette, grâce au fait qu'elle ne soit pas affectée par l'atmosphère. Mais la propagation est quasi optique, c'est-à-dire que le relief et les obstacles empêchent la communication, et réduisent ainsi la portée maximale à quelques centaines de kilomètres seulement (370 km).

Pour pallier à ce problème, et afin d'améliorer la couverture, on utilise des stations déportées, où on installe plusieurs stations réparties sur le territoire, et on les exploite à distance à partir du Centre de Contrôle Régional. Ces stations, appelée également Antennes Avancées, couvrent chacune un rayon de 200 N.m environ.

1.5.3.3. Le Système de Contrôle de la Communication Vocale :

C'est un équipement central qui relie les radios, les pupitres d'exploitation, les enregistreurs et les lignes téléphoniques afin de permettre le contrôle et la commutation des voix de communication. Le **VCCS** permet aussi la gestion opérationnelle du trafic, en affectant à chaque position de travail un rôle avec les tâches qui lui sont dédiées.

1.6. Les enregistreurs ATC :

1.6.1. Le système d'enregistrement :

La chaine de sécurité est primordiale dans l'aviation c'est pour cela qu'un système d'enregistrement a été mis en avant au sol ; tel que la boite noire à bord de l'avion, celui-là permet l'enregistrement des conversations téléphoniques et radiophoniques. En cas d'incident, les enregistrements seront réécoutés et analysés pour déterminer les causes et les responsabilités.

On trouve deux générations, l'une est à base de bandes magnétiques, l'autre est informatisée et offre plus de capacité, d'option d'enregistrement et d'archivage.

Tous les étages de l'enregistreur sont redondé pour plus de sécurité, de plus le système fonctionne seulement sous une licence fournit par le constructeur sur clé **USB**, son système d'exploitation est sous linux comme il peut être sous Windows.

1.6.2. L'enregistreur ATC :

Les communications radiophoniques et téléphoniques doivent être enregistrées, pour permettre de comprendre les évènements et de déterminer les responsabilités lors des enquêtes sur les incidents et les accidents. Les enregistreurs des communications vocales sont donc les équivalents des boîtes noires des aéronefs, mais ils sont installés au sol.

Actuellement, on utilise des équipements complètement informatisés avec une grande capacité de stockage, une facilité d'exploitation et un grand nombre de fonctionnalités et d'options.

1.6.3. Le Fonctionnement des Équipements :

Les tours de contrôle sont toujours équipées des radios, essentiellement des **VHF** utilisées aussi bien pour le contrôle que pour les approches. Les radios sont constituées d'émetteurs et de récepteurs doublés, avec une fréquence principale et au moins une fréquence supplétive.

Le principe de la redondance est appliqué partout en aéronautique, et permet de renforcer la sécurité. Les antennes sont fixées sur le toit de la vigie. En plus des baies radios, on utilise des **VHF** portatives comme ultime secours, et aussi des talkies walkies. Le contrôle route, quant à lui, est assuré par le **CCR**, où l'exploitation des fréquences se fait à distance, et cela grâce aux systèmes de télécommande et aux supports de transmission qui relient toutes les antennes avancées au **CCR**.

Pour les trois types de contrôle, les enregistreurs sont installés dans les salles radio.
En Algérie, deux stations **HF** sont encore utilisées, l'une à Alger et l'autre à Tamanrasset, et elles sont exploitées à partir du **CCR**.

Dans le cas des équipements récents, on utilise un système de contrôle à distance (Remote Monitoring), où on relie l'équipement à un ordinateur pour effectuer des opérations de supervision, de réglage et de configuration. Un grand nombre de ces équipements offre la possibilité d'exploitation et de supervision sous réseau (par voix **IP**).

1.7. La Biométrie :

Les chercheurs s'investissent de plus en plus dans les techniques de sécurité, à savoir sécurité de l'information et sécurité des individus. La détermination de l'identité de ces derniers à travers des moyens simples, fiables, efficaces et peu dispendieux est devenue un problème crucial.

Traditionnellement, il existe deux manières pour y procéder.

a) basée essentiellement sur ce que l'on sait, à savoir mot de passe, code, etc.

b) s'appuie sur ce que l'on a, par exemple un badge, une clé, etc.

Ces deux méthodes présentent des inconvénients majeurs :

- Dans le premier cas, un mot de passe ou bien un code peuvent être oubliés par leur porteur ou devinés par une autre personne.
- Dans le second cas, un badge ou une clé peuvent être perdus, volés ou copiés par des personnes mal intentionnées.

L'inconvénient commun aux deux méthodes est que l'on identifie un objet et non la personne elle-même.

Pour pallier à ce problème, la nouvelle tendance qui apporte simplicité et confort aux utilisateurs consiste à identifier une personne à partir de la « BIOMETRIE ».

1.7.1. Définition :

Un système de contrôle biométrique peut être défini comme étant un système automatique de mesure basé sur la reconnaissance de caractéristiques propres à l'individu, la biométrie peut être vue comme étant une technique globale qui repose sur l'analyse mathématique des caractéristiques biologiques d'une personne, destinée à déterminer son identité de manière irréfutable.

1.7.2. Techniques biométriques :

On recense plus d'une dizaine de technologies biométriques classées en trois catégories : les premières reposent sur l'analyse morphologique ; les secondes sur l'analyse comportementale et les troisièmes enfin sur l'analyse de traces biologiques

1.7.2.1. La Morphologie :

Il existe plusieurs caractéristiques physiques qui se révèlent être uniques pour un individu ; Celles qui sont exploitées jusqu'à nos jours par les systèmes biométriques sont :

a) Empreintes digitales
b) Géométrie de la main
c) Iris
d) Rétine
e) Visage
f) Voix

1.7.2.2. Le Comportement :

Outre les caractéristiques physiques, un individu possède également plusieurs éléments liés à son comportement qui lui sont propres comme:

- Dynamique des frappes au clavier
- manière de parler
- Dynamique des signatures
- Démarche

1.7.2.3. La Biologie :

Comme caractéristiques biologiques, on peut citer :

1. Odeur
2. ADN
3. Salive

1.7.3. La Biométrie Vocale :

La biométrie vocale est la technique de reconnaissance la moins intrusive ; elle n'exige généralement aucun contact physique avec le récepteur du système. C'est une biométrie qui permet une reconnaissance à distance.

Elle se concentre sur les seules caractéristiques de la voix qui sont uniques à la configuration de la parole d'un individu.

1.8. Production de la parole :

1.8.1 La voix humaine :

La parole est depuis tout temps le moyen de communication privilégié de l'Homme. Elle véhicule, en plus du message linguistique prononcé, plusieurs types d'informations. Ces informations servent en particulier à déterminer l'identité du Locuteur ; elles sont exploitées par les humains pour l'identification des personnes qu'ils connaissent en particulier à distance (au téléphone par exemple).

1.8.2. Description Anatomique du Locuteur :

L'appareil vocal est constitué de structures appartenant à l'appareil respiratoire et à l'appareil digestif. On le décompose classiquement en trois étages :

a) **La soufflerie :** Elle comprend la musculature respiratoire, les poumons, et les conduits sus jacents, La soufflerie produit le flux d'air qui sera la matière première de la production vocale, expiré par les poumons et acheminé par la trachée vers le larynx.

b) **Le vibrateur :** Il s'agit du larynx qui est un tube situé à l'extrémité supérieure de la trachée, au niveau de la pomme d'Adam. La colonne d'air produite par la soufflerie est mise en vibration sous l'action des cordes vocales.

c) **Les résonateurs :** Ce sont principalement les cavités supra laryngées, à savoir le pharynx, la cavité buccale et les fosses nasales. La forme et le volume de ces cavités sont très variables selon les individus ; c'est ce qui explique que chaque personne ait un timbre de voix personnel et identifiable. Par ailleurs, les mouvements des muscles du pharynx et de la bouche (notamment : de la langue) permettent des modifications rapides du volume et de la forme de ces résonateurs qui transforment la voix produite par la vibration laryngée en phonèmes constitutifs de la parole articulée et ce, par l'amplification sélective de certaines fréquences laryngées.

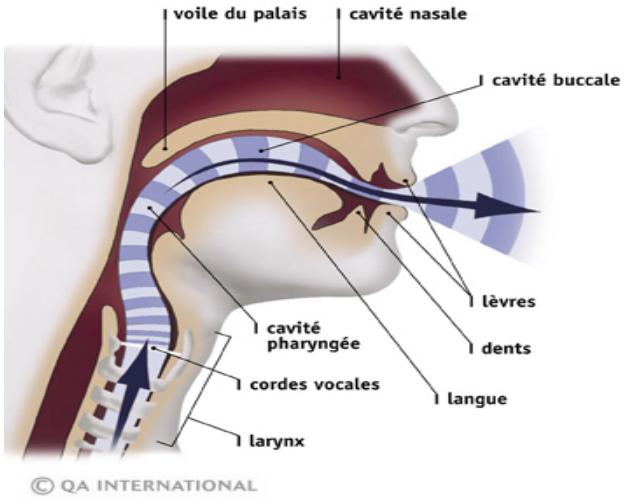


Figure 1.4 : L'appareil phonatoire humain.

1.8.3. Description Physique du Signal Vocal :

En plus du message linguistique servant à la communication entre individus, le signal de parole véhicule des informations caractéristiques de la personne qui l'a émis comme le timbre de sa voix, sa façon de parler, son état émotionnel ou pathologique, etc.

Ces informations caractéristiques du Locuteur peuvent être classées en deux catégories distinctes :

- Les informations de nature statique telles que les paramètres spectraux caractérisant les conduits vocal et nasal, la moyenne et les variations de la fréquence fondamentale.
- Les informations de nature dynamique reflétant les phénomènes de coarticulation, les trajectoires des formants ainsi que les informations temporelles (vitesse d'élocution, distribution des pauses).

Nous parlerons ici des caractéristiques statiques du signal vocal, Ce dernier peut être défini par 4 paramètres principaux:

1. Intensité : L'intensité d'un son correspond à l'amplitude de la vibration acoustique ; elle caractérise le volume sonore qui nous permet de distinguer un son fort d'un son faible.

2. Timbre : Le timbre permet de différencier deux sons de même hauteur et de même amplitude. Il est constitué d'un ensemble de fréquences appelé spectre. La richesse du spectre permettra de dire qu'un son est riche, brillant, profond, etc.

3. Hauteur : La hauteur dépend de la fréquence de la variation de pression acoustique correspondant au son.

4. Fréquence : Elle représente le nombre de vibrations de l'air en une seconde.

1.9. Perception de la parole :

La perception de la parole est effectuée par 1'appareil auditif (oreille) qui est constitué de l'oreille externe, l'oreille moyenne et l'oreille interne. La perception de l'appareil auditif humain a une bande de fréquences qui s'étend entre 800 Hz et 8 KHz et au maximum entre 20Hz et 20KHZ. La figure 1.5 illustre l'appareil auditif:

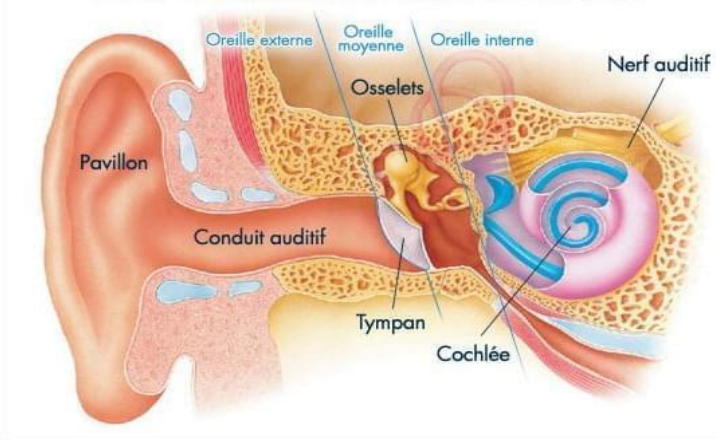


Figure 1.5 : l'appareil auditif humain.

1.10. La Reconnaissance automatique du locuteur :

La Reconnaissance Automatique du Locuteur s'inscrit dans le cadre général du traitement automatique de la parole, la reconnaissance automatique du locuteur consiste à reconnaitre l'identité d'une personne par l'analyse de sa voix. Objet d'un intérêt accru depuis quelque temps au même titre que l'ensemble des méthodes d'authentification dites biométriques, elle ne figure pas parmi les plus fiables de ces techniques, au premier rang desquelles on retrouve l'analyse des empreintes digitales et génétique.

Cependant la RAL présente un certain nombre de qualité qui la distingue de ces dernières notamment en matière de facilité de déploiement. Les systèmes de **RAL** sont sensibles à certains facteurs qui peuvent altérer leur performance ; ces facteurs peuvent être intrinsèques ou extrinsèques au locuteur. On peut citer :

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

1.11. Conclusion :

L'installation et l'exploitation des moyens des télécommunications dans le domaine de l'aviation civile doivent obéir aux normes internationales, selon les recommandations de l'Organisation Internationale de l'Aviation Civile, dans le but d'atteindre l'objectif primordial qui est la sécurité aéronautique. Les efforts et les travaux doivent se concentrer autour de cette idée, surtout avec le développement du trafic aérien, la coopération internationale et les technologies modernes.

Chapitre II : Système de Reconnaissance Automatique Du Locuteur

2.1. Introduction :

Les systèmes de Reconnaissance Automatique du Locuteur s'intéressent précisément aux caractéristiques particulières du signal de la parole. Cette discipline s'inscrit dans le cadre général de la reconnaissance de formes ; c'est un terme générique qui regroupe les problèmes relatifs à l'identification ou à la vérification du Locuteur basé sur l'information contenue dans le signal acoustique ; il est question de reconnaître une personne à partir de sa voix où le champ d'application est très vaste, il va du simple contrôle d'accès, aux applications militaires passant par des applications judiciaires. C'est pour ça on utilise la vérification par la biométrie vocale qui est l'une des méthodes modernes pour la vérification des locuteurs. Elle est particulièrement utilisée dans les communications téléphoniques, la radiocommunication aéronautique, les centres d'appels, les banques, les assurances, et dans les entreprises de télécommunication.

2.2. Processus d'identification biométrique :

Tout système biométrique est composé de deux grandes étapes :

L'enrôlement et le contrôle.

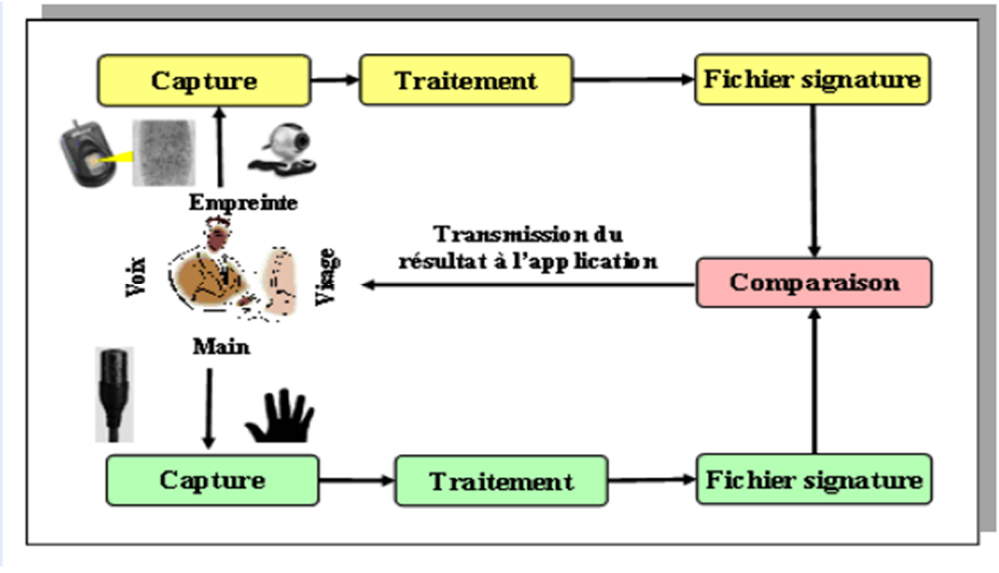


Figure 2.1 : processus d'un système d'identification biométrique

- L'enrôlement des personnes est la phase initiale de création du gabarit biométrique et de son stockage en liaison avec une identité déclarée.

Les caractéristiques physiques sont transformées en un modèle représentatif de la personne et propre au système de reconnaissance, Cette étape n'est effectuée qu'une seule fois.

- Le contrôle d'un autre côté, représente l'action de contrôler les données d'une personne afin de procéder à la vérification de son identité proclamée ou à son identification, Cette étape se déroule à chaque fois qu'une personne se présente devant le système.

2.3. Identification et Vérification :

Il convient de distinguer deux modes de fonctionnement des systèmes biométriques ; systèmes d'identification et système de vérification.

- L'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.
- La vérification (authentification) 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.

2.4. Types de reconnaissance automatique du locuteur :

2.4.1. Reconnaissance Auditive :

Utilisée jusqu'à nos jours dans le domaine juridique, l'identification auditive se base essentiellement sur la capacité naturelle de l'être humain à reconnaître une personne en utilisant seulement l'écoute de sa voix.

2.4.2. Reconnaissance par spectrogramme :

Une 'empreinte vocale' est en fait un terme qui fait référence à un spectrogramme du signal vocal, Il s'agit d'un graphique qui représente le signal en trois dimensions : temps, fréquence et intensité.

Le spectrogramme est un outil utile pour le traitement et l'analyse de la voix, mais n'a cependant aucun lien avec les empreintes digitales ou génétiques. Cette reconnaissance se fait par comparaison spectrale (spectrographiques) des mots.

La voix présente des différences majeures avec les empreintes digitales et génétiques, Elle évolue dans le temps, elle peut être modifiée volontairement par son porteur, elle est

facilement falsifiable, etc. Par conséquent, on ne parle pas d'empreinte vocale mais plutôt de signature vocale.

2.4.3. Reconnaissance phonétique :

Cette méthode utilise une approche linguistique ; L'information recueillie à travers l'étude systématique des sons d'une langue est utilisée par un expert phonéticien pour produire une preuve correspondant à la vraisemblance pour qu'un enregistrement vocal ait été produit par une personne donnée. Plus exactement, il s'agit d'estimer combien de fois il est plus probable d'observer une différence entre les exemples vocaux si ceux-ci proviennent du même Locuteur ou, au contraire, de deux Locuteurs différents (le rapport de vraisemblance Bayésien est utilisé).

2.4.4. Reconnaissance Automatique :

Elle consiste à reconnaitre l'identité d'une personne par l'analyse de sa voix. Cette approche automatique sous tous ses aspects sera détaillée dans la suite de ce chapitre.

2.5. Différentes tâches en RAL :

Les deux tâches pionnières des systèmes de Reconnaissance Automatiques du Locuteur sont l'Identification Automatique du Locuteur **(IAL)** et la Vérification Automatique du Locuteur **(VAL)**.

Récemment, des besoins spécifiques ont stimulé l'apparition de nouvelles tâches comme l'indexation du Locuteur qui consiste à indiquer à quel moment chaque Locuteur intervenant dans une conversation a pris la parole, le suivi de Locuteurs (speaker tracking) ou bien une application connexe à l'indexation qui est la détection d'un Locuteur lors d'une conversation.

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

2.5.1. Identification Automatique du Locuteur :

L'Identification Automatique du Locuteur est le processus qui consiste à déterminer, parmi une population de Locuteurs connus, la personne ayant prononcé un message donné.

Cela est fait en calculant des mesures de similarité entre le signal en entrée et tous les modèles des Locuteurs de la base.

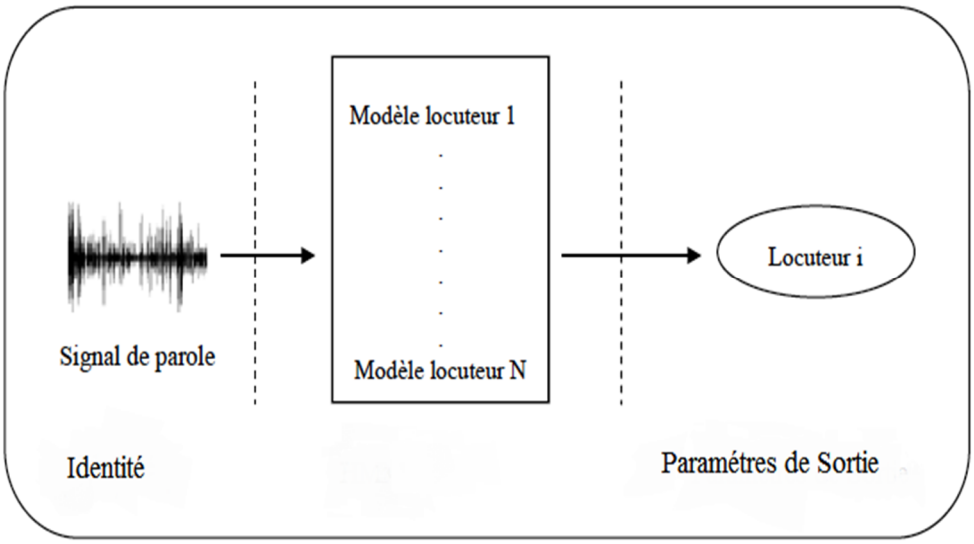


Figure 2.2 : Schéma typique d'un système d'IAL.

L'identité du Locuteur, dont le modèle est le plus proche du signal en entrée, est donnée en sortie du système d'**IAL** (voir Figure 2.2).

Deux modes d'identification sont possibles :

1. Identification en ensemble fermé : c'est le cas où le système doit fournir comme sortie un ensemble d'au moins un Locuteur. En d'autres termes, la séquence fournie en entrée doit être en fait prononcée par un Locuteur connu du système.

2. Identification en ensemble ouvert : le système dans ce cas peut être amené à fournir un ensemble vide, car le Locuteur peut ne pas être connu.

Dans ce mode, le système d'**IAL** doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée. En pratique, la plupart des systèmes d'**IAL** fournissent un ensemble d'un seul Locuteur qui représente le Locuteur le plus proche.

2.5.2. Vérification Automatique du Locuteur :

La Vérification Automatique du Locuteur est une décision en tout ou en rien. Elle consiste à déterminer à partir d'un message vocal, la véracité de l'identité proclamée par un individu.

Les entrées du système sont donc le signal de parole et l'identité proclamée et la sortie une acceptation ou bien un rejet.

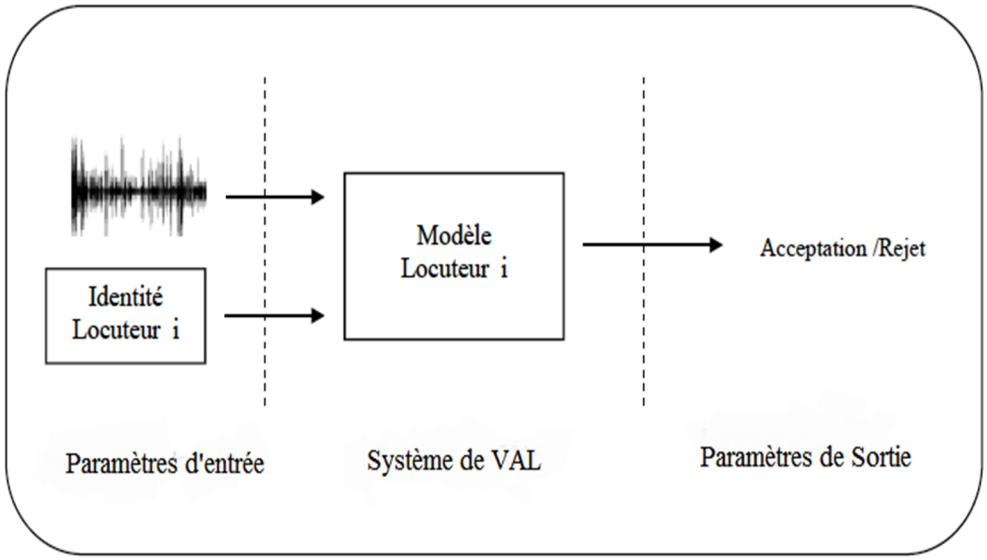


Figure 2.3 : Schéma typique d'un système VAL.

Dans la suite de ce chapitre, nous nous intéresserons à l'Identification Automatique du Locuteur sujet de ce présent travail.

2.6. Structures de systèmes d'IAL:

Un système d'**IAL** se résume à l'enchaînement de trois processus principaux qui sont :

1. La paramétrisation.

2. La modélisation.

3. La décision.

En premier lieu, le message vocal est analysé acoustiquement. À l'issue de cette analyse, on obtient un ensemble de vecteurs de coefficients pertinents qui vont être utilisés pour la modélisation des Locuteurs. À la reconnaissance, une mesure de similarité va être calculée entre les paramètres acoustiques du signal prononcé et les modèles contenus dans la base.

La dernière étape du système est un module de décision qui est basé sur une stratégie de décision donnée et qui fournit la réponse du système.

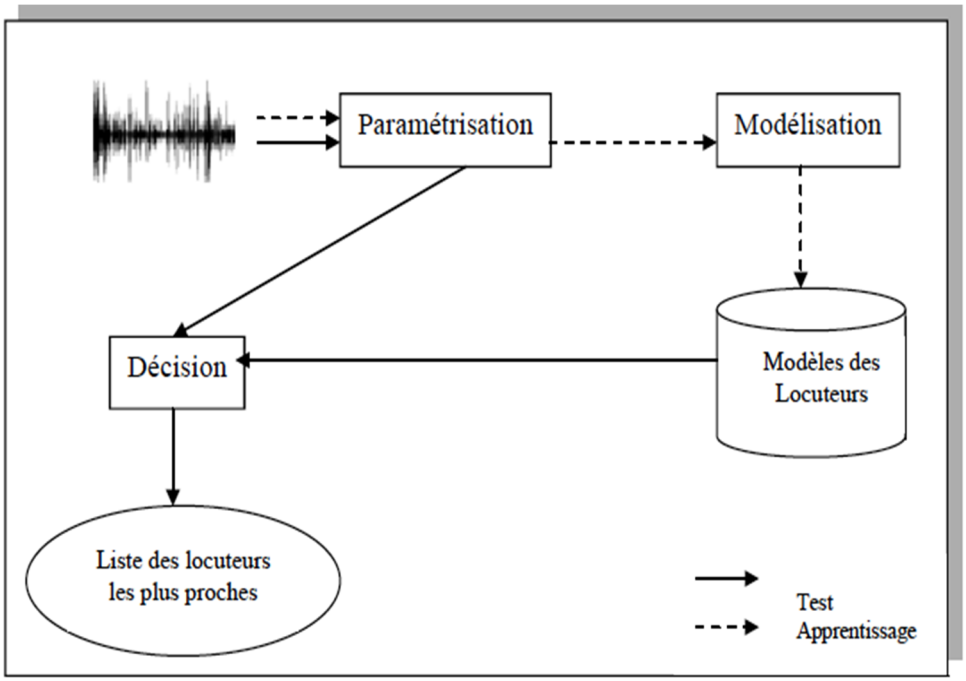


Figure 2.4 : Schéma modulaire d'un système d'IAL.

2.6.1. Paramétrisation Acoustique :

Le processus de paramétrisation consiste à extraire du signal de parole l'information pertinente et à réduire au maximum la redondance en vue de la reconnaissance. C'est une représentation plus simple du signal de parole sous forme de vecteurs de paramètres acoustiques. Le calcul de ces derniers est réalisé en glissant avec une cadence régulière (ex : 10 ms) une fenêtre de pondération d'une longueur variant généralement de 20 à 32 ms. Le fenêtrage le plus utilisé en traitement du signal de parole est en général le fenêtrage de Hamming.

Chaque fenêtre nous permet d'avoir une trame. Les trames ainsi obtenues sur tout le signal de parole sont traitées par la suite afin de produire les vecteurs de paramètres acoustiques. Nous retrouvons dans la littérature trois grandes familles de paramètres :

1. Paramètres de l'analyse spectrale : L'analyse spectrale est l'analyse la plus employée en RAL. Les paramètres qui en découlent sont généralement représentatifs des caractéristiques physiques de l'appareil phonatoire (forme du conduit vocal) de chaque individu.

2. Paramètres prosodiques : Ces paramètres illustrent en général le style d'élocution d'un Locuteur : vitesse d'élocution (débit), durée et fréquence des pauses, ainsi que les caractéristiques de la source glottale (fréquence fondamentale, énergie, ...).

3. Paramètres dynamiques : Le vecteur de paramètres issus des paramétrisations précédentes peut être complété par le vecteur correspondant aux dérivées des premiers et seconds ordres de ces paramètres.

Ces dérivées sont les paramètres dynamiques les plus répandus ; on les appelle aussi coefficients Delta (dérivée première) et Delta-Delta (dérivée seconde).

2.6.2. Modélisation des Locuteurs :

Le processus d'**IAL** se base essentiellement sur la phase de modélisation des caractéristiques des Locuteurs. Cette modélisation est réalisée à partir de données d'apprentissage collectées au cours des sessions d'enrôlement.

Les méthodes existantes de modélisation des Locuteurs peuvent être répertoriées en cinq grandes approches :

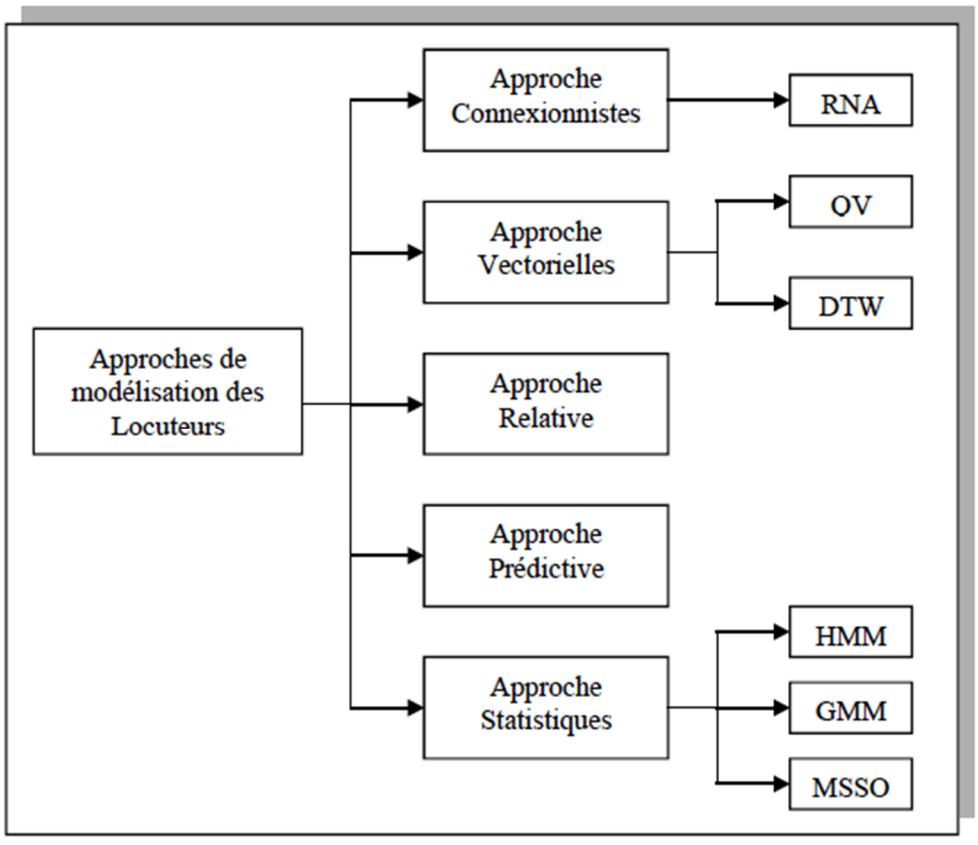


Figure 2.5 : Approches de modélisation des locuteurs.

2.6.2.1. Approches Vectorielles :

Dans l'approche vectorielle, un modèle de Locuteur est un ensemble de vecteurs de paramètres représentatifs de l'espace acoustique issus de la phase de paramétrisation des signaux d'apprentissage.

Dans cette approche, on retrouve deux grandes techniques : la programmation dynamique et la quantification vectorielle.

- **La Déformation Temporelle Dynamique :**

La déformation temporelle dynamique consiste à aligner temporellement une séquence de vecteurs de paramètres de test avec une séquence de vecteurs d'apprentissage.

La programmation dynamique est facile à mettre en œuvre, très rapide et montrant des performances relativement bonnes. La programmation dynamique est toutefois très sensible à la qualité de la déformation et notamment au choix du point de départ des deux formes à comparer.

- **Quantification vectorielle :**

La quantification vectorielle repose sur un partitionnement de l'espace acoustique en sous-espaces. Dans ces conditions, un modèle de Locuteur est composé d'un ensemble de vecteurs centroïdes, appelé dictionnaire de quantification. Lors de la phase de reconnaissance, une distance est calculée entre un vecteur de test et chaque vecteur centroïde du dictionnaire. La distance minimale est retenue. La quantification vectorielle s'applique en mode dépendant ou indépendant du texte.

La rapidité et les performances de cette technique dépendent fortement de la taille du dictionnaire : plus la taille du dictionnaire augmente, meilleures sont les performances, Néanmoins, le processus devient d'autant plus lent.

2.6.2.2. Approches Connexionnistes :

L'approche connexionniste repose sur la discrimination entre Locuteurs. Un ensemble de signaux de parole issus d'une population de Locuteurs clients est fourni en entrée à un réseau de neurones pour une étape d'apprentissage. À l'issue de cette étape, le réseau apprend à discriminer un Locuteur des autres. L'approche connexionniste se résume, par conséquent, à une tâche de classification.

2.6.2.3. Approches Statistiques :

L'inconvénient commun aux méthodes présentées précédemment est qu'elles ne tiennent pas compte de l'ordre dans lequel les vecteurs de paramètres sont présentés.

L'approche statistique résout ce problème en utilisant des techniques qui permettent de construire des modèles qui prennent en considération l'aspect temporel du signal de parole. Les vecteurs acoustiques issus de la paramétrisation sont donc représentés par des statistiques à long terme.

- **Les Modèles de Markov Cachés :**

Les **HMM** ont été largement utilisés en Reconnaissance Automatique de la Parole. Plus récemment, leur utilisation s'est étendue à la Reconnaissance Automatique du Locuteur.

La modélisation dans ce cas de figure se fait par une succession d'états avec des probabilités de transition d'un état à l'autre. La reconnaissance se fait par calcul de la vraisemblance d'une séquence de vecteurs de test qui est issue de la chaîne de Markov.

- **Les Modèles de Mélange de Gaussiennes :**

Les **GMM** sont considérés comme étant la modélisation des systèmes de Reconnaissance Automatique du Locuteur en mode indépendant du texte. Cette approche consiste à modéliser un Locuteur par un mélange de gaussiennes qui représente une somme pondérée de M gaussiennes multidimensionnelles.

- **Méthodes Statistiques du Second Ordre :**

Cette approche est généralement associée à une famille de mesures de similarité entre Locuteurs en vue de la reconnaissance. On peut citer : rapport de vraisemblance, distance de Kullback-Leiber, maximum de vraisemblance, etc.

Le modèle d'un Locuteur se résume au triplet $\{x; X_0; M\}$ où x est un vecteur moyen, X_0 est une matrice de covariance ; tous les deux estimés à partir de la séquence de M vecteurs acoustiques. Les mesures de similarité reposent ainsi essentiellement sur une ressemblance entre les matrices de covariance de test et d'apprentissage

2.6.2.4. Approche Prédictive :

L'approche prédictive repose sur le principe qu'une trame de signal peut être prédite par la seule observation des trames précédentes.

Par ce concept, cette approche est considérée dans la littérature comme une approche dynamique, c'est-à-dire : 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 Locuteur 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 de Locuteur (fonction de prédiction).

2.6.2.5. Approche Relative :

Le principe de la reconnaissance relative des Locuteurs a été initialement appliqué en reconnaissance de la parole dans des techniques d'adaptation rapide, Ces approches ont donné naissance à la notion «d'espace de Locuteurs» où un modèle de Locuteur est représenté par rapport à un ensemble de Locuteurs bien appris. Les principales techniques utilisées dans ce domaine sont : **RMP,** Speaker Clustering**, RSW,** et les voix propres (eigenvoices).

2.6.2.6. Tableau Récapitulatif des Approches de Modélisation des Locuteurs :

Nous reprenons dans le tableau ci-après les approches de modélisation des Locuteurs en mettant l'accent sur leurs avantages et inconvénients.

Tableau 2.1 : étude comparative des Approches de Modélisation des Locuteurs.

Approches	Avantages	Inconvénients
DTW	- Très rapide. - Présente des performances relativement bonnes.	- Utilisée exclusivement en mode dépendant du texte. - Très sensible à la qualité d'alignement des vecteurs et au choix du point de départ.
QV	- S'applique en mode dépendant ou - indépendant du texte.	-Sa rapidité et ses performances dépendent fortement de la taille du dictionnaire.
Approches Connexionnistes	- Bonne performance	Complexité d'apprentissage. - L'ajout d'un nouveau client nécessite le réapprentissage de tous les modèles.

Approche Prédictive	- L'information dynamique transportée par le signal de parole est prise en considération.	- Les performances obtenues ne sont pas assez suffisantes pour un usage pratique.
Approche Relative	- La modélisation d'un nouveau Locuteur ne se fait plus de façon absolue mais relativement à un ensemble de locuteurs bien appris.	- Le taux d'identification dépend de la quantité de données d'apprentissage pour la construction des locuteurs de référence.
HMM	- Prend en considération l'aspect temporel du signal de parole - Excellents résultats en mode dépendant du texte.	- Utilisé uniquement en mode dépendant du texte.
GMM	- Très bonnes performances en mode indépendant du texte.	- Quantité importante de signaux d'apprentissage requise pour une bonne évaluation des paramètres du modèle.
MSSO	- Simplicité de mise en œuvre Performante sur de courtes durées.	- Ne capture que les caractéristiques stables le long du signal de parole. - les variations locales ne sont pas prises en compte.

2.6.3. Décision :

Après avoir comparé le signal de test à tous les modèles de Locuteurs connus du système, on obtient un ensemble de mesures de similarité qui va servir d'entrée au module de décision.

Ce dernier a pour tâche de rechercher la mesure de similarité maximale ou bien minimale en terme de distance et d'indiquer l'identité du Locuteur.

Pour mesurer les performances d'un système d'IAL, on utilise généralement le taux d'identification correcte Ic ou incorrecte Ii qu'on obtient par les formulations suivantes :

$$I_c = \frac{nombre\ de\ tests\ correctement\ identifiés}{nombre\ total\ de\ tentatives} \qquad (2.3)$$

Et

$$I_i = \frac{nombre\ de\ tests\ mal\ identifiés}{nombre\ total\ de\ tentatives} \qquad (2.4)$$

Avec $I_c + I_e = 100\%$ (2.5)

2.7. Conclusion :

Dans ce chapitre, nous avons introduit le principe de la Reconnaissance Automatique du Locuteur.

Nous avons décrit en premier lieu le Locuteur sous ses deux facettes anatomiques et acoustiques afin de justifier l'utilisation de la voix dans le domaine de reconnaissance.

Les différentes étapes d'un système de **RAL** ont été présentées ainsi qu'un état de l'art sur les approches de modélisation du Locuteur.

Chapitre III : Traitement, Analyse et Classification du Signal Vocal

3.1. Introduction :

L'extraction des caractéristiques est une étape importante dans le processus de reconnaissance de la parole. En effet, cette étape permet d'extraire des caractéristiques qui seront ensuite utilisées par le classificateur. La phase d'extraction des caractéristiques doit être faite avec soin, car elle contribue directement aux performances du système global. Nous allons voir dans ce chapitre les paramètres couramment utilisés qui sont les paramètres **MFCC**.

On va voir aussi que le vecteur acoustique est renforcé par les paramètres dynamiques et énergie.

3.2. Approche acoustique :

La phonétique articulatoire essaye de décrire comment des sons de la parole sont produits en termes de gestes articulatoires, tandis que l'approche acoustique vise une conclusion des corrélations acoustiques de la physiologie et des aspects comportementaux des organes de production de la voix. Le son articulé acoustique ne porte pas une agrafe visuelle des mouvements de lèvre.

Cependant, certains paramètres acoustiques ont plus ou moins des corrélations avec l'anatomie et la physiologie des organes de production de parole. Des sons articulés peuvent être analysés dans le domaine temps ou fréquence.

3.3. Les différentes étapes utilisées pour le Traitement du Signal Vocal :

3.3.1. La Transformée de Fourier Discrète :

Supposons que $s[n]$, $n = 0,1,\ldots..N\text{-}1$ une séquence de temps discret de N échantillons. La transformée de Fourier discrète de $s[n]$ est définie comme suite :

$$\hat{S}[k] = f\{s[n]\} = s[\mathrm{n}] \sum_{n=0}^{N-1} s[n] e^{-j2\pi nk/N} \quad , 0 \leq k \leq N\text{-}1 \qquad (3.1)$$

Où k représente la variable discrète de fréquence et j l'unité imaginaire. Le résultat de la **DFT** est un nombre complexe de longueur N.

3.3.2. Transformée de Fourier Rapide :

La **FFT** est un algorithme de calcul de la **TFD**.

Sa complexité varie en O (n log n) avec le nombre n de points, alors que la complexité de l'algorithme « naïf » s'exprime en O (n^2). Ainsi, pour n = 1 024, le temps de calcul de l'algorithme rapide peut être 100 fois plus court que le calcul utilisant la formule de définition de la **TFD**.

Cet algorithme est couramment utilisé en traitement numérique du signal pour transformer des données discrètes du domaine temporel dans le domaine fréquentiel, en particulier dans les analyseurs de spectre.

Soient x_0, ..., x_{n-1} des nombres complexes. La transformée de Fourier discrète est définie par la formule suivante :

$$f_j = \sum_{k=0}^{n-1} x_k \, e^{\frac{2\pi i \; jk}{n}} \qquad j=0,\ldots,n-1 \qquad (3.2)$$

3.3.3. Filtres Numériques :

Un filtre est un système qui modifie le signal d'entrée $s[n]$ en un signal de sortie $y[n]$. Il y a plusieurs manières d'indiquer un filtre numérique.

- Dans le domaine temporel, le filtre est caractérisé par sa réponse impulsionnelle $h[n]$ qui peut être fini (Filtre **RIF**) ou infini (filtre **RII**).
- Dans le domaine fréquentiel, un filtre est spécifié par sa fonction de transfert $H(z)$, où z est une variable complexe.

Dans le domaine temporel, le filtrage est présenté comme convolution entre le signal d'entrée et la réponse impulsionnelle $h[n]$:

$$y[n] = s[n] * h[n] = \sum_{k=-\infty}^{+\infty} s[k]h[n-k] \qquad (3.3)$$

Dans la pratique, ceci est mis en application en utilisant un rapport récursif :

$$y[n] = \sum_{k=0}^{N} a[k]s[n-k] - \sum_{k=1}^{M} b[k]y[n-k] \qquad (3.4)$$

Où les coefficients $a[k]$, $b[k]$ sont déterminés à partir des caractéristiques du filtre.

La dernière somme dans (3.4) représente la partie de rétroaction du filtre et elle est nulle pour les filtres RIF ($b[k] = 0$ pour tout k). La fonction de transfert $H(z)$ de (3.4) est obtenu par prendre la transformée Z des deux côtés et la résoudre pour $H(z)=Y(z) / S(z)$:

$$H(z) = \frac{Y(z)}{S(z)} = \frac{\sum_{k=0}^{N} a[k]z^{-k}}{1+\sum_{k=0}^{N} b[k]z^{-k}} \quad (3.5)$$

On appelle les racines du numérateur de (3.5) les zéros du système, et les racines du dénominateur les pôles du système. Un pôle cause une résonance (crête) dans la réponse d'amplitude du filtre, tandis qu'un zéro cause une antirésonance (vallée).

Par exemple, la fonction de transfert du conduit vocal d'un son de voyelle peut être bien caractérisée seulement par des pôles, qui correspondent aux endroits des formants. D'autre part, les sons nasaux comme [n] ayant en plus des résonances, des antirésonances dans leur spectre, et donc les pôles et les zéros sont nécessaires dans la modélisation.

Dans le domaine fréquentiel, le filtrage est effectué en multipliant point à point la **DFT** du signal d'entrée par la fonction de transfert du filtre. S'accorder au théorème de convolution, tandis que la multiplication dans le domaine fréquentiel correspond à la convolution dans le domaine temporel, et vice versa :

$$s[n]*h[n] \leftrightarrow S(z)\ H(z) \quad (3.6)$$

$$s[n]\ h[n] \leftrightarrow S(z)*\ H(z) \quad (3.7)$$

3.3.4. L'analyse Spectrale :

Puisque le son de la parole change en continu en raison des mouvements articulatoires des organes vocaux de production, le signal doit être traité avec des petits segments, dans lesquels les paramètres quasi-stationnaires demeurent (Figure 3.1).

Le calcul de la **DFT** du signal entier jetterait les propriétés spectrales locales qui présentent des réalisations de différents phonèmes. Au lieu d'exécuter la **DFT** pour le signal entier, une fenêtre **DFT** est calculée. Un segment en général autour de 10-30 millisecondes est multiplié par une fonction fenêtre, et la **DFT** du segment fenêtré est alors calculée. Ce processus est répété jusqu'à la fin du son articulé, de sorte que le segment soit décalé en avant par une quantité fixe des points, en général autour de 30 à 75 % de la longueur du segment.

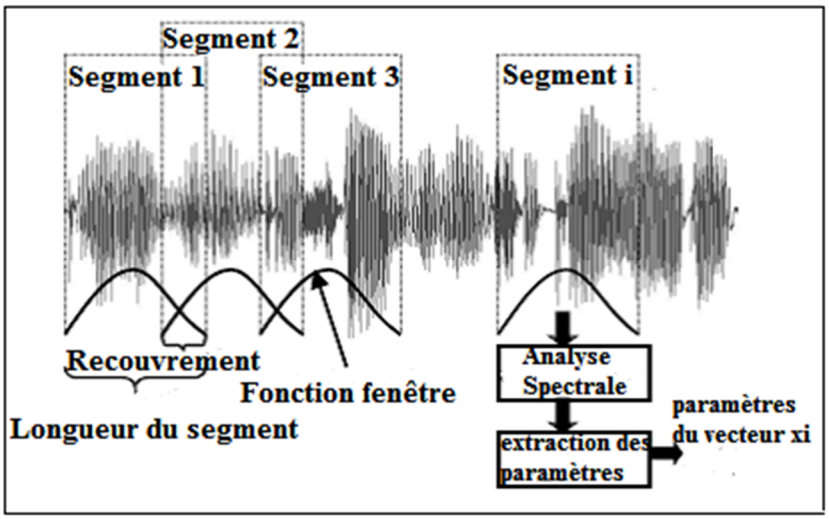


Figure 3.1 : l'analyse spectrale

La transformée de Fourier $X(f)$, qui est habituellement une fonction complexe, décompose le signal en éléments de bases formées par des cosinus et des sinus de durée illimitée.

Elle ne possède pas de localisation temporelle. Cela est dû à son incapacité de définir temporellement les différentes valeurs des composantes fréquentielles. Elle est donc limitée aux signaux stationnaires.

Pour remédier à ce problème, on utilise la transformée de Fourier à fenêtre glissante, appelée aussi **TFCT**. Cette dernière permet de calculer la transformée de Fourier d'un signal sur des segments temporels de durée finie. Ces segments du signal $X(f)$ sont extraits à l'aide d'une fenêtre glissante $g(t\text{-}k)$ de taille fixe. La fenêtre glissante peut être par exemple une fenêtre de Hamming ou une fenêtre gaussienne.

La transformée de Fourier à fenêtre glissante est donnée par l'équation suivante :

$$X_{STFT}(f,k) = \int_{-\infty}^{+\infty} x(t)g(t-k)e^{-j2\pi ft} \qquad (3.8)$$

3.3.5. La Fonction Fenêtre :

Le but du fenêtrage est de réduire l'effet résultant du processus de segmentation. Le fenêtrage dans le domaine temporel est une multiplication point par point entre le segment et la fonction fenêtre. Selon *le théorème de convolution*, ceci correspond à une convolution du spectre à court terme avec la réponse d'amplitude de la fonction fenêtre.

En d'autres termes, la fonction de transfert de la fenêtre sera présente dans le spectre observé. Une bonne fonction fenêtre a un lobe principal étroit et des petits lobes secondaires dans sa fonction de transfert. Il y a une compensation entre ces deux conditions : rendre le lobe

principal plus étroit augmente le niveau des lobes secondaires, et vice versa. En général, une fonction fenêtre appropriée diminue aux bords de segment de sorte que l'effet des discontinuités est diminué.

Intuitivement le fenêtrage le plus simple est le fenêtre rectangulaire.

Définie comme suit :

$$w\{n\}=\begin{cases}1, & 0 \leq n \leq N-1 \\ 0, & autrement\end{cases} \quad (3.9)$$

Bien que la fenêtre rectangulaire préserve la forme d'onde originale sans changement, elle est rarement utilisée en raison de ses effets spectraux. Généralement dans le traitement de la parole, la fonction fenêtre la plus utilisée est la fenêtre de Hamming définie comme suite :

$$w\{n\}=\begin{cases}0.56-0.46\cos(\frac{2\pi n}{N}), & 0 \leq n \leq N-1 \\ 0, & autrement\end{cases} \quad (3.10)$$

Des exemples du fenêtrage sont montrés sur les figures 3.2 et 3.3 qui montre des segments voisés et non voisé de parole prononcé par le même locuteur. On peut voire d'après le segment voisé, que la fenêtre de Hamming donne moins de fuite spectrale. La fenêtre rectangulaire cause aux harmoniques de F_0 des harmoniques voisins, et comme résultat, on lisse les différents harmoniques.

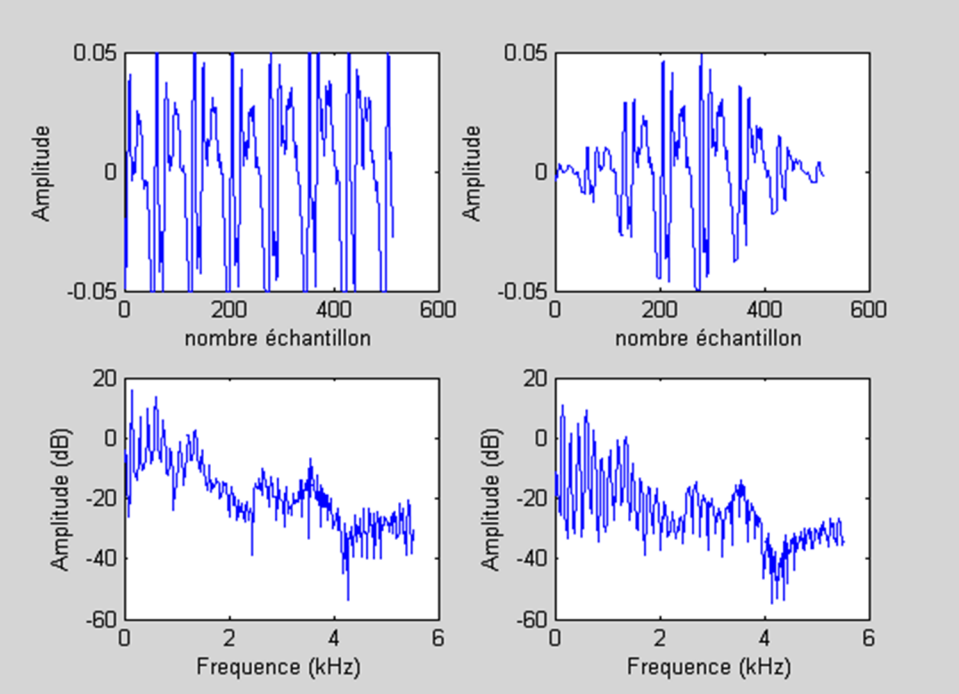


Figure 3.2 : Segment d'un son voisé [voyelle a] fenêtré à gauche par une fenêtre rectangulaire et à droite par une fenêtre de Hamming.

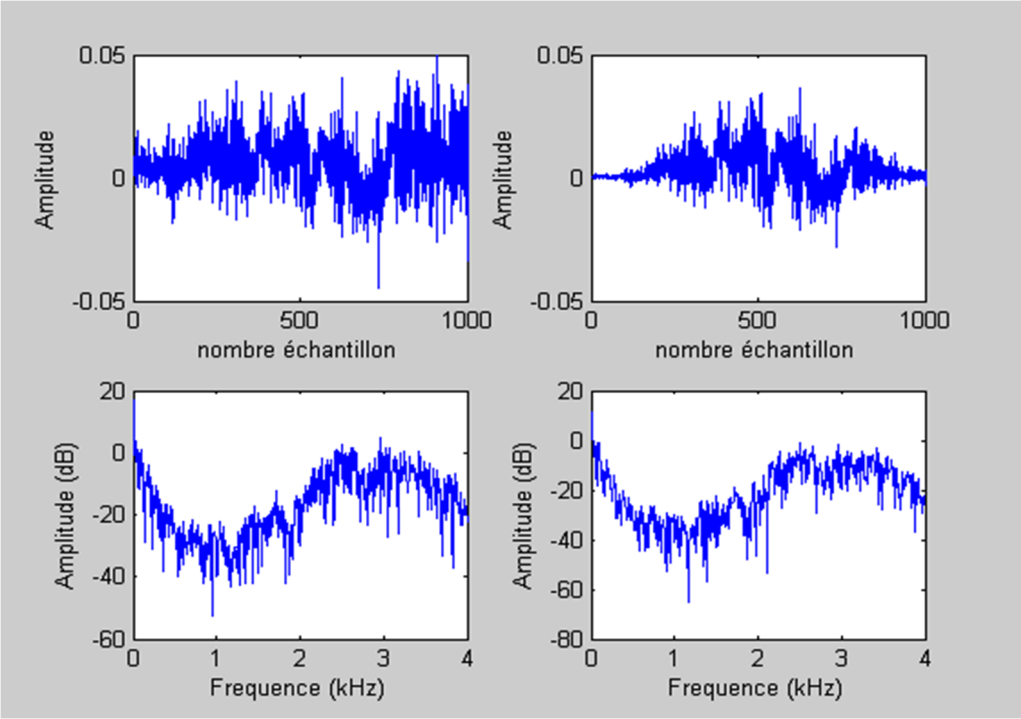


Figure 3.3 : Segment d'un son non voisé [ch] fenêtré à gauche par une fenêtre rectangulaire et à droite par une fenêtre de Hamming.

3.3.6. Longueur et chevauchement des fenêtres :

Le choix de longueur pour la fenêtre est un paramètre décisif pour une bonne analyse spectrale, due à la compensation entre les résolutions temps et fréquence. La fenêtre devrait être assez longue pour une résolution fréquentielle stable, mais d'autre part, elle devrait être assez courte de sorte qu'elle capture les propriétés spectrales locales.

Typiquement une durée de 10-30 millisecondes est utilisée. Pour des femelles et des enfants, le pitch tend à être plus haut, et une fenêtre plus courte employée que pour les locuteurs masculins qui ont un pitch moins petit.

3.4. Représentation du Signal Vocal :

Dans le contexte de la reconnaissance vocale le but essentiel de la phase d'extraction de paramètre est de calculer une séquence des vecteurs de paramètres qui donnent une représentation compacte pour un signal vocal d'entrée donné. La question de base est comment faire pour passer le problème de redondance et la variabilité causée par le signal vocale.

L'extraction des paramètres est habituellement réalisée dans trois phases.

1. La première phase est appelée l'analyse acoustique, elle est prévue pour produire une représentation de l'enveloppe spectre d'énergie court terme. Ce spectre est une version lisse du spectre d'énergie détaillé et montre une variabilité sensiblement plus petite que le spectre origine.
2. La deuxième étape et de compiler un vecteur étendu composé de paramètres statiques et dynamiques.
3. La troisième phase qui n'est pas toujours présente, transforme ces vecteurs étendus en des vecteurs plus compacts qui sont alors fournis au système de reconnaissance.

3.5. Les Bancs de Filtres :

Les bancs de filtres ont l'avantage par rapport à la plupart des autres représentations spectrales parce que les paramètres ont une interprétation physique directe. Ceci permet, par exemple, l'utilisation de la connaissance a priori des puissances de discrimination des sous-bandes pour pondérer les sous-bandes. En outre, si des sous-bandes sont bruitées, les sous-bandes propres peuvent toujours être employées.

En général deux approches basées sur les bancs de filtres sont utilisées dans les systèmes **RAP**

- Fusion au niveau des paramètres (fusion en entrée).
- Fusion au niveau du classificateur (fusion en sortie).

Dans le premier cas, les sorties des sous-bandes sont combinées dans un seul paramètre de dimension *M*. Le vecteur représente une unité de son de parole. Dans l'autre cas, chaque sous-bande est considérée comme indépendante et un modèle est généré séparément pour chaque sous-bande.

L'approche la plus simple à l'extraction des paramètres ; est de considérer les sorties des sous-bandes directement comme paramètres. La prolongation normale de cette approche est de pondérer chacune des sous-bandes par une masse représentant la puissance de discrimination de chaque sou bande, Figure 3.4.

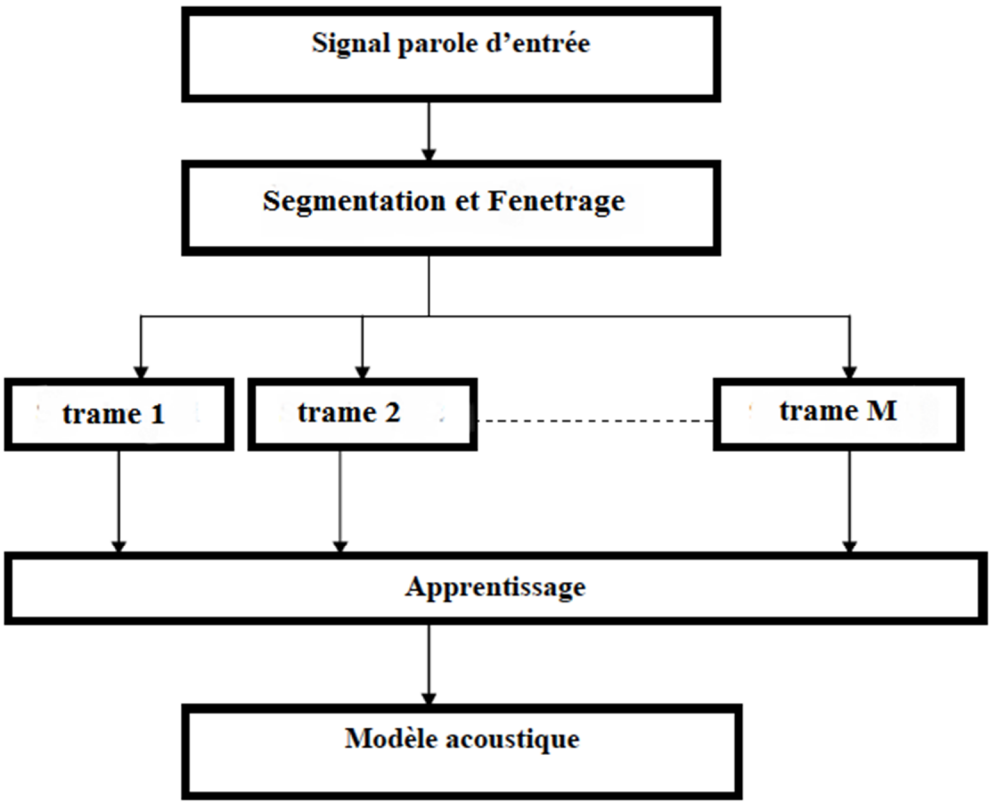


Figure 3.4 : Extraction des paramètres par banc de filtres (fusion au niveau des paramètres).

Une autre approche pour combiner l'information de plusieurs sous-bandes, est de modeler chaque sous-bande indépendamment des autres, en suite faire une combinaison des points des modèles des sous-bandes dans le classificateur (Figure 3.6). Le résultat de combinaison au niveau de la sortie du classificateur est flexible.

Pour chaque sous bande de chaque classe, un modèle séparé doit être stocké, et dans l'étape d'identification, chaque classificateur doit calculer ses propres points. Le temps global augmente avec le nombre de sous bandes et la complexité des classificateurs. La figure 3.5 montre un des principes utilisés dans l'architecture des systèmes à base d'un banc de filtre et la fusion au niveau du classificateur.

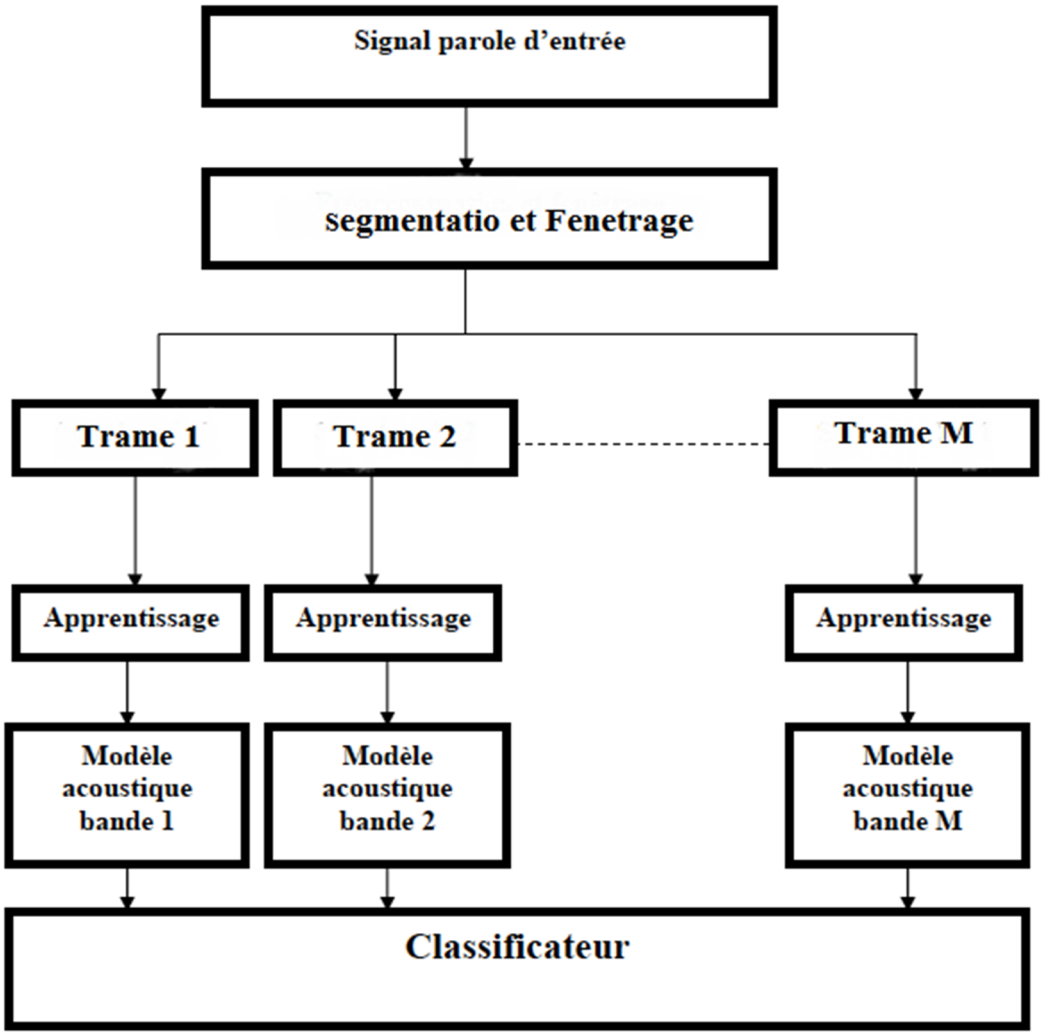


Figure 3.5 : Extraction des paramètres par banc de filtre (fusion au niveau du classificateur).

3.6. L'Analyse Cepstrale :

La prédiction linéaire emploie le modèle tout-pôle du spectre. Une méthode alternative au **LPC** est la prétendue analyse cepstrale. Dans l'analyse cepstrale, le spectre d'amplitude est représenté comme combinaison des fonctions de base cosinus avec des fréquences variables. Les coefficients cepstraux sont les amplitudes des fonctions de base. La figure 3.6 montre une comparaison de l'évaluation de l'enveloppe spectrale en utilisant le **LPC** et la représentation cepstrale.

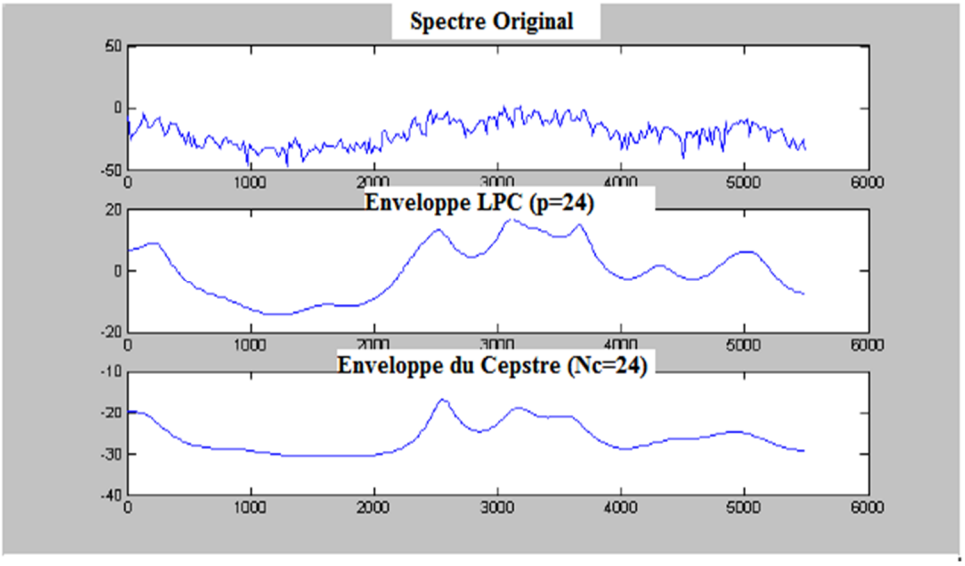


Figure 3.6 : Exemple d'estimation d'enveloppe spectrale par le LPC et le Cepstre de la FFT.

Notons que les crêtes dans le modèle du **LPC** sont très claires, tandis que le cepstre présente une enveloppe plus lisse. Dans ce sens, le modèle de **LPC** préserve plus de détails au sujet du spectre avec le même nombre de coefficients.

Formellement, le vrai cepstre du signal numérique $s[n]$ est défini comme l'inverse de la Transformée de Fourier du logarithme du spectre d'amplitude :

$$C[n]=\frac{1}{2\pi}\int_{-\pi}^{\pi} Cs(\omega)e^{j\omega\pi}d\omega \qquad (3.11)$$

Ou $C_s= \log|S(e^{j\omega})|$ dénote le logarithme du spectre d'amplitude, les coefficients $c[n]$ sont les coefficients de série de Fourier du Log Spectre. Le Log spectre est représenté comme une infinité d'addition des cosinus de différentes fréquences, et les coefficients cepstraux sont les amplitudes des fonctions de base. Les petits coefficients cepstraux représentent les changements lents du spectre, et les coefficients plus élevés représentent les composants rapidement variables du spectre.

Dans les sons voisés, il y a des composants périodiques dans le spectre d'amplitude, la structure fine harmonique résulte de la vibration des cordes vocales. Les variations lentes résultent du filtrage du conduit vocal, et de la descente spectrale de la source.

Similaire à l'analyse de **LPC**, l'augmentation du nombre de coefficients a comme conséquence plus de détails. La raison de prendre le logarithme du spectre peut être expliquée comme suit. Selon la théorie du filtre source :

$$\left| S\left(e^{j\omega}\right) \right| = \left| U\left(e^{j\omega}\right) \right| . \left| H\left(e^{j\omega}\right) \right| \quad (3.12)$$

Où *S, U* et *H* correspondent au son articulé, à la source et au filtre, respectivement. En prenant le logarithme, les composants multiplicatifs sont convertis dans des composants additifs:

$$\log\left| S\left(e^{j\omega}\right) \right| = \log\left| U\left(e^{j\omega}\right) \right| + \log\left| H\left(e^{j\omega}\right) \right| \quad (3.13)$$

Prendre le logarithme correspond à exécuter une transformation homomorphique, les séquences multiplicatives sont converties en un nouveau domaine, où ils deviennent additifs.

La formule pratique pour calculer le cepstre réel est obtenue par l'emploie de la **DFT** et la **IDFT** :

$$c\left[n\right] = f^{1}\left\{\log\left| f\left\{\text{segment signal}\right\}\right|\right\} \quad (3.14)$$

Donc le cepstre réel est obtenu en appliquant la **DFT** inverse au logarithme de l'amplitude de la **DFT** d'un segment du signal parole à analyser.

3.7. Coefficients MFCC :

La méthode **MFCC** est une méthode d'extraction des paramètres selon l'échelle de Mel.

En effet, la perception de la parole par le système auditif humain est fondée sur une échelle fréquentielle semblable à l'échelle de Mel. Cette échelle est linéaire aux basses fréquences et logarithmique en hautes fréquences.

Le signal acoustique contient de différentes sortes de renseignements sur le locuteur. La paramétrisation **MFCC** est basé sur la perception humaine de son. Les renseignements portés par les composantes de la fréquence basse du signal de parole sont plus importants phonétiquement pour les humains que les composantes à haute fréquence. La fréquence perceptive humaine est représentée dans l'échelle de Mel, qui est l'espacement de fréquence linéaire au-dessous de 1000 Hz, et l'espacement logarithmique au-dessus de 1000 Hz. On suppose que la perception humaine de son se compose du banc de filtres. Chaque filtre a une forme triangulaire. Les bancs de filtre triangulaires dans l'échelle Mel sont espacés uniformément.

Après études sur l'oreille humaine, il a été montré que l'homme se base sur une échelle fréquentielle spécifique. La formule de transfert est :

$$Mel(f) = 2595 * \log 10\left(1 + \frac{f}{700}\right) \quad (3.15)$$

f_{low} et f_{high} sont les limites basses et à haute fréquence de banc de filtre, donnés par :

Où : $f_{low}=f_s/N$ (3.16)

Et N est la longueur de trame de 160 échantillons (corresponds à 20ms).

$f_{high}=f_s/2$ (3.17)

Où: $f_{low}=100Hz$ et $f_{high}=8000Hz$

Quand la fréquence réelle f est au-dessous de 1000Hz, le rapport est linéaire ; cependant, le rapport des traits devient logarithmique quand f est au-dessus de 1000Hz, la bande passante du filtre individuel (le banc de filtres) augmente logarithmiquement dans l'échelle normale.

Chaque filtre triangulaire a une longueur de 1000 (Arbitrairement choisit) dans le domaine de fréquence, notez aussi que le 1000 ème échantillon correspond à : $fs/2$.

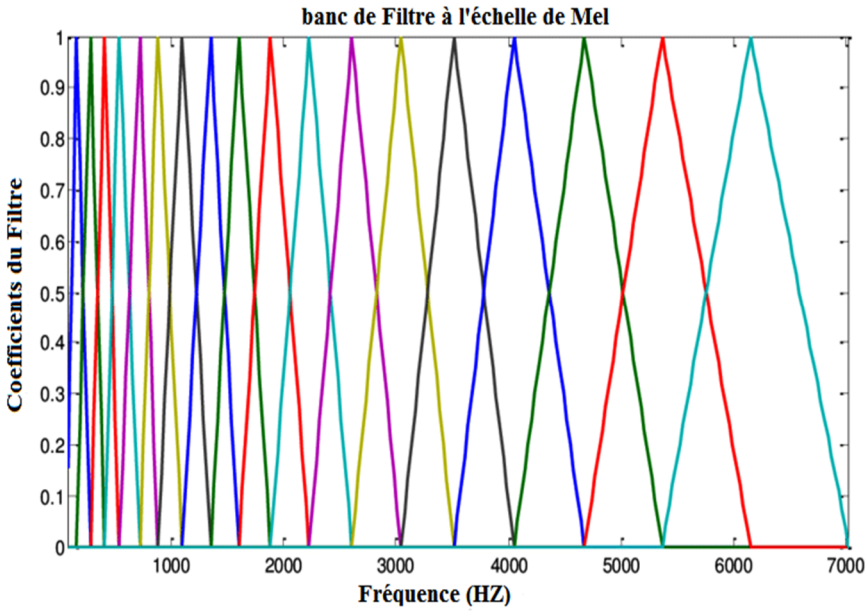


Figure 3.7 : Banc de filtres à l'échelle de Fréquence Mel.

Une fois le spectre Mel a été calculé, il doit être converti en arrière de l'intervalle de temps en utilisant la **DCT**. On appelle le résultat des fréquences Mel par les coefficients cepstraux. En utilisant la même procédure, un ensemble de fréquences Mel est calculé pour chaque trame de signal parole d'environ 20 millisecondes avec un chevauchement. Le calcul de **MFCC** est montré dans la Figure 3.8.

Le calcul des MFCCs se décompose en cinq phases :

1. Découper le signal en plusieurs fenêtres avec chevauchement.

2. Afin de diminuer la distorsion spectrale souvent on applique une fenêtre de Hamming :

$$W(n) = 0.54\text{-}0.46*\cos((2\pi\, n)/(N-1)) \qquad (3.18)$$

Cette fonction est multipliée par le signal à transformer, nous minimisons ainsi la distorsion spectrale créée par le chevauchement (le recouvrement).

3. Pour appliquer la **FFT** aux fenêtres ; On passe à l'échelle de Mel (équation 3.15).

4. Pour simuler l'oreille humaine, faut passer par un banc de Filtres, un filtre pour chaque fréquence que l'on cherche. Ces filtres ont une réponse de bande passante triangulaire. Pour connaître l'intervalle entre chaque filtre, on utilise une constante : l'intervalle de fréquence de Mel.

5. Convertissons de spectre logarithmique de Mel en temps comme suit :

$$Y(k) = \sum_{n=1}^{N} \boldsymbol{w(n)x(n)}\cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right) \quad \text{avec } k=1,\ldots,N \qquad (3.19)$$

$$\text{Où : } w(n) = \sqrt{\frac{1}{N}} \quad \text{si } n=1 \text{ et } w(n) = \sqrt{\frac{2}{N}} \text{ si non} \qquad (3.20)$$

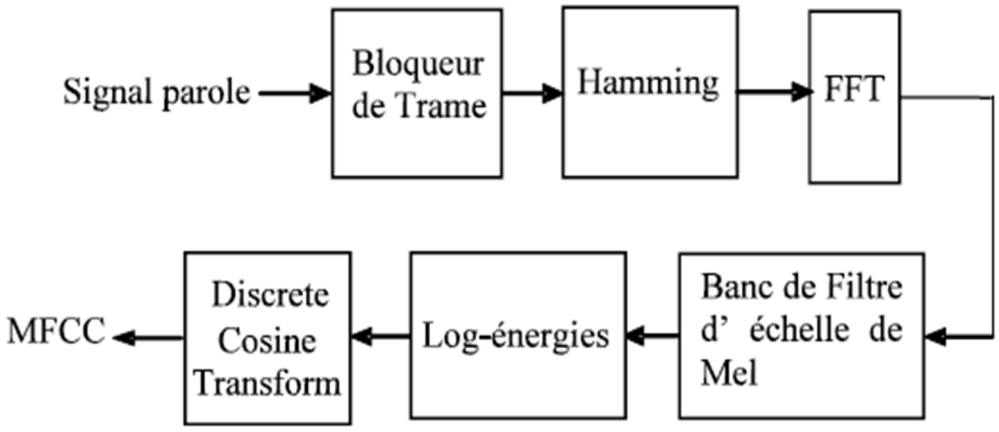


Figure 3.8 : calcul des coefficients MFCCs avec une échelle Mel.

3.8. Spectrogramme :

Le spectrogramme permet de mettre en évidence les différentes composantes fréquentielles du signal à un instant donné. L'amplitude du spectre apparaît sous la forme de niveaux en gris dans un diagramme en deux axes : temps et fréquence. Ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants.

3.9. Classification du modèle vectoriel :

Le but de la classification automatique est de déterminer automatiquement l'identité des données non classifiées. La classification automatique de modèle prend place en deux étapes, apprentissage et classification du modèle. La figure 3.9 donne une vue graphique globale du processus.

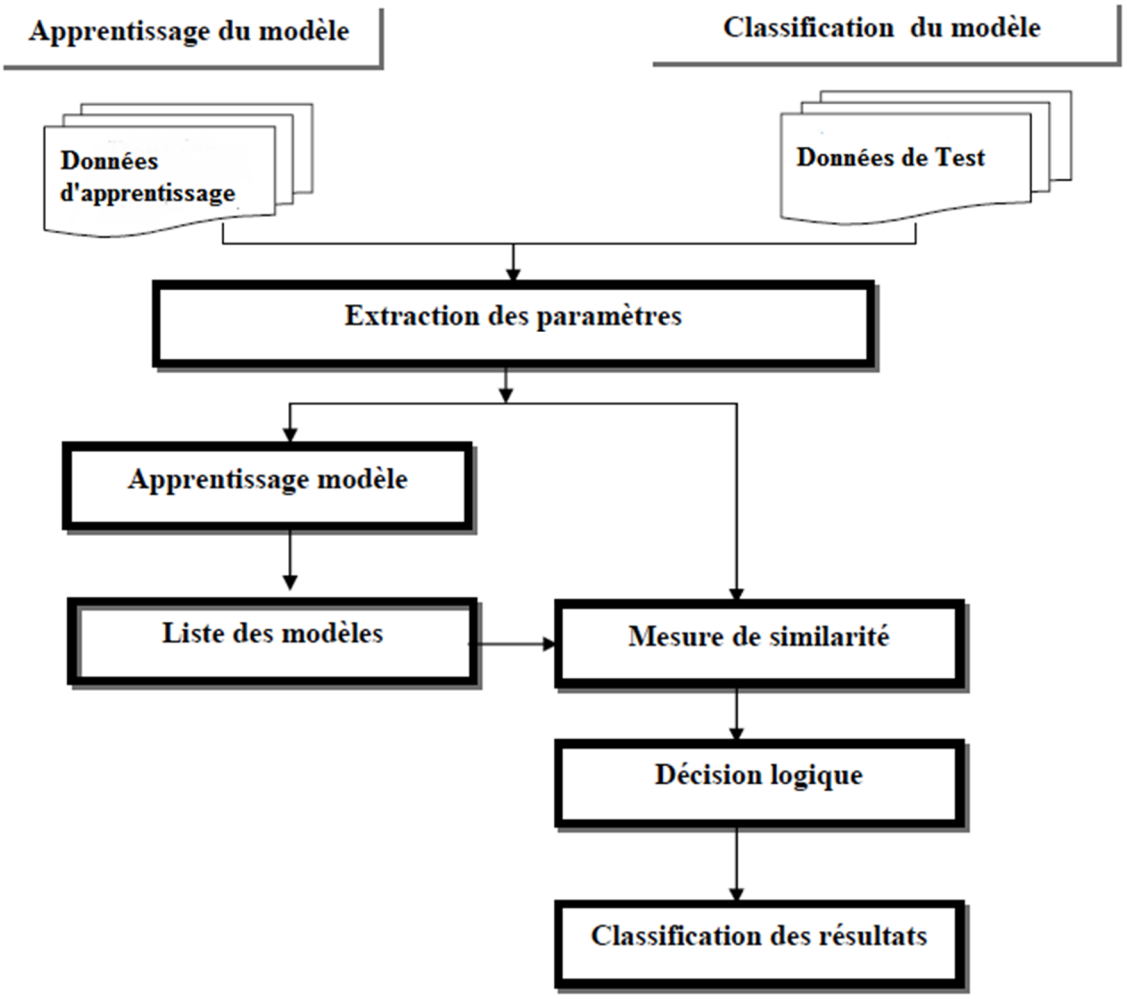


Figure 3.9 : Un graphique globale d'un processus de classification automatique.

3.9. Approche vectorielle :

Dans l'approche vectorielle, les vecteurs paramétriques d'apprentissage et de test sont comparés, sous l'hypothèse que les vecteurs d'une des séquences sont une réalisation imparfaite des vecteurs de l'autre séquence. La distorsion entre les deux séquences représente leur degré de similarité. Cette approche comporte deux grandes techniquese :

1. **La Déformation Temporelle Dynamique.**
2. **La Quantification Vectorielle**.

Qui ont été respectivement proposés pour les applications dépendantes et indépendantes du texte.

Le **DTW** aligne temporellement les suites d'observations, tandis que la VQ représente le locuteur par un dictionnaire.

La figure 3.14 montre un diagramme conceptuel pour illustrer le processus de reconnaissance. Sur cette figure, seules deux voix et deux dimensions de l'espace acoustique sont représentées.

Les cercles se réfèrent aux vecteurs acoustiques de la voix 1 tandis que les triangles proviennent de la voix 2.

Dans la phase d'apprentissage, un dictionnaire **VQ** spécifique au locuteur est généré pour chaque locuteur connu en regroupant ses vecteurs acoustiques d'entraînement, Le centroïde résultant est représenté par des cercles noirs et des triangles noirs pour les voix 1 et 2, respectivement.

La distance entre un vecteur et le centroïde le plus proche d'un dictionnaire est appelée « distorsion VQ ».

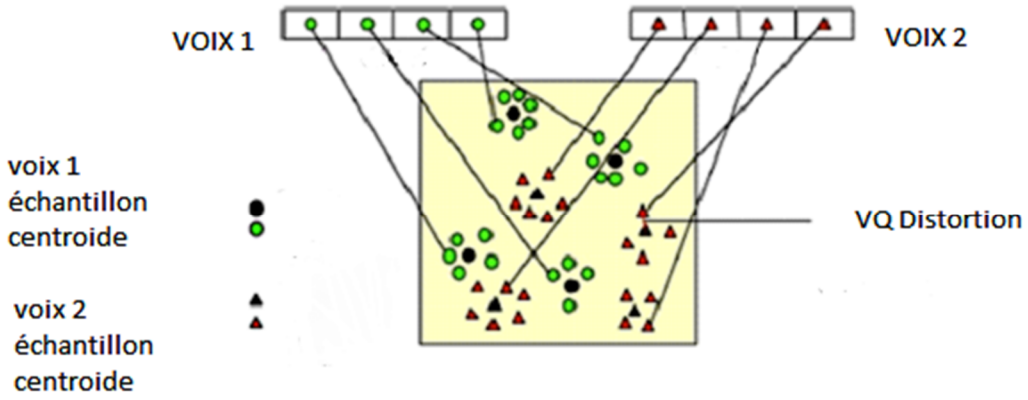


Figure 3.10 : Diagramme conceptuel illustrant la formation d'un dictionnaire de quantification vectorielle.

3.9.1. La Déformation Temporelle Dynamique :

D'une façon générale, le **DTW** est une méthode qui recherche un appariement optimal entre deux séries temporelles. L'alignement temporel, plus connu sous l'acronyme de **DTW**, est une méthode fondée sur un principe de comparaison d'un signal à analyser avec un ensemble de signaux stockés dans une base de référence. Le signal à analyser est comparé avec chacune

des références, et il est classé en fonction de sa proximité avec l'une des références stockées. Le **DTW** est en fait une application au domaine de la reconnaissance de la parole. Cette méthode d'alignement de séries temporelles est souvent utilisée dans le contexte de modèles de Markov cachés.

3.9.2. La Quantification Vectorielle :

La quantification vectorielle décompose l'espace acoustique d'un locuteur donné X, en un ensemble de M sous-espaces représentés par leur vecteurs centroides $C = \{c_1, c_2,\ldots,c_M\}$.

Ces vecteurs centroides forment un dictionnaire (de taille M) qui modélise ce locuteur, et ils sont calculés en minimisant l'erreur de quantification moyenne (distorsion) induite par le dictionnaire sur les données d'apprentissage du locuteur $\{x_1, x_2,\ldots,x_M\}$.

$$D(X,C)=\frac{1}{T}\sum_{t=1}^{T} \min d(xt, cm) \qquad 1\leq m \leq M \qquad (3.21)$$

Où $d(x_t , c_m)$ est une mesure de distance au sens d'une certaine métrique liée à la paramétrisation.

L'apprentissage vise à réduire l'erreur de quantification. On peut mieux représenter le locuteur en augmentant la taille du dictionnaire, mais le système sera moins rapide et plus demandeur de mémoire. Il faut donc trouver un bon compromis.

3.10. Conclusion :

Dans ce chapitre on a vu quelques outils utilisés dans le traitement des signaux vocaux (acquisition, filtrage, segmentation, transformation, classification etc.). Ces outilles sont nécessaire dans le prétraitement et l'analyse du signale qui sont des étapes essentiel dans un système **RAL**. Aussi on a donné une idée globale sur les principes utilisés et les opérations de base dans l'extraction des paramètres, comment ces paramètres sont utilisés dans un système **RAL** et les obstacles techniques qu'on peut rencontrer.

Chapitre IV : Implémentation et mise en œuvre de la méthode

4.1. Introduction :

Dans ce projet, nous expérimenterons la construction et l'essai d'un système de reconnaissance vocale automatique.

Pour mettre en œuvre un tel système, il faut passer par plusieurs étapes, qui ont été décrites en détail dans les sections précédentes. Sans oubliant de noter que la plupart des tâches ci-dessous sont implémentées dans Matlab.

4.2. Principes de reconnaissance des locuteurs / voix :

La reconnaissance des locuteurs peut être classée en identification et vérification. L'identification du locuteur / voix est le processus permettant de déterminer quel locuteur enregistré fournit un énoncé donné. D'autre part, la vérification est le processus d'acceptation ou de rejet de la revendication d'identité d'un locuteur.

Au plus haut niveau, tous les systèmes de reconnaissance de locuteur contiennent deux modules principaux (voir Figure 4.1 et 4.2) :

1. Extraction et appariement des caractéristiques : C'est le processus qui extrait une petite quantité de données du signal vocal, et qui peut ensuite être utilisé pour représenter chaque locuteur.

2. La correspondance des caractéristiques : Elle implique la procédure actuelle pour identifier le locuteur inconnu en comparant les caractéristiques extraites de son signal vocal avec celles d'un ensemble de locuteurs connus.

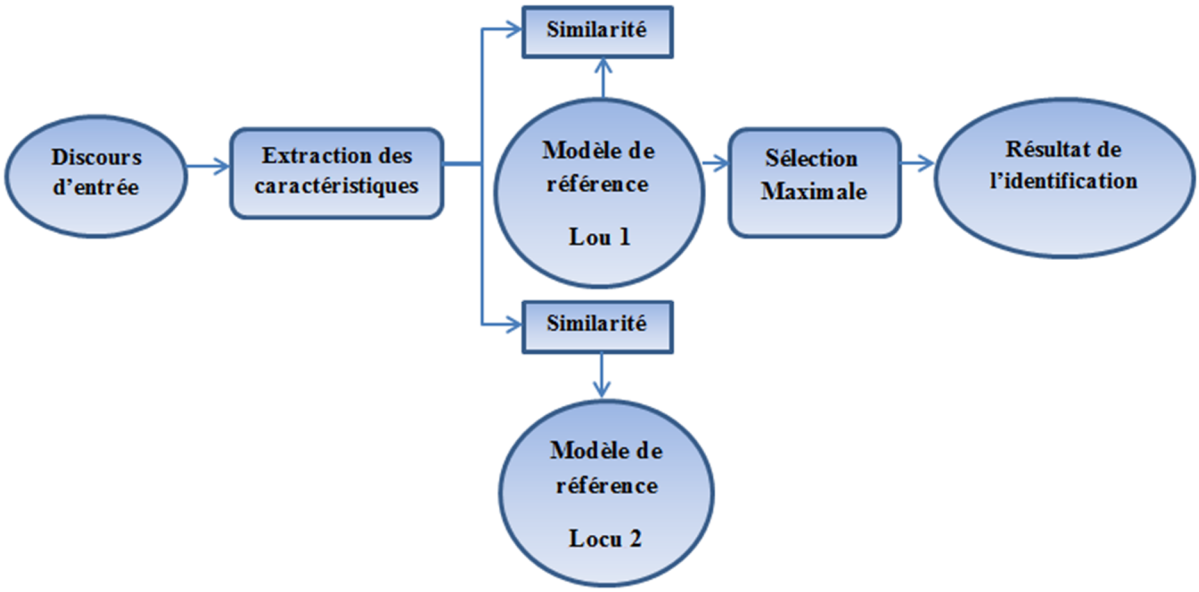


Figure 4.1 : Structures de base des systèmes de reconnaissance de locuteur/Identification.

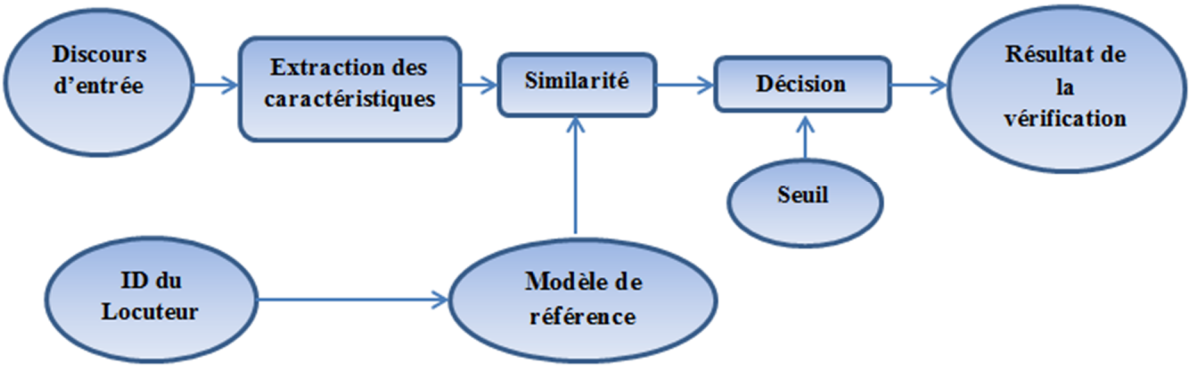


Figure 4.2 : Structures de base des systèmes de reconnaissance de locuteur/vérification.

Tous les systèmes de reconnaissance de locuteur doivent contenir deux phases distinctes :

La première est référée aux sessions d'inscription ou à la phase d'entraînement tandis que **la seconde** est appelée session d'opération ou phase de test.

Dans la phase d'entraînement, chaque locuteur enregistré doit fournir des échantillons de son discours, afin que le système puisse construire ou former un modèle de référence pour ce locuteur. En outre, dans le cas des systèmes de vérification du locuteur, un seuil spécifique au locuteur est également calculé à partir des échantillons d'apprentissage.

Pendant la phase de test le discours d'entrée est mis en correspondance avec le modèle de référence mémorisé, et donc la décision de reconnaissance est prise.

La reconnaissance des locuteurs est une tâche difficile et c'est toujours un domaine de recherche actif, elle fonctionne sur la base du principe selon lequel le discours d'une personne présente des caractéristiques propres au locuteur. Cependant, cette tâche a été remise en cause par la très grande diversité des signaux vocaux d'entrée. La principale source de variance vient des locuteurs eux-mêmes.

Les signaux vocaux dans les séances d'entraînement et de test peuvent être très différents en raison de nombreux facteurs tels que : les changements de la voix, les conditions de santé (par exemple, Le rhume), les taux d'élocution, etc.
Il existe également d'autres facteurs, au-delà de la variabilité des locuteurs, qui constituent un défi pour la technologie de reconnaissance des locuteurs.

Des exemples de ceux-ci sont le bruit acoustique et les variations dans les environnements d'enregistrement (par exemple, le locuteur utilise différents combinés téléphoniques / microphones).

4.3. Extraction des caractéristiques de la parole :

Le but de ce module est de convertir la forme d'onde de la parole en un type de représentation paramétrique (à un débit d'information considérablement plus faible) pour une analyse et un traitement plus poussés. Ceci est souvent appelé l'interface de traitement du signal.

La fonction **findvoice** permet de faire appel à la parole du locuteur (signal audio de format brut .wav), et de transformer ce signal du mode stéréo en mode mono et ensuite, elle éliminera le silence (pause) qui sera entre les mots en réduisant la durée de la parole.

```
function [normalized,slot]=findvoice(suono,fs,fsnew)
%[suono,fs] = wavread('train/s11.wav');
[dimx, dimy] = size(suono);
% conversion from stereo 2 mono
if dimy==2
suono = suono(:,1);
end
if dimx==2
suono = suono(1,:);
end
%figure,plot(suono)
% fsnew = 8000;
suono_ricampionato = resample(suono,fsnew,fs);
```

```
% elimino pause
sr = suono_ricampionato;
srdct = dct(sr);
srdct(1) = 0;
sr = idct(srdct);
%------------------------------------------------------------- first
output
normalized = sr;
%-------------------------------------------------------------------------
-
sr = abs(sr);
minimo = min(sr);
massimo = max(sr);
L = length(sr);
percentuale = 5;
valmin = minimo+(massimo-minimo)/100*percentuale;
% posinf = find(sr<valmin);
possup = find(sr>=valmin);
Lok = length(possup);
thresh = 100;
cont = 1;
slot(cont,1) = possup(1);
slot(cont,2) = possup(1);
for ii=2:Lok
dv = possup(ii)-possup(ii-1);
if dv<=thresh
slot(cont,2) = possup(ii);
else
cont = cont+1;
slot(cont,1) = possup(ii);
end
end
```

Le signal de parole est un signal variant lentement (appelé quasi-stationnaire).

Un exemple de signal vocal est représenté dans le code suivant et sur la figure 4.3, Lorsqu'il est examiné sur une période de temps suffisamment courte (entre 5 et 100 ms), ses caractéristiques sont relativement stationnaires.

Cependant, sur de longues périodes (de l'ordre de 1/5 seconde ou plus), les caractéristiques du signal changent pour refléter les différents sons vocaux parlés, Par conséquent, l'analyse spectrale à court terme est la façon la plus courante de caractériser le signal vocal

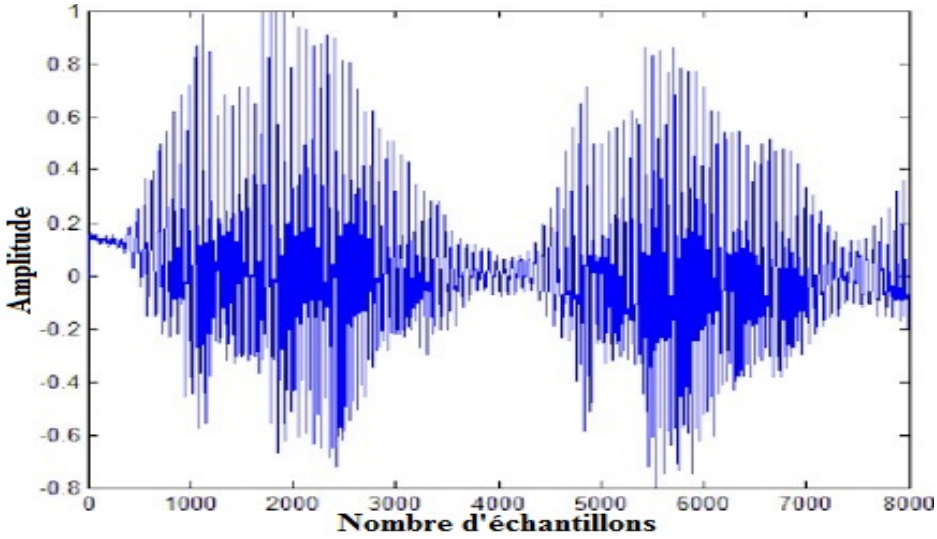


Figure 4.3 : Un exemple de signal vocal

Il existe un large éventail de possibilités de représenter les paramètres du signal de parole pour la tâche de reconnaissance du locuteur, tel que le codage de prédiction linéaire, les coefficients cepstraux de fréquence Mel et autres.

MFCC est peut-être le plus connu et le plus populaire, et ceux-ci seront utilisés dans ce projet.

Les **MFCCs** sont basés sur la variation connue des largeurs de bande critiques de l'oreille humaine avec la fréquence.

Les filtres espacés linéairement aux basses fréquences et logarithmiquement en hautes fréquences ont été utilisés pour capturer les caractéristiques phonétiquement importantes de la parole.

Ceci est exprimé dans l'échelle de fréquence Mel, qui est un espacement de fréquence linéaire inférieur à 1000 Hz et un espacement logarithmique supérieur à 1000 Hz. Le processus de calcul des **MFCCs** est décrit plus en détail ci-après.

4.4. Processeurs de coefficients cepstraux de fréquence Mel :

Un schéma qui montre le principe de la structure d'un processeur **MFCC** est donné dans la figure 4.4.

L'entrée vocale est typiquement enregistrée à une fréquence d'échantillonnage supérieure à 10 kHz, Cette fréquence d'échantillonnage a été choisie pour minimiser les effets de repliement dans la conversion analogique-numérique.

Ces signaux échantillonnés peuvent capturer toutes les fréquences jusqu'à 5 kHz, qui couvrent la plupart de l'énergie des sons générés par les humains. Comme cela a été discuté précédemment, l'objectif principal du processeur **MFCCs** est d'imiter le comportement des oreilles humaines. En outre, plutôt que les formes d'onde de la parole elles-mêmes, les **MFFCs** sont moins sensibles aux variations mentionnées.

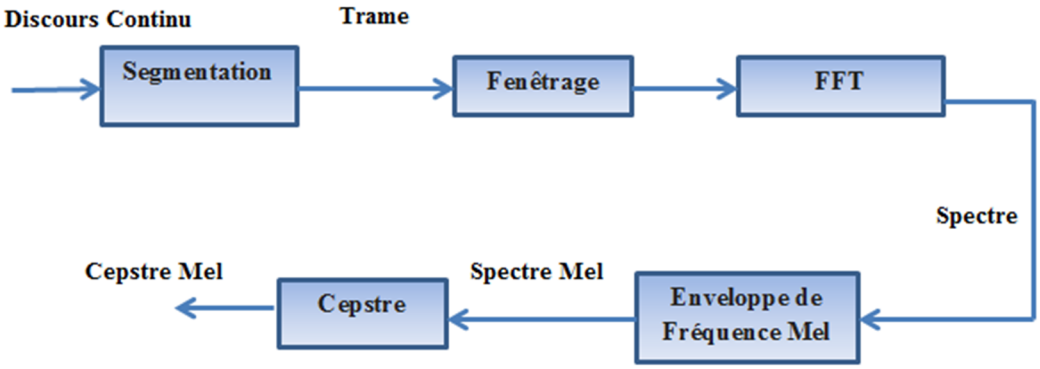


Figure 4.4 : Schéma de principe du processeur MFCC.

4.4.1. Segmentation de trame :

Dans cette étape, le signal vocal continu est segmenté dans des trames de N échantillons, les trames adjacentes étant séparées par M (M <N).

- La première trame est constituée des N premiers échantillons.
- La deuxième trame commence à partir de M échantillons après la première trame, et se chevauche par des échantillons N - M.
- De même, la troisième trame commence à partir de 2M échantillons après la première trame (ou à partir de M échantillons après la deuxième trame) et se chevauche par N - 2M échantillons.

Ce processus se poursuit jusqu'à ce que tout le discours soit pris en compte dans une ou plusieurs trames, Les valeurs typiques pour N et M sont N = 256 (ce qui équivaut à un fenêtrage d'environ 30 ms et facilite la **FFT** de Radix 2 rapide) et M = 100.

La fonction **enframe** coupe le signal vocal en trames avec chevauchement. Le résultat est une matrice où chaque colonne est une trame de N échantillons du signal de parole original.

```
function f=enframe(x,win,inc)
nx=length(x);
nwin=length(win);
if (nwin == 1)
   len = win;
else
   len = nwin;
end
if (nargin < 3)
   inc = len;
end
nf = fix((nx-len+inc)/inc);
f=zeros(nf,len);
indf= inc*(0:(nf-1)).';
inds = (1:len);
f(:) = x(indf(:,ones(1,len))+inds(ones(nf,1),:));
if (nwin > 1)
    w = win(:)';
    f = f .* w(ones(nf,1),:);
end
```

4.4.2. Le fenêtrage :

Cette étape de traitement consiste à fenêtrer chaque trame individuelle de manière à minimiser les discontinuités du signal au début et à la fin de chaque trame. Le concept ici est de minimiser la distorsion spectrale en utilisant la fenêtre pour la réduire à zéro au début et à la fin de chaque trame.

Si nous définissons, où N est le nombre d'échantillons pour chaque $N \leq n$ la fenêtre comme W(n), alors le résultat de la fenêtre est le signal.

$$Y_l(n)=X_l(n)W(n),\ 0<=n<=N-1 \qquad (4.1)$$

Typiquement la fenêtre de Hamming est utilisée, qui a la forme ci-dessous :

$$W(n)=0{,}54 - 0{,}46 \cos (2\pi\ n/N),\ 0<=n<=N-1 \qquad (4.2)$$

4.4.3. La Transformée de Fourier rapide :

Cette étape de traitement est la transformée de Fourier Rapide, qui convertit chaque trame de N échantillons du domaine temporel dans le domaine fréquen1tiel, La **FFT** est un algorithme

rapide pour implémenter la Transformée de Fourier Discrète qui est définie sur l'ensemble des N échantillons $\{X_n\}$.

Le résultat obtenu après cette étape est souvent appelé 'Spectre' ou 'Périodogramme du signal'.

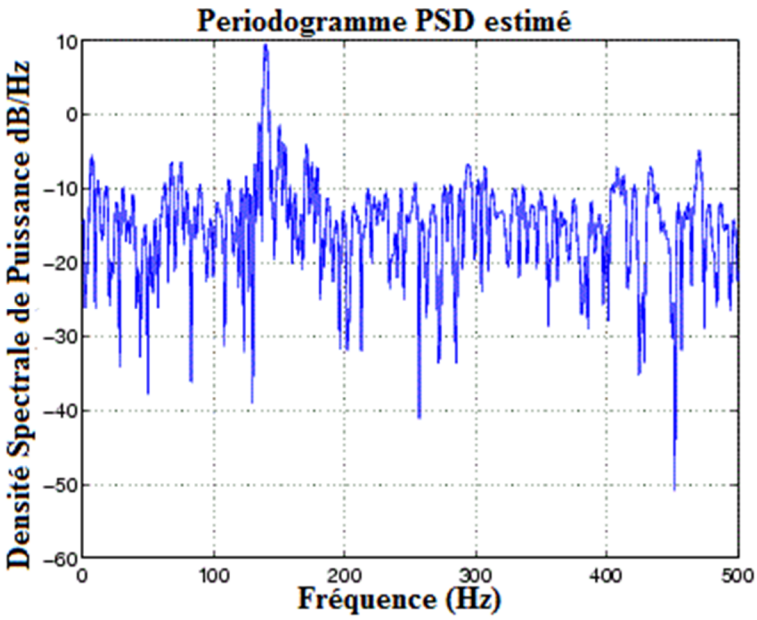


Figure 4.5 : Densité spectrale de puissance du périodogramme.

La fonction **rfft** Applique les étapes du fenêtrage et la **FFT** pour transformer le signal dans le domaine fréquentiel; ce processus est utilisé dans de nombreuses applications différentes et référé dans la littérature sous le nom de transformée de Fourier fenêtrée ou de transformée de Fourier à court terme. Le résultat est souvent appelé le spectre ou le périodogramme.

```
function y=rfft(x,n,d)
s=size(x);
if prod(s)==1
    y=x
else
    if nargin <3
        d=find(s>1);
        d=d(1);
        if nargin<2
            n=s(d);
        end
    end
    if isempty(n)
```

```
        n=s(d);
    end
    y=fft(x,n,d);
    y=reshape(y,prod(s(1:d-1)),n,prod(s(d+1:end)));
    s(d)=1+fix(n/2);
    y(:,s(d)+1:end,:)=[];
    y=reshape(y,s);
end
```

4.4.4. L'Enveloppe de Fréquence Mel :

Des études psychophysiques ont montré que la perception humaine du contenu fréquentiel des sons pour les signaux vocaux ne suit pas une échelle linéaire. Ainsi, pour chaque tonalité avec une fréquence réelle, f, mesurée en Hz, un pas subjectif est mesuré sur une échelle appelée « L'échelle Mel ».

L'échelle de fréquence Mel est un espacement de fréquence linéaire inférieur à 1000 Hz et un espacement logarithmique supérieur à 1000 Hz, Comme point de référence, la hauteur d'une tonalité de 1 kHz, 40 dB au-dessus du seuil d'audition perceptuelle, est définie comme 1000 Mel.

Par conséquent, nous pouvons utiliser la formule approximative suivante pour calculer le Mel pour une fréquence donnée f en Hz:

$$\text{Mel}(f) = 2595 * \log10(1 + f / 700) \qquad (4.3)$$

Une approche pour simuler le spectre subjectif consiste à utiliser un banc de filtres, un filtre pour chaque composante de fréquence Mel souhaitée (voir la figure 4.5), ce banc de filtres a une réponse en fréquence de bande passante triangulaire, et l'espacement ainsi que la bande passante sont déterminés par un intervalle de fréquence Mel constant.

Notez que ce banc de filtres est appliqué dans le domaine fréquentiel. Par conséquent, cela revient simplement à prendre ces fenêtres en forme de triangle sur la figure 4.5 du spectre.
La fonction **melbankm** applique ce banc de filtres d'enveloppe Mel et faite voir chaque filtre comme un binogramme d'histogramme (où les bins se chevauchent) dans le domaine fréquentiel.

```
function [x,mn,mx]=melbankm(p,n,fs,fl,fh,w)


if nargin < 6
```

```
  w='tz';
  if nargin < 5
    fh=0.5;
    if nargin < 4
      fl=0;
    end
  end
end
f0=700/fs;
fn2=floor(n/2);
lr=log((f0+fh)/(f0+fl))/(p+1);
% convert to fft bin numbers with 0 for DC term
bl=n*((f0+fl)*exp([0 1 p p+1]*lr)-f0);
b2=ceil(bl(2));
b3=floor(bl(3));
if any(w=='y')
  pf=log((f0+(b2:b3)/n)/(f0+fl))/lr;
  fp=floor(pf);
  r=[ones(1,b2) fp fp+1 p*ones(1,fn2-b3)];
  c=[1:b3+1 b2+1:fn2+1];
  v=2*[0.5 ones(1,b2-1) 1-pf+fp pf-fp ones(1,fn2-b3-1) 0.5];
  mn=1;
  mx=fn2+1;
else
  b1=floor(bl(1))+1;
  b4=min(fn2,ceil(bl(4)))-1;
  pf=log((f0+(b1:b4)/n)/(f0+fl))/lr;
  fp=floor(pf);
  pm=pf-fp;
  k2=b2-b1+1;
  k3=b3-b1+1;
  k4=b4-b1+1;
  r=[fp(k2:k4) 1+fp(1:k3)];
  c=[k2:k4 1:k3];
  v=2*[1-pm(k2:k4) pm(1:k3)];
  mn=b1+1;
  mx=b4+1;
end
if any(w=='n')
  v=1-cos(v*pi/2);
elseif any(w=='m')
```

```
  v=1-0.92/1.08*cos(v*pi/2);
end
if nargout > 1
  x=sparse(r,c,v);
else
  x=sparse(r,c+mn-1,v,p,1+fn2);
end
```

4.4.5. Cepstre :

Dans cette dernière étape, nous convertissons le logarithme du spectre Mel dans le temps, Le résultat est appelé les Coefficients Cepstraux de Fréquence Mel, La représentation cepstrale du spectre de la parole fournit une bonne représentation des propriétés spectrales locales du signal pour l'analyse de trame donnée.

La fonction **melcepst** fait la conversion du spectre de puissance en coefficient MFCC.

```
function c=melcepst(s,fs,w,nc,p,n,inc,fl,fh)
if nargin<2 fs=11025; end
if nargin<3 w='M'; end
if nargin<4 nc=12; end
if nargin<5 p=floor(3*log(fs)); end
if nargin<6 n=pow2(floor(log2(0.03*fs))); end
if nargin<9
   fh=0.5;
   if nargin<8
     fl=0;
     if nargin<7
        inc=floor(n/2);
     end
  end
end

if length(w)==0
   w='M';
end
if any(w=='R')
   z=enframe(s,n,inc);
elseif any (w=='N')
   z=enframe(s,hanning(n),inc);
else
   z=enframe(s,hamming(n),inc);
```

```
end
f=rfft(z.');
[m,a,b]=melbankm(p,n,fs,fl,fh,w);
pw=f(a:b,:).*conj(f(a:b,:));
pth=max(pw(:))*1E-6;
if any(w=='p')
   y=log(max(m*pw,pth));
else
   ath=sqrt(pth);
   y=log(max(m*abs(f(a:b,:)),ath));
end
c=rdct(y).';
nf=size(c,1);
nc=nc+1;
if p>nc
   c(:,nc+1:end)=[];
elseif p<nc
   c=[c zeros(nf,nc-p)];
end
if ~any(w=='0')
   c(:,1)=[];
   nc=nc-1;
end
if any(w=='e')
   c=[log(sum(pw)).' c];
   nc=nc+1;
end

% calculate derivative

if any(w=='D')
  vf=(4:-1:-4)/60;
  af=(1:-1:-1)/2;
  ww=ones(5,1);
  cx=[c(ww,:); c; c(nf*ww,:)];
  vx=reshape(filter(vf,1,cx(:)),nf+10,nc);
  vx(1:8,:)=[];
  ax=reshape(filter(af,1,vx(:)),nf+2,nc);
  ax(1:2,:)=[];
  vx([1 nf+2],:)=[];
```

```
  if any(w=='d')
     c=[c vx ax];
  else
     c=[c ax];
  end
elseif any(w=='d')
  vf=(4:-1:-4)/60;
  ww=ones(4,1);
  cx=[c(ww,:); c; c(nf*ww,:)];
  vx=reshape(filter(vf,1,cx(:)),nf+8,nc);
  vx(1:8,:)=[];
  c=[c vx];
end

if nargout<1
   [nf,nc]=size(c);
   t=((0:nf-1)*inc+(n-1)/2)/fs;
   ci=(1:nc)-any(w=='0')-any(w=='e');
   imh = imagesc(t,ci,c.');
   axis('xy');
   xlabel('Time (s)');
   ylabel('Mel-cepstrum coefficient');
    map = (0:63)'/63;
    colormap([map map map]);
    colorbar;
end
```

Comme les coefficients du spectre Mel (et donc leur logarithme) sont des nombres réels, nous pouvons les convertir dans le domaine temporel en utilisant la transformée en cosinus discrète

```
function y=rdct(x,n,a,b)
fl=size(x,1)==1;
if fl x=x(:); end
[m,k]=size(x);
if nargin<2 n=m;
end
if nargin<4 b=1;
    if nargin<3 a=sqrt(2*n);
    end
```

```
    end
if n>m x=[x; zeros(n-m,k)];
elseif n<m x(n+1:m,:)=[];
end


x=[x(1:2:n,:); x(2*fix(n/2):-2:2,:)];
z=[sqrt(2) 2*exp((-0.5i*pi/n)*(1:n-1))].';
y=real(fft(x).*z(:,ones(1,k)))/a;
y(1,:)=y(1,:)*b;
if fl y=y.'; end
```

4.4.6. Résultats :

En appliquant la procédure décrite ci-dessus, pour chaque trame vocale d'environ 30 msec avec chevauchement, un ensemble de Coefficients Cepstraux de Fréquence Mel est calculé.

C'est le résultat d'une transformée en cosinus du logarithme du spectre de puissance à court terme exprimé sur une échelle de fréquence Mel.
Cet ensemble de coefficients est appelé un vecteur acoustique, Par conséquent, chaque énoncé d'entrée est transformé en une séquence de vecteurs acoustiques.

Dans la section suivante, nous verrons comment ces vecteurs acoustiques peuvent être utilisés pour représenter et reconnaître les caractéristiques vocales du locuteur.

La fonction **findfeatures** transforme les coefficients **MFCCs** des signaux de la parole en vecteurs acoustiques.

```
function [out] = findfeatures(ingresso,fs)
suono              = double(ingresso);
fsnew              = 8000;
[normalized,slot] = findvoice(suono,fs,fsnew);
C = [];
L = size(slot,1);
for ii=1:L
    if slot(ii,2)-slot(ii,1)>=128
        s = normalized(slot(ii,1):slot(ii,2));
        c = melcepst(s,fsnew,[],40);
        C = [C;c];
    end
end
out = C;
```

4.5. Correspondance des caractéristiques :

Le problème de la reconnaissance du locuteur appartient à un sujet beaucoup plus large dans le domaine de la reconnaissance des formes scientifiques et techniques, Le but de la reconnaissance de formes est de classer les objets d'intérêt dans l'une des catégories ou classes.

Les objets d'intérêt sont génériquement appelés modèles et dans notre cas, ce sont des séquences de vecteurs acoustiques qui sont extraites d'un discours d'entrée en utilisant les techniques décrites dans la section précédente.

Les classes ici se réfèrent à des locuteurs individuels, Puisque la procédure de classification dans notre cas est appliquée aux entités extraites, elle peut également être appelée correspondance de caractéristiques.

De plus, s'il existe un ensemble de modèles dont les classes individuelles sont déjà connues, alors on a un problème de reconnaissance de modèle supervisé. C'est exactement notre cas puisque pendant la séance d'entraînement, nous étiquetons chaque voix d'entrée avec **l'ID** (s_1 à s_n).

Ces modèles comprennent l'ensemble d'apprentissage et sont utilisés pour dériver un algorithme de classification.

Les modèles restants sont ensuite utilisés pour tester l'algorithme de classification ; ces modèles sont collectivement appelés l'ensemble de test. Si les classes correctes des modèles individuels dans l'ensemble de test sont également connues, alors on peut évaluer la performance de l'algorithme.

L'état de l'art dans les techniques d'appariement de caractéristiques utilisées dans la reconnaissance de locuteur inclut le **DTW**, **HMM**, et la **VQ**.

Dans ce projet, l'approche VQ sera utilisée, en raison de la facilité de mise en œuvre et de la grande précision de cette approche, VQ est un processus de mappage de vecteurs d'un grand espace vectoriel à un nombre fini de régions dans cet espace. Chaque région est appelée un «Cluster» et peut être représentée par son centre appelé «un Centroïde», La collection de tous les centroïdes est appelée «un dictionnaire».

La fonction **createnn** initie la technique de reconnaissance de formes basée sur la **VQ** en utilisant l'algorithme **LBG** pour construire des modèles de référence de locuteur à partir des vecteurs acoustiques durant la phase d'apprentissage.

```
function [net] = createnn(P,T)
alphabet = P;
targets  = T;

[R,Q]  = size(alphabet);
[S2,Q] = size(targets);
S1     = 100;
% traingda
net = newff(minmax(alphabet),[S1 S2],{'tansig' 'tansig'},'traingda');
net.LW{2,1}           = net.LW{2,1}*0.01;
net.b{2}              = net.b{2}*0.01;
net.performFcn        = 'mse';
net.trainParam.goal   = 0.000000001;
net.trainParam.show   = 100;
net.trainParam.epochs = 500000;
net.trainParam.mc     = 0.95;
net.trainParam.showWindow=0;
P                     = alphabet;
T                     = targets;
[net,tr]              = train(net,P,T);
```

Dans la phase de reconnaissance, un énoncé d'entrée d'une voix inconnue est "quantifié par vecteur" en utilisant chaque dictionnaire entraîné et la distorsion **VQ** totale est calculée.

Le locuteur correspondant au dictionnaire **VQ** avec la plus petite distorsion totale est identifié.

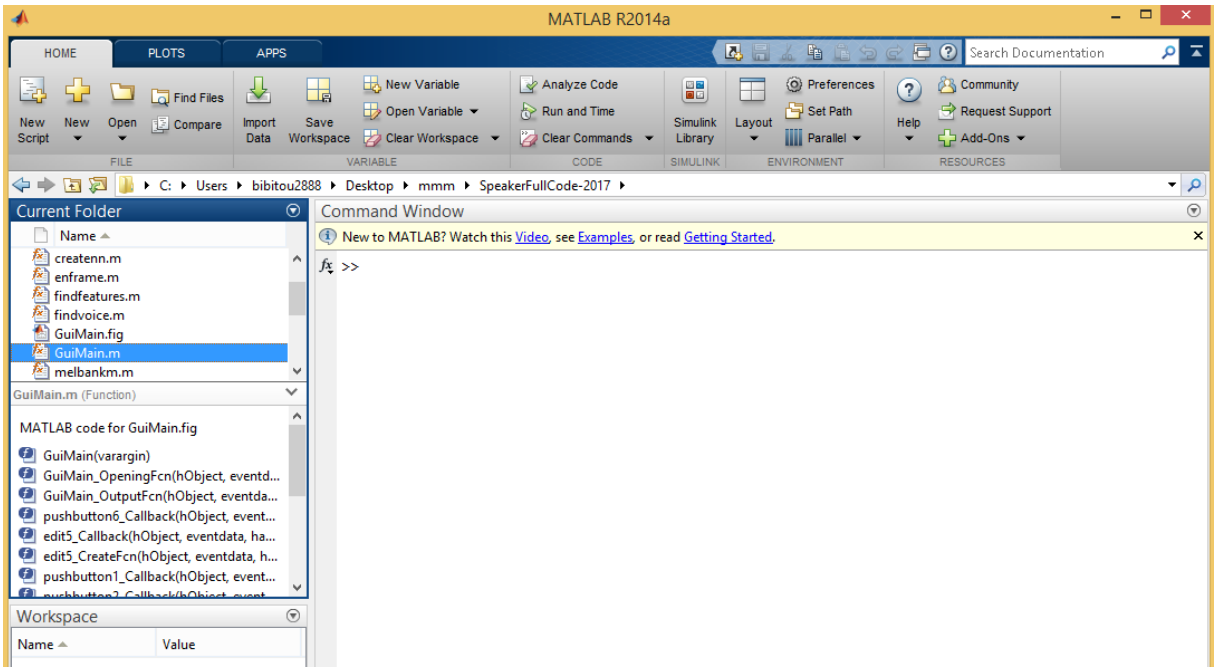
Nous pourrons identifier toute séquence de vecteurs acoustiques prononcés par des locuteurs inconnus grâce à la fonction **ann_face_matching.**

```
function [out] = ann_face_matching(features)
P = [];
T = [];
L = features_size;
for ii=1:L
    C = features_data{ii,1};
    [dimx,dimy] = size(C);
    P = [P C'];
```

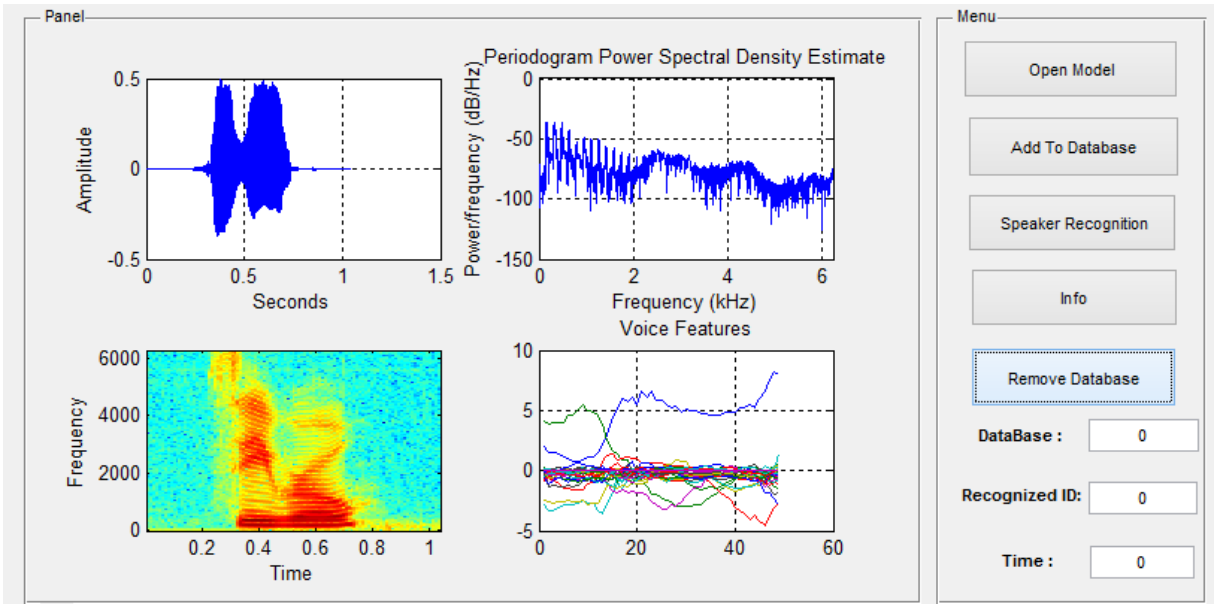
```
    t   = zeros(max_class-1,dimx);
    pos =  features_data{ii,2};
    for jj=1:(max_class-1)
        if jj==pos
            t(jj,:) = 1;
        else
            t(jj,:) = -1;
        end
    end
    T = [T t];
end

input_vector = features';
%Normalization
for ii=1:size(P,1)
    v = P(ii,:);
    v = v(:);
    bii = max([v;1]);
    aii = min([v;-1]);
    P(ii,:) = 2*(P(ii,:)-aii)/(bii-aii)-1;
    input_vector(ii,:) = 2*(input_vector(ii,:)-aii)/(bii-aii)-1;
end
[net]       = createnn(P,T);
risultato   = sim(net,input_vector);
[dimx,dimy] = size(risultato);
vettore     = zeros(dimy,1);
for jj=1:dimy
    c           = risultato(:,jj);
    [val,pos]   = max(c);
    vettore(pos) = vettore(pos)+1;
end
[val,pos] = max(vettore);
out       = pos;
```

4.6. Simulation et évaluation :

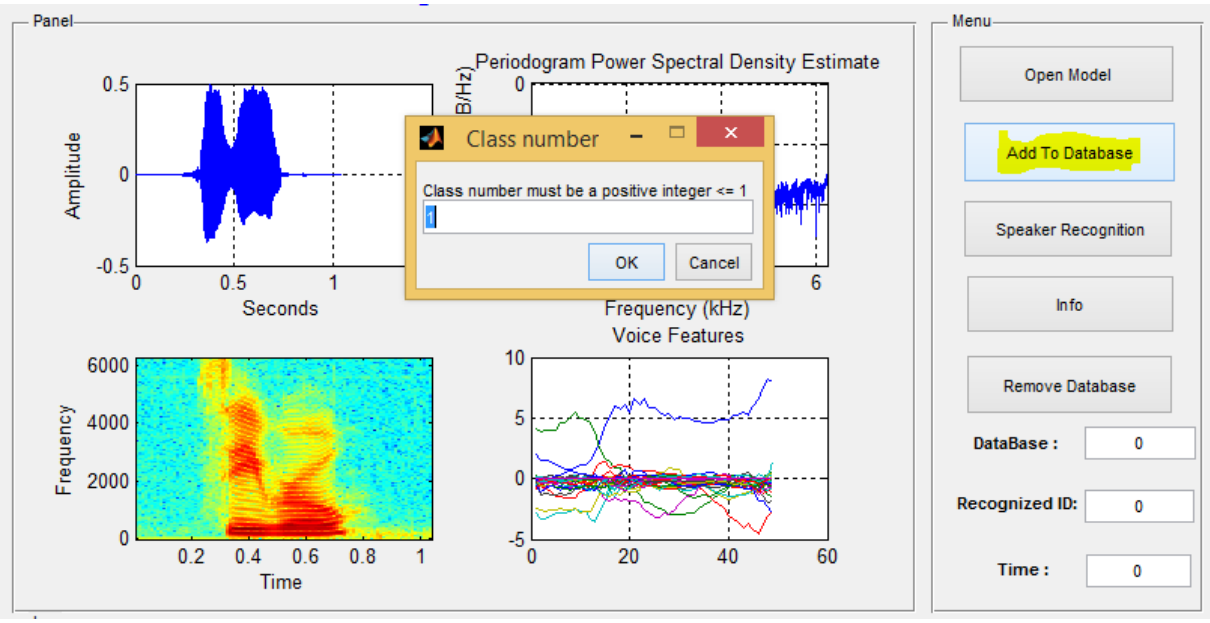
1. Copiez tous les fichiers dans le répertoire courant de Matlab et tapez "GuiMain" dans la fenêtre de commande de Matlab.



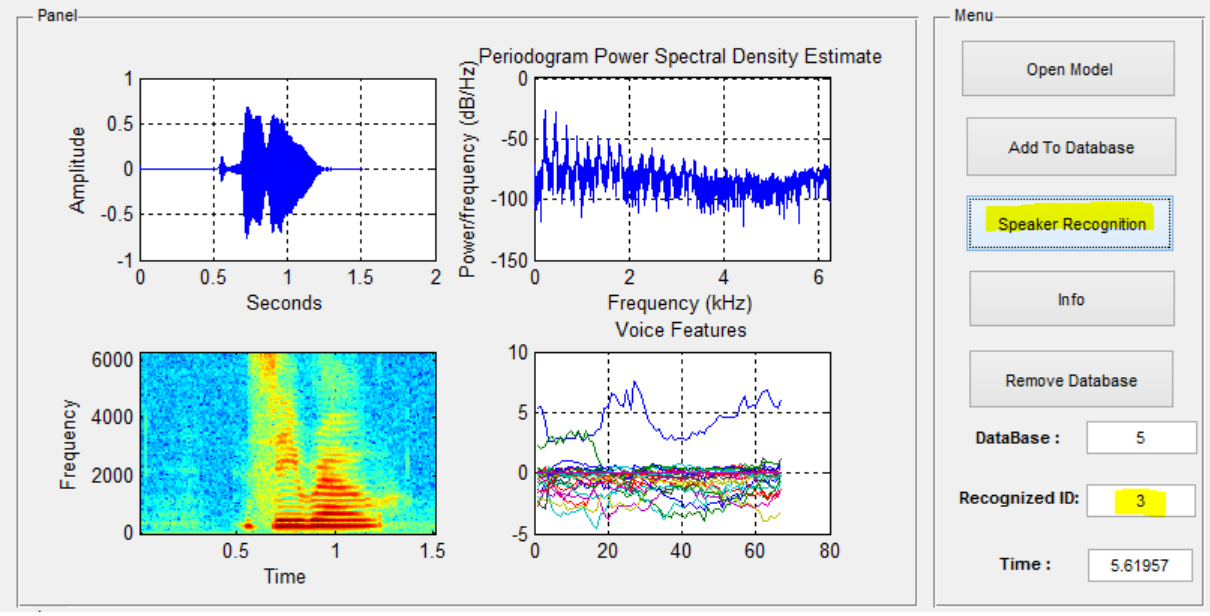
2. Sélectionnez un son d'entrée en cliquant sur **"Open Model"**. Vous pouvez ensuite ajouter ce son à la base de données en cliquant sur **"Add To Database"**.



3. Si vous choisissez d'ajouter un son à la base de données, un entier positif (ID du locuteur) est requis. Cet entier positif est un nombre progressif qui identifie une personne (chaque personne correspond à une classe).



4. La reconnaissance donne comme résultats l'identifiant de la personne la plus proche présente dans la base de données.



5. Interprétation :

Le Premier graphe : Nous montre le signal acoustique du mot isolé prononcé et représenté dans le domaine temporel, cette représentation montre l'évolution de l'intensité du signal dans le temps

Le Deuxième graphe: Nous montre le periodogramme représenté dans le domaine fréquentiel, cette représentation permet de visualiser la composition fréquentielle d'une voix mais également l'énergie de chaque fréquence

Le Troisième graphe : Nous montre le spectrogramme représenté en 3 Dimensions, il s'agit de la représentation temps-fréquence d'une voix. Ce graphe montre l'évolution de la fréquence et de l'intensité dans le temps

L'intensité est définie par la couleur : plus la couleur évolue vers le rouge plus l'intensité est importante.

Le Quatrième graphe : Nous montre les Coefficients Cepstraux de Fréquence Mel, qui ont pour rôle de montrer la représentation paramétrique de notre signal acoustique qu'on a besoin pour construire notre vecteur caractéristique qui sert à différencier la voix de chaque locuteur

4.7. Conclusion :

Dans ce chapitre, nous avons présenté les différentes étapes de réalisation d'un système de reconnaissance automatique du locuteur. Les différents blocs du système ont été programmés avec le logiciel MATLAB®. Nous avons testé la méthode et analysé les résultats des simulations, obtenues avec un nombre d'échantillons de voix différentes (fichiers .wav). Nous avons obtenu des résultats intéressants qui peuvent être améliorés.

Conclusion

Ce travail avait comme objectif principal de renforcer la sécurité dans le domaine des communications aéronautiques, Plus précisément de réaliser un système de reconnaissance automatique du locuteur simple, complet et représentatif.

En effet de nos jours, de nombreuses applications utilisent la reconnaissance automatique du locuteur grâce à son efficacité et flexibilité. Parmi ces applications, on note les systèmes d'enregistrements **ATM/ATC**. Pour cela nous avons proposé une méthode vectorielle simple de segmentation et de fenêtrage avec recouvrement, ensuite un traitement qui convertit chaque trame du domaine temporel au domaine fréquentiel en utilisant la **FFT**.

D'une part nous avons étudié l'extraction des caractéristiques ; qui est une étape importante dans le processus de reconnaissance du locuteur. Cette étape permet d'extraire des caractéristiques qui seront ensuite utilisées par le classificateur. Elle doit être faite avec soin, car elle contribue directement aux performances du système global. D'autre part, Nous avons vu les paramètres couramment utilisés qui sont les paramètres **MFCC**, convertit par la suite avec la **DCT** dans le domaine temporel et organisés dans le vecteur acoustique.

Ainsi que L'approche **VQ** qui a été utilisée, en raison de sa facilité de mise en œuvre et de sa grande précision comme méthode de classification des vecteurs acoustiques à un nombre fini de régions.

Nous avons présenté les différentes étapes de réalisation d'un système de reconnaissance Automatique Du Locuteur. Les différents blocs du système ont été programmés avec le logiciel MATLAB®. Nous avons testé la méthode et analysé les résultats des simulations, et nous avons obtenu des résultats intéressant et satisfaisants. Qui peuvent être améliorés.

Bibliographie :

[01] M. Kunt,: Techniques modernes de traitement numérique des signaux, Presses polytechniques et universitaires Romandes, 2011

[02] A. Sakina Reconnaissance de la parole par HMM , Mémoire de magister, institut d'électronique , université d'Annaba, 2014

[03] C. Snani , Conception d'un système de reconnaissance de mots isolés a base de l'approche stochastique en temps réel. Application: commande vocale d'une calculatrice Mémoire de magister, institut d'électronique, université d'Annaba, 2010

[04] K. Tomi , Spectral Features for Automatic Text-Independent Speaker Recognition Thèse University of Joensuu Department of Computer Science Finland December 21, 2010.

[05] N. Badri, Utilisation de la transformée de Fourier et de la transformée en ondelettes pour la reconnissace du locuteur, ÉTS, Montréal (Qc), 2009.

[06] M. Djemili, " reconnaissance de mots isolés arabes par DTW & HMM". Mémoire de Magister, Institut d'électronique, Université d'Annaba, 2001.

[07] P. Premarkanthan , W. B . Mikhael. Speaker verification /recognition and the importance of selective feature extraction: review. Department of Electrical Engineering, University of central Florida, Orlando, september 2008.

[08] C. Lévy, Reconnaissance de chiffres isolés embarquée dans un téléphone portablle, Laboratoire Informatique d'Avignon, France, 2010

[09] S. FURUI, MEMBER, IEEE Cepstral Analysis Technique for Automatic Speaker Verification, IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, APRIL 2002