الجمهورية الجزائرية الديمقراطية الشعبية

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR
ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE DR. YAHIA FARES DE MEDEA

FACULTE DES SCIENCES ET DE LA TECHNOLOGIE

Département du Génie Électrique et de l'Informatique



Projet de Fin d'Etudes

Présenté pour l'obtention du Diplôme
de Master en Génie Électrique

Spécialité : *Systèmes de Télécommunications*

THEME

RECONNAISSANCE DES MARQUES DE TELEPHONES PORTABLES A PARTIR DES SIGNAUX DE LA PAROLES

Proposé et dirigé par,

M. ALIMOHAD Abdenour

Préparé et présenté par,

M. BOUZEMLAL Hicham Abdelaziz.

M. SOUSSI Rabie

Dédicaces

Tout d'abord je tiens à remercier ***ALLAH*** *le tout puissant de m'avoir donné la santé, la volonté et le courage.*
Ensuite, à mes ***chers parents*** *pour leur soutien et leurs encouragements et qu'aucune dédicace ne pourra compenser leur sacrifice pour mon bien être et mon bonheur.*

A mon adorable sœur ***Madina.***

A mes frères ***Mohamed,*** *et* ***Islam.***

A mes oncles, mes tantes et toute ma famille.

A mes amis ***BENGHERBIA Ismail et DEBAGHI Bilal.***

A notre encadreur ***M. ALIMOHAD Abdenour.***

A Mon binôme **SOUSSI Rabie.**

Et à tous ceux qui nous ont soutenus
pour l'achèvement de ce projet avec réussite.
Je dédie ce modeste travail.

BOUZEMLAL Hicham Abdelaziz



Dédicaces

*Merci **ALLAH** de m'avoir donné la capacité d'écrire et de réfléchir, la force d'y croire, la patience d'aller jusqu'au bout du rêve et le bonheur de lever mes mains vers le ciel et de dire "**Ya Kayoum**"*

Je dédie ce modeste travail à celle qui m'a donné la vie, le symbole de tendresse, qui s'est sacrifiée pour mon bonheur et ma réussite,

*A **ma très chère mère** ...*

*A **mon très cher père**, école de mon enfance, qui a veillé tout au long de ma vie à m'encourager, à me donner l'aide.*

*Que **ALLAH** les gardes et les protège.*

*A mes adorables sœurs **Nassima, Latifa**.*

*A mes frères **Merouane, Mohamed Elhadi***

*A mes amis, en particulier mon binôme de ce projet **M. BLAIFI Yasmine**.*

*A notre encadreur **M. ALIMOHAD Abdenour**.*

A Mon binôme BOUZEMLAL Hicham Abdelaziz.

A tous ceux qui me sont chères.

A tous ceux qui m'aiment et ceux que j'aime.

Je dédie ce modeste travail.

SOUSSI Rabie



Remerciements

*Nous tenons tout d'abord à remercier **ALLAH** le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce Modeste travail.*

*En second lieu, nous tenons à remercier tout le corps enseignant qui a contribué à notre formation, plus particulièrement notre encadreur **M. ALIMOHAD Abdenour** pour son aide et les précieux conseils qu'il a bien voulu nous donner.*

Nos remerciements s'étendent également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail.

*On tient à remercier **M. LEBOUKH Aboubakr (dhaif allah)** pour son contribution à l'arrangement de ce mémoire*

الملخص:

ويتناول هذا العمل مشكلة الاعتراف على الهواتف الخلوية باستخدام إشارات خطاب مسجل. ولهذا الغرض، مختلف الآليات لاستخراج ألميزات الصوتية مثل ميل معاملات سبسترل ، ومعاملات الإدراك الحسي و التنبؤ الخطي. تم استخدام تقنية آلات المتجهات لكل الهاتف الخليوي. تم بناء قاعدة بيانات لاستخدامها في التجارب..

كلمات مفتاحية: إشارات كلام مسجله، آلات لناقلات دعم ، معاملات سبسترو، التنبؤ الخطي، الاعتراف على الهاتف الخليوي.

RÉSUMÉ

Ce travail traite le problème de la reconnaissance de téléphones cellulaires en utilisant les signaux de parole enregistrées. Dans cet objectif l'extraction les caractéristique acoustiques telles que les coefficients Mel cepstraux (MFCC), et les coefficients de prédiction linéaire perceptuelle (PLP) sont considérés. Pour modéliser ces paramètres, on a utilisé la technique de machines à vecteurs de support (SVM) pour chaque téléphone cellulaire. Une base de données a ete construite pour être employée dans les expériences effectuées.

Mots clés : *Parole, Machines à vecteurs de support (SVM), Coefficients Cepstraux MFCC, Prédiction Linéaire Perceptuelle PLP, téléphone cellulaire.*

ABSTRACT

This work deals the cell-phone recognition problem using recorded speech signals. For this, acoustic feature extraction methods such as Mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction (PLP) are considered. To model these parameters, we used support vector machine (SVM) technique for each cell-phone. A database is constructed to be used in all performed experiments.

Key words: *speech, support vector machine (SVM), Mel Frequency cepstral Coefficients MFCC, Perceptual Linear Prediction PLP, recognition of cell phone.*

Table des métiers

Liste des figures..... I
Liste des tableaux..... II
Liste des acronymes…… III
Introduction générales……... 1
CHAPITRE I : Concepts Généraux sur la Reconnaissance Basée sur la Parole 3
1. Introduction……... 4
2. Reconnaissance de la parole…… 5
3. Reconnaissance du locuteur…….. 5
4. Reconnaissance des téléphones mobiles……... 6
5. Production de la parole…… 8
6. Analyse de la parole…….. 11
6.1 Analyse Spectral à court terme…….. 11
6.1.1 Transformée de Fourier court terme……... 11
6.2 Fenêtrage……... 12
6.3 Préaccentuation…… 14
6.4 Extraction des paramètres…… 15
6.4.1 Les paramètres MFCC (Mel Frequency cepstral Coefficients)…….. 16
6.4.2 Les paramètres LPC (Linear Predictive Coefficients)…… 16
6.4.3 Les paramètres PLP (Perceptual Linear Prediction) …….. 17
6.4.4. Delta and Delta-Delta Coefficients……..... 17
7. Conclusion…….. 17
CHAPITRE II : Système de reconnaissance 18
1. Introduction……... 19
2. Méthodes d'extraction des paramètres…… 20
2.1. Coefficients Perceptuelles de prédiction linéaire (PLP) …….. 21
2.1.1. Fenêtrage de Hamming et FFT…….. 21
2.1.2. Banc du filtre Bark…… 22
2.1.3. Courbe isosonique…… 24
2.1.4. Compression intensité-volume…….. 25
2.1.5. L'IDFT et l'analyse prédictive linéaire…….. 25
2.1.6. L'analyse cepstrale…… 27
2.2. Les Coefficients Mel Fréquences Cepstraux (MFCC) …… 29
2.2.1. Banc du filtre MEL…… 29
2.2.2. Compression logarithmique……... 31
2.2.3. DCT……... 31
3. Méthodes de modélisation…… 32
3.1. Déformation temporelle dynamique DTW (dynamic time warping) ……. 32
3.2.Modèle de mélange gaussien GMM (Gaussian Mixture Model) ……. 33

3.3. Machines à vecteurs de support SVM (Support Vector Machine) …….. 33
3.3.1. Cas des données linéairement séparables…….. 35
3.3.2. Cas des données non-linéairement séparables……... 38
3.3.3. Type de noyau……... 39
3.3.4. SVMs binaires……... 39
3.3.5. SVMs multiclasse…….. 40
3.3.1.1. Une-contre-reste (1vsR) …… 40
3.3.1.2. Une-contre-une (1vs1) ……... 42
4. Décision……... 44
5. Conclusion. …… 44
CHAPITRE III : Implémentation du système de reconnaissance 45
1. Introduction……... 46
2. Base de données et Protocol d'évaluation…… 46
2.1. Construction de la base de données…….. 46
2.2. Protocol d'évaluation…… 48
2.2.1. Collection des échantillons et extraction des paramètres acoustique…… 48
2.2.2. Reconnaissance des marques de téléphones portables par la méthode(SVM).. 48
2.2.3. Fonctionnalités de la LIBSVM…….. 49
3. Expériences et résultats…… 50
4. Conclusion…….. 52
Conclusion générale……... 53
Refereces…… 54

Liste des figures

Figure I.1. *Vue schématique de l'appareil vocal11, dans le plan sagittal médian…….* 9

Figure I.2. *Modèle de production de la parole……...* 10

Figure I.3. *Signal pondéré par une fonction fenêtre : (a) signal ; (b) fenêtre de durée infinie ; (c) fenêtre symétrique de durée finie* ***N*****……..** 11

Figure I.4. *Les étapes de segmentation et de e fenêtrage.* **…….** 11

Figure I.5. *Le découpage en trames.* **……..** 13

Figure I.6. *La fenêtre de Hamming et son spectre.* **……** 14

Figure I.7. *Le filtre de la préaccentuation.* **……** 15

Figure II.1. *Structure d'un système de reconnaissance.* **……** 19

Figure II.2. *Schéma de l'algorithme PLP.* **……..** 21

Figure II.3. *Un exemple de la distribution de la fréquence centrale dans un Banc du filtre Bark (fréquence de Nyquist =5 kHz (16,9 Barks)).* **……** 22

Figure II.4. *Banc du filtre Bark.* **……** 23

Figure II.5. *Illustration des étapes de l'algorithme MFCC***……** 29

Figure II.6. *Le banc du filtre MEL (Fréquence d'échantillonnage =16 kHz)***……...** 30

Figure II.7. *Schéma d'un système de reconnaissance basé sur la comparaison dynamique. .* **……** 32

Figure II.8. *Principe des SVM. .* **……** 34

Figure II.9. *Hyperplans séparateurs : H est un hyperplan quelconque,* ***HO*** *est l'hyperplan optimal, VS : sont les Vecteurs Support.* **……** 35

Figure II.10. *Hyperplans séparateurs dans le cas de données non-linéairement séparables : H est un hyperplan quelconque,* ***HO*** *est l'hyperplan optimal, VS : sont les Vecteurs Sup…***……** 38

Figure II.11. *SVM binaire.* **……** 40

Figure II.12. *Approche une-contre-reste avec des zones d'indécision…...……* 41

Figure II.13. *Résolution des cas d'indécision dans la méthode 1vsR.* **……...** 42

Figure II.14. *Approche une-contre-une.* **……** 43

Liste des tableaux

Tableau III.1. *La marque et le modèle des téléphones cellulaires*…… 47

Tableau III.2. *Effet de l'énergie sur le taux d'identification (D=12).* ……..… 50

Tableau III.3. *Effet des caractéristiques dynamiques et l'augmentation du nombre de paramètres sur le taux d'identification.* ……..……..…….. 51

Liste des acronymes

IHM:	Human-Machine Interfaces
LPC:	Linear Predictive Coding
LPCC:	Linear Predictive Coding Cepstra
MFCC:	Mel-frequency Cepstral Coefficients
PLP:	Perceptual Linear Prediction
FFT:	Fast Fourier Transform
IDFT:	Inverse Discrete Fourier Transform
DCT:	Discrete Cosine Transform
DTW:	Dynamic Time Warping
GMM:	Gaussian Mixture Model
SVM:	Support Vector Machine
Hz:	Hertz
DAP:	Acoustic-Phonetic Decoding
EM:	Expectation-Maximization
TIMIT:	Texas Instruments/Massachusetts Institute of Technology
SNR:	Signal Noise Ratio
OCR	La reconnaissance des caractères optiques

INTORODUCTION GENERALE

Le développement technologique dans des appareils électroniques a mené à leur utilisation courante par les personnes dans leur vie quotidienne. Des dispositifs tels que les ordinateurs, les appareils photo numériques, les téléphones cellulaires, les imprimantes et les scanners permettent de produire des données numériques qui peuvent être stockées ou transmises à un autre dispositif. Cependant, ces développements apportent une remise en question sur l'ensemble des problèmes posés. De nos jours, dans beaucoup d'affaires en juridique, ces appareils numériques et outils logiciels sophistiqués qui sont produits par telles ou telles marques sont présentés comme preuves. Ainsi, le cheminement de la source de ces appareils serait essentiel pour une cour de justice.

La recherche légale a été en grande partie concentrée sur la médecine légale d'image durant les dernières décennies. Le dispositif de source (appareil-photo, imprimante ou identification de scanner), le problème d'identifier le dispositif de source utilisant des images, la détection de la manipulation d'image, la stéganalyse d'image qui vise à détecter la présence des messages cachés dans les images sont les plus populaires applications des médecines légales d'image.

La criminalistique audio est un autre domaine de la recherche légale, mais qui a été beaucoup moins étudiée. Les criminalistiques audio ont été classées en trois groupes : i) authentification des preuves audio, ii) amélioration des enregistrements audio et iii) interprétation des preuves que les enregistrements audio contiennent. L'authentification audio est la capacité de déterminer si un enregistrement audio est original et s'il a été trifouillé. Généralement, l'analyse spectrale, l'analyse de spectrogramme et l'analyse de continuité de phase sont employées pour l'authentification audio. La steganalysis audio, la tâche de détecter la présence des messages cachés dans les signaux audio, est une autre application intéressante la criminalistique audio. L'amélioration de la parole qui vise à enlever le bruit additif non désiré des signaux audio, est généralement employée comme étape de prétraitement dans beaucoup d'applications de la parole telles que la reconnaissance de la parole et de locuteurs. La reconnaissance de la parole, identifiant le contexte d'un signal audio, la reconnaissance de locuteur, la tâche d'identifier le locuteur, et la reconnaissance de son articulé et de la langue étant parlée sont les applications les plus populaires des médecines légales audio. Dans une autre étude récente, il a montré que la technique de codage et de décodage (codec) peut

également être identifiée en utilisant les sons articulés. Des techniques légales peuvent être employées pour identifier l'origine d'un matériel numérique. Ceci est également connu en tant que caractérisation légale qui signifie identifier le type du dispositif, du modèle et d'autres caractéristiques. Un dispositif de source peut être identifié à l'aide de données qui sont produites par cet appareil particulier. Dans notre étude, nous avons abordé un récent problème de la criminalistique, qui est l'identification de la source audio du téléphone portable en utilisant des enregistrements audio. Ceci en montrant que signal de parole transmet des informations sur le dispositif d'enregistrement de la source.

La raison de notre motivation de considérer les téléphones cellulaires, c'est le fait qu'ils font partie intégrante de notre vie humaine. Puisque les téléphones cellulaires ont leurs microphones intégrés qui peuvent être utilisés en tant qu'enregistreurs ordinaires de la parole, la fonction de transfert de différents microphones de téléphone cellulaire sera différente entre eux. La différence dans des fonctions de transfert sera plus grande quand des composants électroniques sont produits par différents fabricants. En raison des tolérances des composants, il y aura des différences même que des composants produits par le même fabricant. Ainsi, chaque téléphone cellulaire laisse ses empreintes sur les sons articulés enregistrés et ceci peut être employé pour la reconnaissance de dispositif de source en utilisant les signaux audio enregistrés.

Dans ce document, nous utiliserons les caractéristiques acoustiques pour la reconnaissance de téléphone cellulaire de source.

Pour cela, Nous allons commencer dans le premier chapitre par des notions de base sur la reconnaissance ainsi que la parole. Nous étudierons plus en détail le système de reconnaissance de téléphones portables à travers ces diverses parties comme l'extraction de paramètres, la modélisation et la prise de décision. Nous analyserons et nous discuterons les différents résultats obtenus dans les expériences réalisées dans ce travail et nous terminerons par une conclusion générale.

CHAPITRE I : Concepts Généraux sur la Reconnaissance Basée sur la Parole

1. Introduction

Le signal de parole n'est pas un signal ordinaire, il est le vecteur d'un phénomène extrêmement complexe : communication parlée. Il est difficile de modéliser le signal de parole car ses propriétés statistiques varient au cours du temps. Sa redondance lui confère une robustesse à certains types de bruit. Il possède aussi une très grande variabilité. Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution peut varier, la durée du signal est alors modifiée. Toute altération de l'appareil phonatoire peut modifier la qualité de l'émission (par exemple rhume, fatigue). La variabilité interlocuteur est encore plus évidente, la hauteur de la voix, l'intonation, l'accent différent selon le sexe, l'origine social, régionale ou nationale [1].

Le signal de parole est un signal acoustique produit par des vibrations des couches d'air. Les fluctuations de la pression de l'air produisent des variations de ce signal, en fonction du temps, qui peuvent être enregistrées de façon analogique ou digitale.

La parole est formée de phonèmes et de transitions entre ces phonèmes. Les phonèmes sont les éléments les plus brefs qui permettent de distinguer différents mots. Plusieurs formes du conduit vocal peuvent produire le même phonème. La forme que le conduit vocal prend pour la production d'un phonème, dans un contexte donné, est assez variable et surtout dépendante de ce contexte. De ce fait, nous remarquons que les formes acoustiques associées à un phonème déterminé sont variables. Plusieurs types de phonèmes existent : les voyelles, les consonnes fricatives et les consonnes plosives, les nasales et les liquides.

L'acoustique du milieu (milieu protégé ou environnement bruité), la qualité du microphone ou de la ligne téléphonique, les bruits de la bouche, les hésitations, les mots hors-vocabulaire sont autant d'interférences supplémentaires sur le signal de parole que le système de reconnaissance doit compenser.

L'aspect continu du signal de parole complique encore la tâche de reconnaissance. En effet, lorsqu'on écoute parler une personne, on perçoit une suite de mots, alors que l'analyse du signal vocal ne permet de déceler aucun séparateur. Le même problème de segmentation se retrouve à l'intérieur du mot lui-même. Celui-ci est perçu comme une suite de sons élémentaires.

Aussi, l'analyse du signal ne permet pas de découper en segments distincts le signal acoustique afin d'identifier les différents phonèmes qui le composent. [2]

2. **Reconnaissance de la parole**

La reconnaissance automatique de la parole (souvent improprement appelée reconnaissance vocale) est une technique informatique qui permet d'analyser la parole captée au moyen d'un microphone pour la transcrire sous la forme d'un texte exploitable par une machine. La reconnaissance de la parole, ainsi que la synthèse de la parole, l'identification du locuteur ou la vérification du locuteur, font partie des techniques de traitement de la parole. Ces techniques permettent notamment de réaliser des interfaces vocales c'est-à-dire des interfaces homme-machine (IHM) où une partie de l'interaction se fait à la voix. Parmi les nombreuses applications, on peut citer les applications de dictée vocale sur PC où la difficulté tient à la taille du vocabulaire et à la longueur des phrases, mais aussi les applications téléphoniques de type serveur vocal, où la difficulté tient plutôt à la nécessité de reconnaître n'importe quelle voix dans des conditions acoustiques variables et souvent bruyantes (téléphones mobiles dans des lieux publics). [3]

3. **Reconnaissance du locuteur**

La reconnaissance du locuteur est une des branches de l'authentification biométrique, qui se réfère à la reconnaissance automatique de l'identité des personnes en utilisant certaines de leurs caractéristiques intrinsèques. Outre la voix, il y a beaucoup d'autres modèles physique et comportementaux pour l'authentification biométrique, par exemple : les yeux, le visage, les empreintes, la signature, …etc. Pratiquement, la sélection d'un modèle biométrique prometteur devrait prendre en compte au moins les considérations suivantes : la robustesse, la précision, l'accessibilité et l'acceptabilité. Par rapport à ces critères de sélection, parmi toutes les technologies d'authentification biométriques, la reconnaissance du locuteur est probablement la plus naturelle et économique pour les systèmes de communication homme-machine et cela en raison de 1) la collecte de données parole est beaucoup plus pratique que les autres motifs, et 2) plus important encore, la parole et le mode dominant d'échange d'information pour les êtres humains et tend à être le mode dominant pour l'échange d'information pour les système de communication homme-machine.

Le développement de la technologie de traitement de la parole a boosté les nombreuses applications de reconnaissance du locuteur, plus particulièrement dans les domaines suivants:

Contrôle d'accès aux installations physiques ou des réseaux de données ;
Achats par téléphone ou d'autres transactions bancaires ;
Recherche d'information, par exemple, renseignements sur les clients pour les centres d'appels et l'indexation audio ;
La surveillance à distance ;
Criminalistique.

La reconnaissance du locuteur peut être divisée en deux taches: vérification et identification. La tâche de vérification est de décider si oui ou non une voix inconnue appartient au locuteur revendique. Il n'y a que deux décisions possibles : soit accepter la voix comme étant celle de la personne clamée ou bien la rejeter si la voix estconsidérée comme étant celle d'un imposteur. La tâche d'identification consiste à classer la voix inconnue comme appartenant à l'un des locuteurs de la base. Le nombre de décisions alternatives dans l'identification du locuteur est égal a la taille de la population de la base de locuteurs N, généralement la performance du système et inversement proportionnelle à N. Par conséquent, l'identification est généralement une tache plus difficile qu'une vérification. Cette identification est connue comme une identification en ensemble fermé (closed-set en anglais). Par contraste, l'identification en ensemble-ouvert (open-set en anglais) intègre la possibilité que la voix inconnue n'appartient à aucun locuteur de la base. Par conséquent, le nombre de décisions alternatives est de $N + 1$, ce qui veut dire que la voix peut appartenir à un locuteur n'appartenant pas à la base de référence. L'identification en ensemble-ouvert est une combinaison entre l'identification et la vérification. [4]

4. **Reconnaissance des téléphones mobiles**

Reconnaissance des téléphones mobiles présente la comparaison et l'optimisation des caractéristiques acoustiques de la source de reconnaissance de téléphone cellulaire en utilisant des signaux de parole enregistrées. Différentes méthodes d'extraction de caractéristiques acoustiques tels que les coefficients Mel cepstraux (Mel frequency cepstral coefficients MFCC) et coefficients de prédiction linéaire perceptuelle (perceptual linear predictive PLP)

peuvent être considérés, En plus de ces coefficients, les effets dynamiques peuvent être introduites par l'ajout des coefficients delta et double delta (Δ et Δ^2).

La criminalistique audio est un défi de la recherche médico-légale qui a été beaucoup moins étudiée. La médecine légale audio a été classé en trois groupes: i) l'authentification des preuves audio, ii) l'amélioration des enregistrements audio et iii) l'interprétation des preuves que les enregistrements audio contiennent. L'authentification audio est la capacité de déterminer si un enregistrement audio est original ou a été altérée. En général, l'analyse spectrale, l'analyse de spectrogramme et l'analyse de la continuité de phase sont utilisés pour l'authentification audio. La reconnaissance vocale, tenant compte du contexte d'un signal audio, la reconnaissance du locuteur, la tâche de reconnaître le haut-parleur du signal de parole et la reconnaissance de la langue, qui vise à identifier la langue étant parlée sont les applications les plus populaires de la médecine légale audio.

Les techniques médico-légales peuvent être utilisées pour identifier l'origine d'un matériel numérique. Ceci est également connu comme la caractérisation de la médecine légale qui consiste à identifier le type de caractéristiques de l'appareil, et d'autres modèles. Un dispositif de source peut être identifié en utilisant les données qui sont produits par cet appareil particulier. Dans notre étude récente, nous avons abordé un problème de la criminalistique, la source audio identification du téléphone portable en utilisant des enregistrements audio et montré que le signal de parole transmet des informations sur le dispositif d'enregistrement de la source.

Dans une autre étude récente, il a été montré que les signaux de parole peuvent être utilisés pour identifier les appareils de téléphonie fixe (carbone ou à électret). La raison de cette motivation est qu'ils sont devenus une partie intégrante de la vie humaine. Du point de vue médico-légal, l'utilisation généralisée des téléphones cellulaires dans la vie quotidienne fournira beaucoup de preuves dans les signaux de parole enregistrées par les téléphones cellulaires et il y a une tendance à la hausse sur la présentation de ces enregistrements comme preuve à la cour de lois. Comme chaque téléphone cellulaire a un microphone intégré permettant de faire des enregistreurs de la parole ordinaire, la fonction de transfert de ce téléphone sera différente d'un autre possédant un autre modèle de microphone. La différence de la fonction de transfert sera plus grande lorsque les composants électroniques sont fabriqués par différents fabricants. En raison des tolérances des composants, il y aura même une différence de composants produits par le même fabricant. Ainsi, chaque téléphone

cellulaire laisse ses empreintes révélatrices sur des signaux de parole enregistrés. Ceci peut être utilisé pour la reconnaissance de l'appareil source à l'aide des signaux audio enregistrés. [5]

5. **Production de la parole**

La production de la parole fait intervenir différents organes. La source de la parole provient des poumons qui émettent un flux d'air. Ce flux d'air va traverser le larynx pour faire vibrer ou non les cordes vocales. Il va ensuite traverser le conduit vocal (cavité nasale et buccale) et les articulateurs tels que les lèvres et la langue (Figure I.1).

La production de la parole est un processus complexe dans lequel sont impliqués de nombreux organes dont les principaux sont représentés sur la Figure I.1. Le contenu fréquentiel du signal acoustique de parole produit par un locuteur est fortement dépendant des caractéristiques morphologiques de son appareil phonatoire. Celui-ci peut être divisé en quatre parties : le générateur, le vibrateur, le résonateur et les modulateurs.

Le générateur **:** L'air expulsé des poumons est le moteur de la parole. Cet air traverse l'appareil phonatoire comme un instrument à vent et crée la pression nécessaire à la génération d'un signal acoustique.

Le vibrateur **:** L'air expulsé des poumons traverse la trachée pour arriver dans le larynx où se trouvent les cordes vocales. Les cordes vocales sont une paire de muscles dont la longueur moyenne se situe entre 20 et 25 millimètres. Cette longueur varie d'un individu à l'autre. L'air traversant le larynx met en vibration les cordes vocales. La fréquence de vibration des cordes vocales est modulée en fonction de leur degré de contraction. Le locuteur peut ainsi moduler la hauteur des sons qu'il émet.

Le résonateur **:** Ces vibrations sont modifiées par le passage de l'air dans les différentes cavités qui composent le pharynx mais aussi dans les fosses nasales, la bouche et le larynx avec qui il communique. Ces résonateurs influent sur le son en atténuant certaines fréquences et en amplifiant d'autres. La forme et le volume de ces cavités, spécifiques au locuteur, modifient fortement le son produit.

Les modulateurs **:** Enfin les organes modulateurs que sont la langue, les lèvres et la mâchoire sculptent le sont pour produire les phonèmes qui composent la parole. La position de ces différents organes est le mécanisme final qui permet la production de parole articulée.

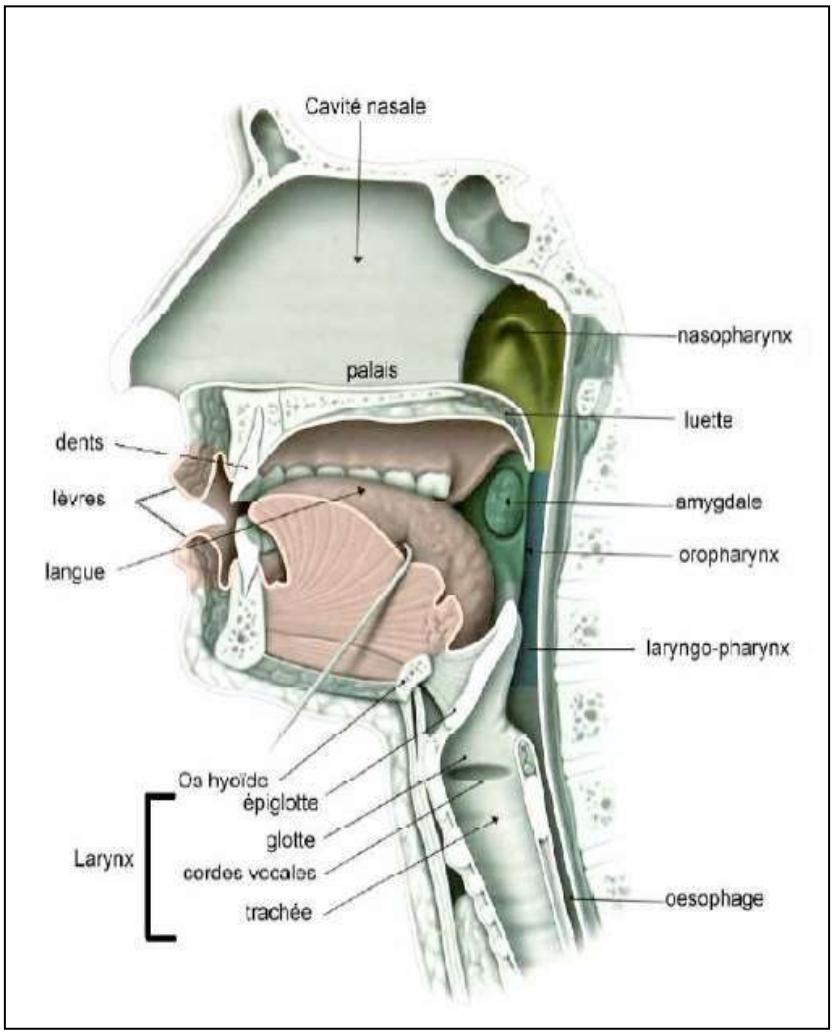


Figure I.1. *Vue schématique de l'appareil vocal11, dans le plan sagittal médian***.**

L'ensemble des organes agit comme un filtre, considéré comme linéaire, dont la réponse impulsionnelle comporte des fréquences de résonance caractérisées par des pics, appelés formants, dans le spectre du signal de sortie. Le signal résultant est globalement non stationnaire mais peut être considère comme stationnaire sur de très courtes périodes, de l'ordre de 20ms (signal pseudo-stationnaire). Sur un segment de parole de cette longueur la voix est habituellement et schématiquement séparée en deux classes distinctes :

1. *voisée :* lorsqu'il y a vibration des cordes vocales, le signal est alors quasi-périodique,

2. *non voisée :* dans le cas d'un simple soufflement, le signal est alors considère comme aléatoire.

Dans le premier cas, la source d'excitation est modélisée par un train d'impulsions périodique, de fréquence dite de voisement F_0, qui correspond à la fréquence de vibration des cordes vocales, la fréquence fondamentale ou pitch ;

$$\mathbf{u(n) = \sum_k \delta(n - k.p)} \qquad \textbf{(I.1)}$$

Où p représente la période de la fréquence fondamentale (en anglais, ≪ pitch ≫).

Dans le second cas, la source est modélisée par un bruit blanc.

Cette représentation binaire de la production de la parole a été introduite par [6]. Elle est reprise sur la Figure I.2.

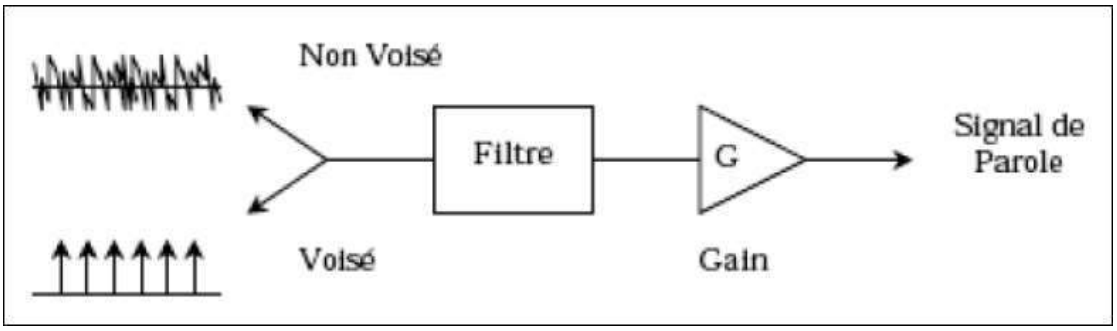


Figure I.2. *Modèle de production de la parole***.**

Certains organes impliqués dans la production de la parole se trouvent en partie visibles pour un observateur extérieur (lèvres, langue, position des mâchoires). La configuration des organes modulateurs et leurs mouvements fournissent donc une information visuelle caractéristique de la parole produite Ainsi la parole est un phénomène temporel bi-modal structuré, dont les manifestations acoustiques et visuelles peuvent être observées au cours du temps. L'utilisation de la parole en biométrie rend possible l'acquisition simultanée du signal acoustique de parole et du signal vidéo correspondant. Cette acquisition est d'autant plus facile à réaliser que le matériel nécessaire (microphone et camera) est bon marché et présent

sur un grand nombre d'appareils commercialises tels que les téléphones cellulaires, ordinateurs de bureau ou portables ... etc. [4]

6. **Analyse de la parole**

Le traitement de la parole est un champ de recherche riche mais difficile. En effet, le signal de parole est complexe car sa structure résulte de l'interaction entre la production des sons et leur perception par l'oreille. Les chercheurs du domaine ont réussi malgré tout à réaliser quelques systèmes de reconnaissance ou de synthèse exploitables dans des conditions réelles. [2]

6.1. ***Analyse Spectral à court terme***

L'analyse Spectral à court terme est définie à partir de la transformée de Fourier court terme.

6.1.1. Transformée de Fourier court terme.

On va tenir compte du caractère non stationnaire d'un signal x en le pondérant par une fonction fenêtre w repérée par rapport à un instant particulier n (Figure 0I.3) ; cette fenêtre peut-être de durée infinie a décroissance exponentielle (Figure 0I.3 (b)) ou de durée finie, et en général symétrique (Figure 0I.3 (c)).

La transformée de Fourier court-terme d'un signal échantillonné est par définition la transformée du signal pondéré :

$$\mathbf{X(n,\theta) = \sum_m x(m).w(n-m).exp(-jm\theta)} \quad ; \qquad \textbf{(I.2)}$$

$$\mathbf{O\grave{u}\ \theta = wT}$$

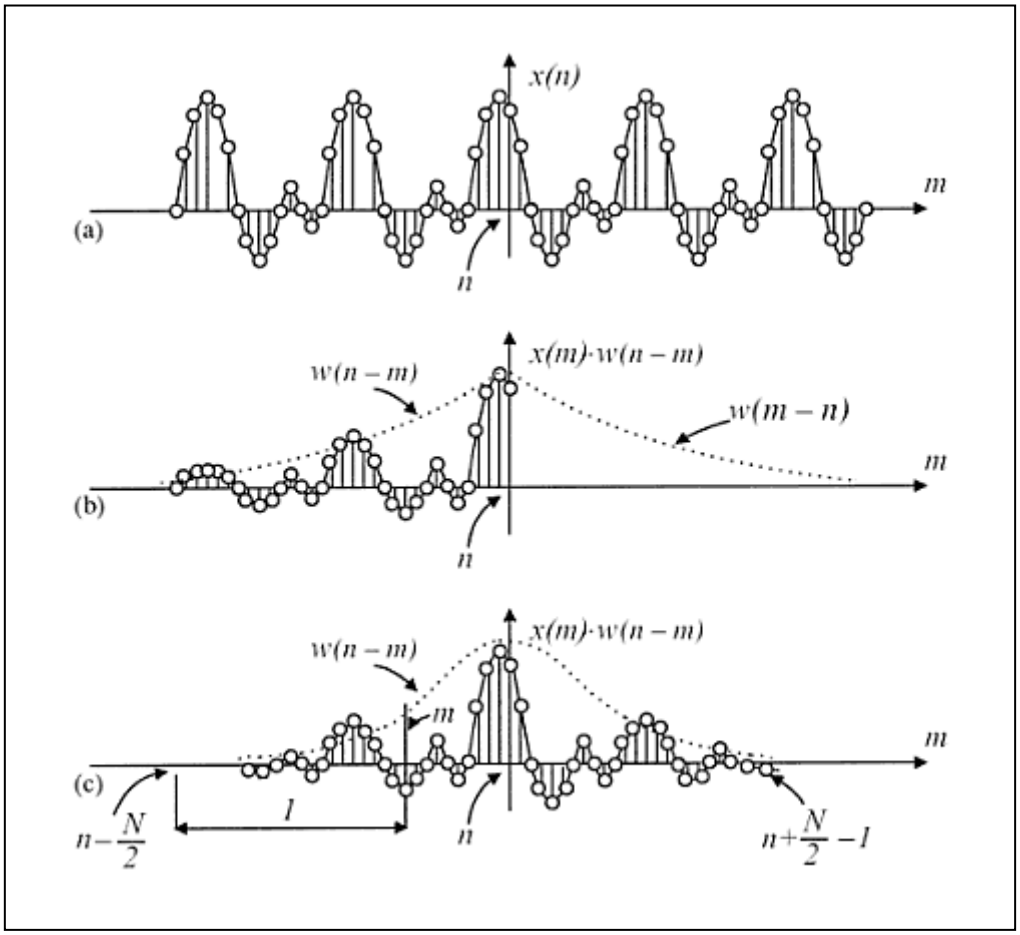


Figure I.3. *Signal pondéré par une fonction fenêtre : (a) signal ; (b) fenêtre de durée infinie ; (c) fenêtre symétrique de durée finie* ***N***

La transformée de fourrier est fonction de l'instant n : si l'on fixe $n = n_0$, la fonction $X(n_0, \theta)$ est la transformée de Fourier du signal $x(m).w(n_0 - m)$ défini pour $m \in]-\infty, \infty[$. [7]

6.2. ***Fenêtrage***

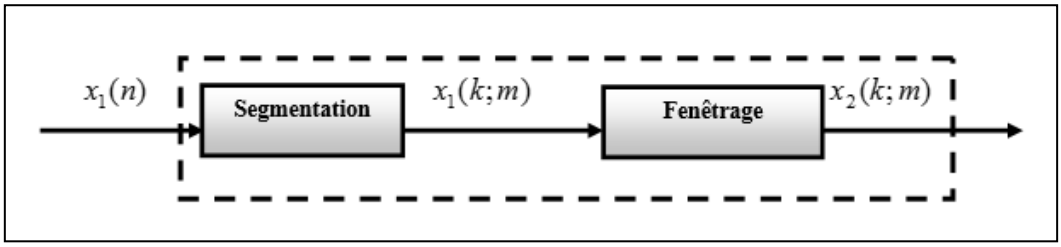


Figure I.4. *Les étapes de segmentation et de e fenêtrage*

Le signal vocal est un signal non stationnaire. Il présente une évolution lente dans le temps.

Le but du découpage est de découper le signal de parole en petites tranches (chacune de durée 10 à 30 ms) où il peut être considérer localement comme quasi-stationnaire.

En outre, et pour profiter de l'évolution lente du signal vocal, le fenêtrage permet le traitement en temps réel et il facilite aussi l'analyse des signaux sur la machine. Les ressources d'une machine étant limitées, le signal ne peut pas être traité dans sa globalité.

Chaque fenêtre est de longueur de N échantillons, telles que les trames adjacentes sont chevauché par un nombre d'échantillon échantillons (de 5 à 15 millisecondes environ), cela est illustré à la Figure I.5.

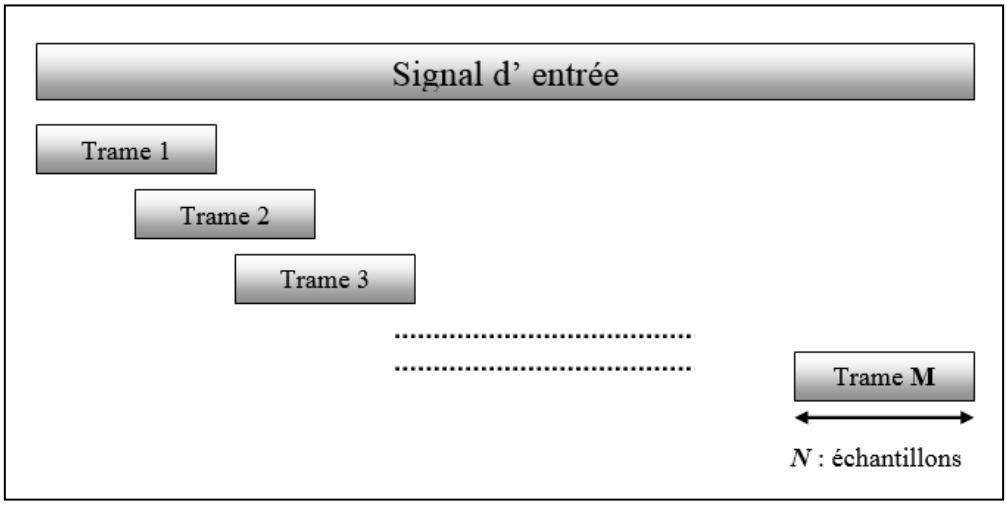


Figure I.5. *Le découpage en trames*

Il existe plusieurs types de fenêtres d'analyse, parmi ces fenêtres, la fenêtre de *Hamming* est la plus convenable à la parole, car elle entraîne un minimum de distorsion spectrale du signal de parole, par rapport aux autres fenêtres, (Atténuation du rapport du lobe principal au lobe secondaire faible, c'est à dire que la concentration de l'énergie est dans le lobe principal, voir Figure I.6). Elle est définie comme suit :

$$\mathbf{w[n]} = \begin{cases} \mathbf{0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right)} & \mathbf{0 \leq n \leq N-1} \\ \mathbf{0} & \mathbf{autrement} \end{cases} \qquad \textbf{(I.3)}$$

Où N est la longueur de la fonction de fenêtrage.

Dans cette étape on applique sur chaque trame une pondération de *Hamming*, cette pondération ayant pour but d'une part de concentrer la répartition de l'énergie sur les basses fréquences et d'autre part d'amoindrir les fortes variations du signal sur les bords de la fenêtre, variations qui entraînent une mauvaise estimation des coefficients du filtre si elles n'étaient pas atténuées. [8]

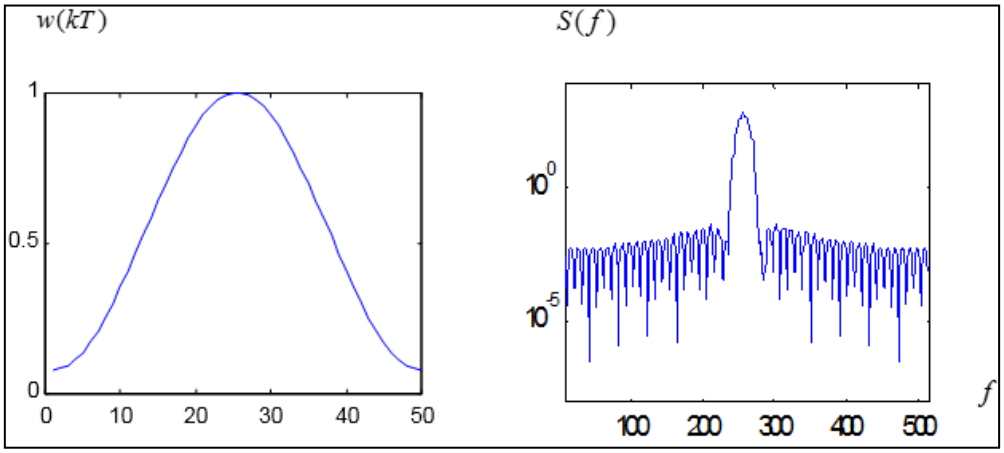


Figure I.6. *La fenêtre de Hamming et son spectre*

6.3. ***Préaccentuation***

Le spectre du signal de la parole a une décroissance globale de l'énergie de -6 dB/Octave, due à l'influence de la source d'excitation et au rayonnement des lèvres. Pour compenser cette décroissance on fait une préaccentuation en utilisant un filtre passe-haut. Le filtre le plus utilisé est le filtre à réponse impulsionnelle finie décrit ci-dessous [9] :

$$\mathbf{H(z) = 1 - 0.95z^{-1}} \qquad \textbf{(I.4)}$$

La réponse de ce filtre est représentée dans la Figure I.7.

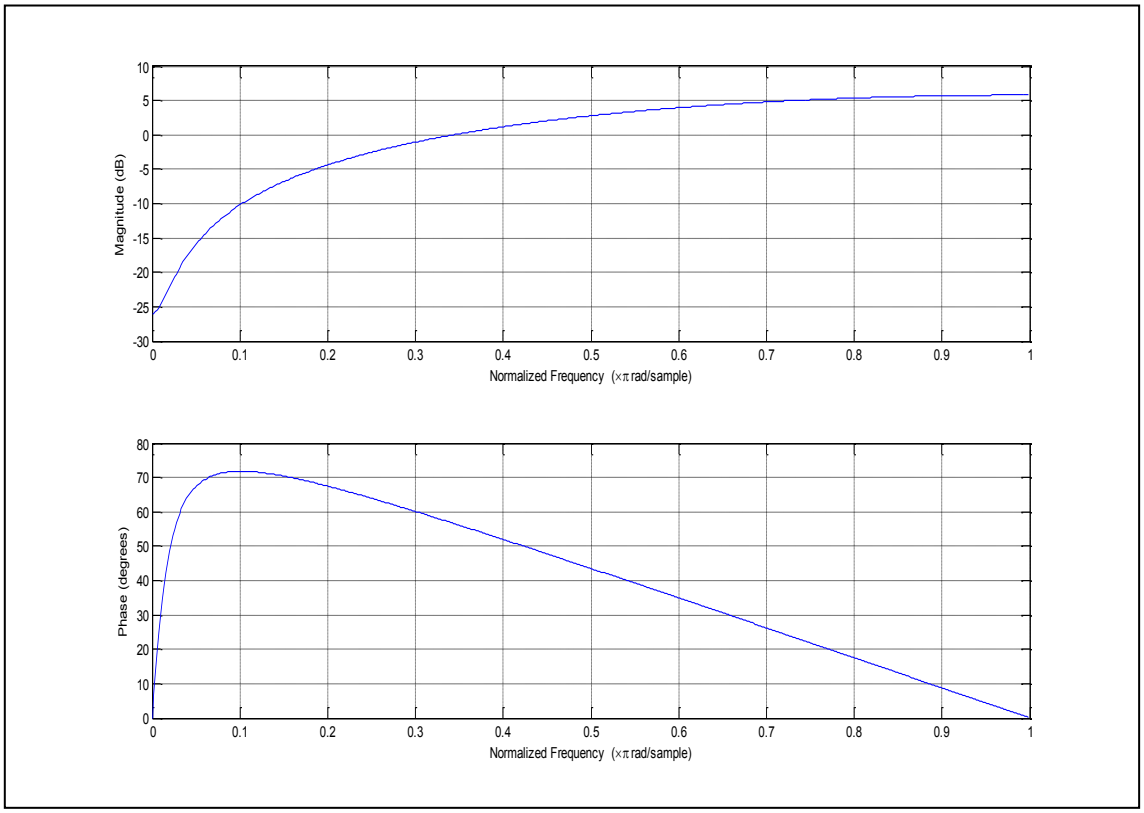


Figure I.7. *Le filtre de la préaccentuation.*

6.4. ***Extraction des paramètres***

Le signal de parole est complexe et redondant. Il possède une grande variabilité. Pour que le système de reconnaissance de la parole fonctionne efficacement, les informations caractéristiques et invariantes doivent être extraites du signal de parole. Cette procédure consiste à associer au signal de parole une ***série de vecteurs de paramètres généralement acoustiques***.

La phase d'extraction de caractéristiques doit être faite avec soin, car elle contribue directement aux performances du système global. Les coefficients les plus couramment utilisés sont les coefficients linéaire prédictifs (Linear Predictive Coefficients « LPC ») et les coefficients linéaires prédictifs cepstraux (Linear Predictive cepstral Coefficients « LPCC »), les coefficients mel frequences cepstraux (Mel Frequency Cepstre Coefficients « MFCC ») ou bien les coefficients linéaires prédictifs perceptuels (Perceptual Linear Predictive « PLP »). Les coefficients MFCC et PLP ont la propriété d'intégrer des connaissances du modèle auditif humain [2].

6.4.1. Les paramètres MFCC (Mel Frequency cepstral Coefficients)

Les paramètres MFCC sont des coefficients cepstraux calculés par une transformée en cosinus discrète appliquée au spectre de puissance d'un signal. Les bandes de fréquence de ce spectre sont espacées logarithmiquement selon l'échelle de Mel.

Les étapes de MFCC sont:

Calcul de la transformée de Fourier de la trame à analyser
Pondération du spectre d'amplitude (ou de puissance selon les cas) par un banc de filtres triangulaires espacés selon l'échelle de Mel
Calcul de la transformée en cosinus discrète du log-mel-spectre Les coefficients résultants de cette DCT sont les MFCCs.

Le cepstre d'un signal est utilisé par exemple en traitement de la parole et en reconnaissance vocale. [10]

6.4.2. Les paramètres LPC (Linear Predictive Coefficients)

L'analyse par prédiction linéaire LPC est fortement utilisée dans le codage de parole dans le but de réduire la redondance du signal vocal, ou pour extraire des paramètres pertinents pour la reconnaissance de parole. L'estimation des coefficients de la fonction de transfert du conduit vocal est faite en supposant connaître le signal d'excitation. Pour les sons non voisés, le signal d'excitation est un bruit blanc de moyenne nulle et de variance unité. Pour les sons voisés, cette excitation est une suite d'impulsions d'amplitude unité.

Le signal de parole à la sortie du modèle est donc représenté par la somme d'une combinaison linéaire des échantillons précédents et de la fonction d'excitation.

Le modèle de prédiction exploite le fait que les échantillons successifs du signal de parole sont corrélés ; d'où l'intérêt de ce modèle dans le codage de la parole dans le sens où il permet de représenter la parole juste par ses paramètres pertinents, sans redondance. Signalons également que les coefficients sont choisis de façon à minimiser l'erreur quadratique de prédiction sur chaque segment de la fenêtre d'analyse [2].

Des paramètres cepstraux LPCC (Linear Prediction Cepstral Coefficients) sont ensuite calculés à partir de ces coefficients de prédiction linéaire (Linear Prediction Coefficients LPC) [11].

6.4.3. Les paramètres PLP (Perceptual Linear Prediction)

La prédiction linéaire perceptuelle exploite les connaissances du système auditif humain pour paramétrer la parole en introduisant des mécanismes psycho acoustiques de l'oreille humaine.

L'analyse perceptive repose sur l'analyse par prédiction linéaire (LPC). Dans cette analyse, les coefficients sont calculés suite à la résolution d'équations obtenus par la transformée de Fourier inverse du module au carré de la transformée de Fourier du signal.

L'analyse perceptive introduit au niveau fréquentiel des bandes critiques. Cette intégration se fait avec un banc dont les fréquences centrales sont espacées linéairement selon l'échelle Bark. Cette intégration se justifie par le fait que l'oreille se comporte comme un banc de filtres [2].

6.4.4. Delta and Delta-Delta Coefficients

Après avoir extrait les vecteurs de caractéristiques, Il est très important d'avoir une information temporelle concernant les coefficients dérivés de la prédiction linéaire ou de l'analyse cepstral. Ces caractéristiques dynamiques peuvent également être ajoutées aux vecteurs de caractéristiques originales. Les caractéristiques dynamiques donnent des informations sur la façon de variation temporelle des vecteurs caractéristiques. Le $1^{ère}$ et la $2^{ème}$ dérivé dans le temps des vecteurs de caractéristiques (Δ et Δ^2) sont principalement utilisés comme des informations dynamiques [5].

7. **Conclusion**

Dans ce chapitre, nous avons présenté notions générales sur les différents types des systèmes de reconnaissance, à savoir, la reconnaissance de la parole, la reconnaissance du locuteur, et la reconnaissance des téléphones mobiles. Nous avons également cité et définit les méthodes les plus utilisés dans une analyse acoustique.

Enfin, l'extraction des paramètres acoustiques qui est une étape très importante dans les systèmes de reconnaissance a été introduite à travers les paramètres LPCC, PLP et MFCC.

CHAPITRE II : Système de reconnaissance de la parole

1. Introduction

En général, un dispositif de reconnaissance de la parole typique peut être divisé en deux parties principales : le prétraitement et le post traitement. Le prétraitement du système de reconnaissance est un dispositif d'extraction de caractéristiques, tandis que le post traitement se compose d'un modèle/classificateur et d'une base de données bien formée. La Figure II.1 illustre la structure d'un système de reconnaissance typique.

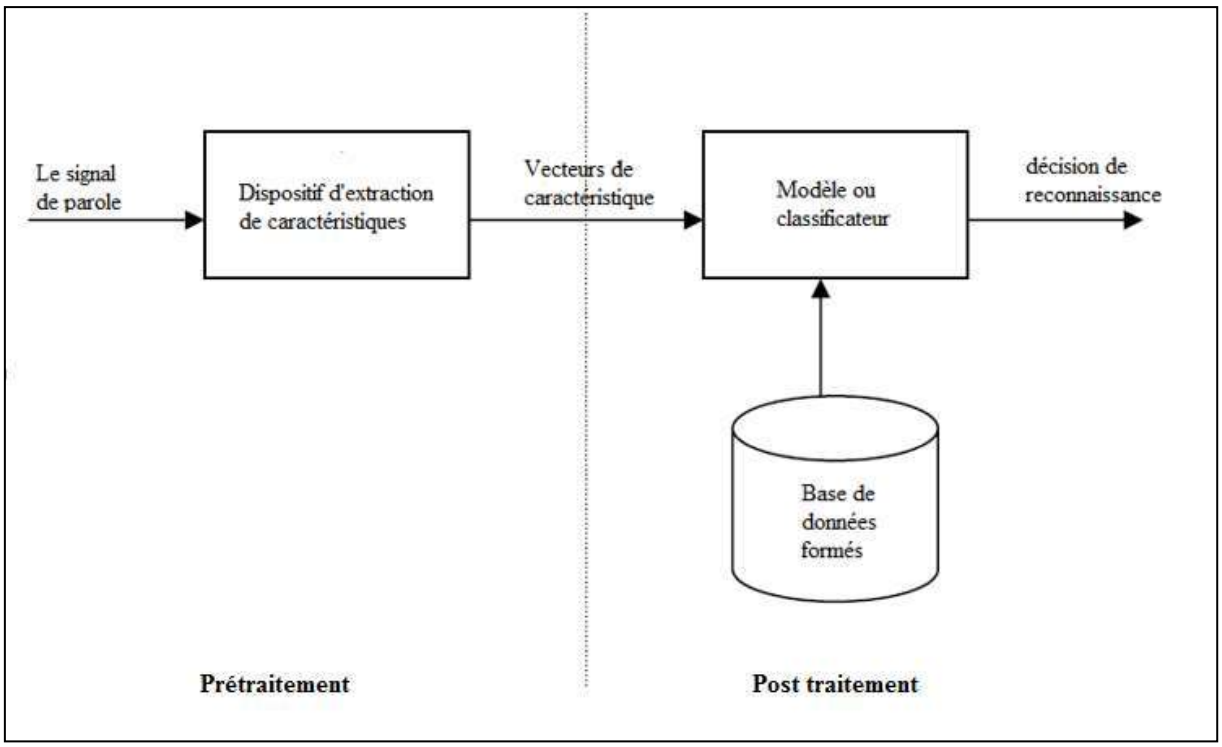


Figure II.1. *Structure d'un système de reconnaissance*

Le but de l'étage d'extraction des paramètres (caractéristiques) est de bien représenter les caractéristiques du signal de parole avec une redondance réduite. Ces caractéristiques sont extraites sur la base de trames (fenêtres). En d'autres termes, un court segment de parole, typiquement de 20ms, est tout d'abord obtenu à partir du signal de parole entier. L'algorithme d'extraction de la parole est appliqué sur un segment pour calculer un certain nombre de coefficients. Ces coefficients sont regroupés pour forme d'un vecteur de caractéristiques. Cette procédure sera répète pour chaque trame du signal parole.

Afin d'avoir une bonne performance de la reconnaissance, l'étage d'extraction de paramètres devrait fournir des vecteurs de caractéristiques qui peuvent capturer les caractéristiques importantes d'un signal.

Il existe deux grands types d'extraction de caractéristiques qui sont couramment utilisés dans dispositifs de reconnaissance de la parole. Ils sont les coefficients Perceptual Linear Prediction (PLP) et Mel Frequency Cepstral Coefficients (MFCC) [12].

De nombreux travaux de recherche ont décrit des expériences, qui testent les performances de différents types de paramètres dans des différentes conditions défavorables. Quand un système de reconnaissance est modélisé et testé dans un bon environnement, le taux reconnaissance peut être facilement bien au-dessus de 90%. Toutefois, lorsque le système est modélisé dans un bon environnement et testé dans un environnement bruyant (15dB SNR par exemple), les performances des deux MFCC et PLP chutent de manière importante à 40-57% [12]. On conclut que ces algorithmes d'extraction de paramètres sont peux efficaces dans un environnement défavorable.

Après l'extraction de caractéristiques, la séquence de vecteurs de caractéristiques est transmise à l'extrémité arrière de la reconnaissance, qui est essentiellement un étage de modélisation en apprentissage ou un classificateur dans l'étage de test.

En apprentissage, il existe trois grands types de modèles qui sont couramment utilisés dans les systèmes de reconnaissance. Le premier type est appelé déformation temporelle dynamique DTW (dynamic time warping), le second type est le modèle de mélange gaussien GMM (Gaussian Mixture Model) et le troisième type est le modèle de machines à vecteurs de support SVM (Support Vector Machine).

2. Méthodes d'extraction des paramètres

Plusieurs approches ont été proposées pour l'analyse de la parole, ayant toutes pour but d'extraire le minimum d'information pouvant définir complètement le signal de parole. Parmi ces approches il y a celles qui agissent dans le domaine fréquentiel, celles qui agissent dans le temps, d'autres permettant d'avoir une analyse conjointe temps-fréquence, et finalement les méthodes basées sur la modélisation du système des phonations.

L'extraction de paramètre est une étape très importante, puisque les performances d'un système de reconnaissance dépendent essentiellement de la qualité des paramètres choisis. La plupart de ces systèmes utilise les coefficients :

- Perceptual Linear Prediction Coefficients (PLP)
- Mel Frequency Cepstral Coefficients (MFCC)

2.1. Coefficients Perceptuelles de prédiction linéaire (PLP)

Traduite de l'anglais Perceptual Linear Prediction(PLP) coefficients est une autre technique d'extraction de caractéristiques, qui tente d'émuler le système auditif humain. Il existe trois principaux concepts derrière la technique PLP.

Dans cette technique, la FFT d'une trame fenêtré de parole est d'abord obtenue. Ensuite, le spectre va subir un banc de filtres Bark, qui modélise la sélectivité critique de fréquence de la bande à l'intérieur de la cochlée humaine. Les sorties de filtre sont pondérées en fonction de la courbe d'isosonie, qui se rapproche de la sensibilité de l'oreille humaine. Des filtres pondérés sont ensuite comprimés selon la loi de puissance sur l'intensité sonore, qui décrit la relation entre l'intensité du signal et de l'intensité sonore perçue. La transformée de Fourier inverse est appliquée aux sorties des filtres et l'analyse prédictive linéaire est effectuée. Enfin, l'analyse cepstrale est appliquée et un ensemble de coefficients PLP est obtenu. La Figure II.2 illustre le schéma de principe de l'algorithme de PLP entier.

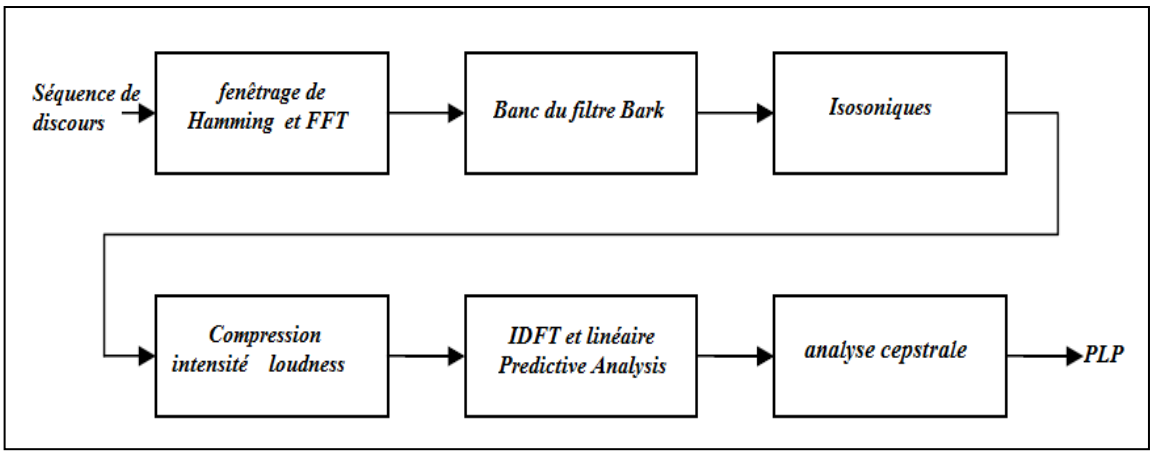


Figure II.2. *Schéma de l'algorithme PLP.*

2.1.1. Fenêtrage de Hamming et FFT

La première étape de l'algorithme consiste à subdiviser le signal de parole en trames en utilisant la fenêtre de Hamming. La Transformée de Fourier Rapide (FFT) est effectuée sur la trame fenêtrée pour obtenir le spectre de la parole.

2.1.2. Banc du filtre Bark

Le banc de filtres dans l'analyse PLP est basée sur une échelle de fréquence non linéaire appelée échelle de Bark. L'équation suivante montre la relation mathématique entre l'échelle Bark et l'échelle de fréquence linéaire,

$$f_{Bark} = 6ln\left(\frac{f}{600} + \left(\left(\frac{f}{600}\right)^2 + 1\right)^{0.5}\right) \quad \textbf{(II.1)}$$

Où f est la fréquence en Hz et f_{Bark} est la fréquence en Bark.

Les fréquences centrales des filtres dans la banque de filtres sont répartir uniformément sur l'échelle de Bark. En outre, Hermansky a suggéré que les fréquences centrales de ces filtres doivent être d'environ 1 Bark. Il a également suggéré que le premier et le dernier filtre doit être situé à 0 Bark et à la fréquence de Nyquist, respectivement, et leurs formes de filtrage sont identiques à leur filtre voisin le plus proche. Ainsi, par exemple, si la fréquence de Nyquist est de 5 kHz (16,9 Barks), le premier et le dernier filtre seront situés à 0 et 16,9 Barks respectivement, avec 16 filtres entre les deux.

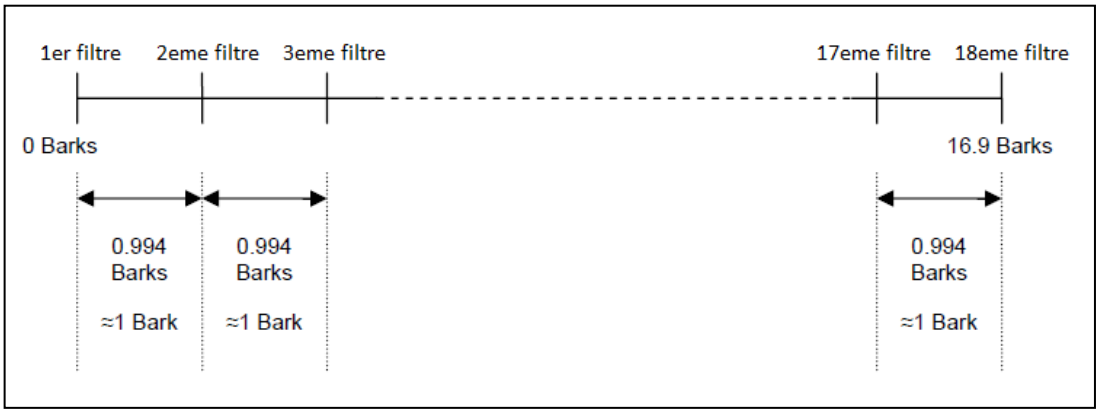


Figure II.3.*Un exemple de la distribution de la fréquence centrale dans un Banc du filtre Bark (fréquence de Nyquist =5 kHz (16,9 Barks))*

Il y aura un total de 18 filtres avec des tailles de pas de 0,994 Barks (Figure II.3). La forme du 1ier filtre à 0 Bark est identique au 2ème filtre, tandis que la forme du dernier filtre

à 16,9 Barks est le même que le 17ème filtre.Les formes des filtres sont identiques sur l'échelle de Bark. L'équation suivante définie la forme du filtre,

$$\psi = \begin{cases} 0 & \text{pour} \quad f_{Bark} - f_{c(Bark)} < -2.5 \\ 10^{(f_{Bark} - f_{c(Bark)} + 0.5)} & \text{pour} \quad -2.5 \le f_{Bark} - f_{c(Bark)} \le -0.5 \\ 1 & \text{pour} \quad -0.5 < f_{Bark} - f_{c(Bark)} < 0.5 \\ 10^{-0.25(f_{Bark} - f_{c(Bark)} + 0.5)} & \text{pour} \quad 0.5 \le f_{Bark} - f_{c(Bark)} \le 1.3 \\ 0 & \text{pour} \quad f_{Bark} - f_{c(Bark)} > 1 \end{cases} \qquad \textbf{(II.2)}$$

Où $f_{c(Bark)}$ est la fréquence centrale d'un filtre en échelle de Bark. ψ Est le poids du filtre à f_{Bark} Barks. La Figure II.4 illustre la banque de filtres en échelle de Bark sur une échelle de fréquence linéaire, avec la fréquence d'échantillonnage de 16 kHz. Notez que, bien que les formes de filtres soient identiques sur l'échelle de Bark, les filtres sont plus larges à la fréquence la plus élevée sur l'échelle de fréquence linéaire. Cela est dû à la non-linéarité entre l'échelle de Bark et l'échelle de fréquence linéaire (équation (II.1)).

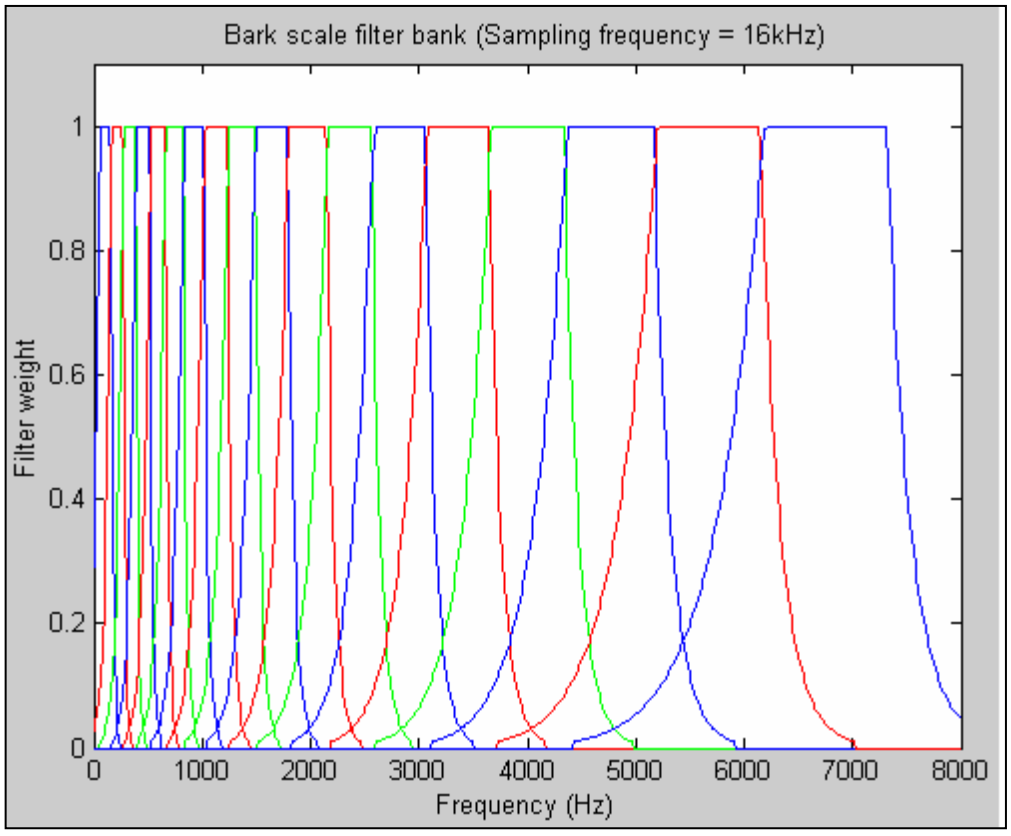


Figure II.4. *Banc du filtre Bark*

La sortie du filtre est définie comme la somme du produit de chacun des points de la FFT sur le spectre de puissance de la parole et les poids des filtres. Mathématiquement,

$$\boldsymbol{X_m = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 |\psi_m[k]|} \qquad \boldsymbol{2 \leq m \leq M-1} \qquad \textbf{(II.3)}$$

où X_m est la sortie de filtre du m^{ieme} filtre. $|S[k]|^2$ Est le spectre de puissance du N-point de la trame de parole fenêtré et $|\psi_m[k]|$ sont les poids de filtrage du m^{eme} filtre, le long de l'échelle de fréquence linéaire discret. On notera que les sorties du premier (m = 1, 0 à Barks) et la dernière (m =M, à la fréquence de Nyquist) les filtres ne sont pas calculés à ce stade. C'est parce que les formes de ces deux filtres seront les mêmes que leurs voisins les plus proches. Les sorties de ces deux filtres sont calculés à l'étape suivante.

2.1.3. Courbe isosonique

La courbe d'isosonie se rapproche de la sensibilité de l'oreille humaine aux différentes fréquences. Il est livré avec deux versions différentes [12]. L'un est pour les applications, dont la fréquence de Nyquist est jusqu'à 5 kHz (équation (II.4)),

$$\mathbf{E} = \frac{(w^2 + 56.8 * 10) w^6}{(w^6 + 6.3 * 10^6)^2 \, (w^2 + 0.38 * 10^9)} \qquad \textbf{(II.4)}$$

Où w est la fréquence angulaire en rad^{-1}, c'est-à-dire $w = 2\pi f$. L'autre version est destinée aux applications dont la fréquence de Nyquist est au-dessus de 5kHz (équation(II.5)).

$$\mathbf{E} = \frac{(w^2 + 56.8 * 10) w^6}{(w^6 + 6.3 * 10^6)^2 (w^2 + 0.38 * 10^9)(w^6 + 9.58 * 10^{26})} \qquad \textbf{(II.5)}$$

Le poids d'isosonie E, de chaque filtre peut être calculé en remplaçant sa fréquence centrale (sous forme de pulsation) dans l'une des deux équations ci-dessus, en fonction de la fréquence de Nyquist.

Chaque sortie du filtre dans la banque de filtre est pondérée par le poids d'isosonie correspondant.

$$\boldsymbol{X_{m(e)} = E_m X_m} \qquad \boldsymbol{2 \leq m \leq M-1} \qquad \textbf{(II.6)}$$

$X_{m(e)}$ est la sortie du m^{eme} filtre après l'avoir multiplié par le poids d'isosonie, E_m.

Alternativement, les poids isosoniques peuvent être implicitement appliqués aux poids des filtres (équation(II.7)). Par conséquent, les résultats de filtrage peuvent être obtenus directement à partir des poids de filtre modifiés (équation (II.8)).

$$\boldsymbol{\psi_{m(e)}[k] = E_m \psi_m[k] 2 \leq m \leq M-1} \qquad \textbf{(II.7)}$$

$$\boldsymbol{X_{m(e)} = \sum_{k=0}^{\frac{N}{2}-1} |S[k]|^2 \left|\psi_{m(e)}[k]\right|} \qquad \boldsymbol{2 \leq m \leq M-1} \qquad \textbf{(II.8)}$$

Puisque, les formes du premier filtre (à 0 Bark) et le dernier filtre (à la fréquence de Nyquist) sont identiques à leurs plus proches voisins, la première et la dernière des sorties de filtres sont rendus égaux à leur sortie la plus proche du filtre, c'est-à-dire,

$$\boldsymbol{X_{1(e)} = X_{2(e)}}$$
$$\boldsymbol{X_{M-1(e)} = X_{M(e)}} \quad \textbf{(II.9)}$$

2.1.4. Compression intensité-volume

La compression intensité-volume se rapproche de la perceptibilité de l'homme sur l'intensité sonore. Elle décrit la relation non linéaire entre l'intensité du signal et de l'intensité sonore perçue. Dans la technique PLP, cette relation est une relation de racine cubique, et elle peut être exprimée par la formule suivante [12],

$$\boldsymbol{\phi_m = \left(X_{m(e)}\right)^{0.33} \quad 1 \leq m \leq M} \quad \textbf{(II.10)}$$

Où φ_m est la sortie du filtre correspondant après compression.

2.1.5. L'IDFT et l'analyse prédictive linéaire

Il s'agit d'effectuer une analyse par prédiction linéaire. Tout d'abord, il est nécessaire de trouver la fonction d'autocorrélation des sorties de filtre.

Il existe de nombreuses méthodes pour évaluer les coefficients PLP et le gain du filtre. Parmi ces méthodes, on cite la méthode d'autocorrélation. Elle implique le calcul d'une matrice d'équations simultanées et l'autocorrélation des trames de parole en fenêtres. L'équation suivante illustre la matrice d'équations qui doivent être résolus,

$$\begin{bmatrix} R[0] & R[1] & \dots & R[p-1] \\ R[1] & R[0] & \dots & \dots \\ \dots & \dots & \dots & R[1] \\ R[p-1] & \dots & R[1] & R[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix} = \begin{bmatrix} R[1] \\ R[2] \\ \dots \\ R[p] \end{bmatrix} \quad \textbf{(II.11)}$$

Où $R[n]$ est la fonction d'autocorrélation d'une trame de parole fenêtré.

Le gain du filtre tous pôles peut être trouvé en résolvant l'équation suivante.

$$G = \sqrt{R[0] - \sum_{k=1}^{p} a_k R[k]} \quad \textbf{(II.12)}$$

Une transformée de Fourier discrète inverse (IDFT) est appliquée à la sortie du filtre φ_m. Notez que φ_m n'est que la moitié de l'ensemble du spectre (à la fréquence de Nyquist). En outre, le premier et le dernier sont réalisés pour qu'ils soient égaux à leurs voisins les plus proches. Par conséquent, la totalité du spectre est obtenue en ajoutant l'image en miroir de la première moitié en excluant la première et de la dernière φ_m. En plus simple, la totalité du spectre peut être exprimé par le vecteur suivant,

$$\Phi = [\phi_1 \phi_2 \ \ldots \ldots \phi_{M-1} \phi_M \phi_{M-1} \ \ldots \ \phi_3 \phi_2] \quad \textbf{(II.13)}$$

IDFT est appliquée sur le vecteur Φ et les premiers coefficients p+1 sont considérés comme fonction d'autocorrélation R[n] pour 0 ≤ n ≤ p, où p est l'ordre de la prédiction linéaire. La fonction d'autocorrélation est utilisée pour évaluer les coefficients de prédiction linéaire en résolvant les équations (II.14) - (II.15).

Etant donné que la matrice située à gauche de l'équation (II.11) est une matrice de Toeplitz, des algorithmes récursifs efficaces peuvent être utilisés pour résoudre l'équation ci-dessus. Un algorithme appelé procédure récursive de Durbin peut être utilisé, il est représentée par le jeu d'équations suivant,

$$E^0 = R[0]$$

$$k_i = \frac{R[i] - \sum_{j=1}^{i-1} a_j^{(i-1)} R[i-j]}{E^{i-1}} \, 1 \le i \le p$$

$$a_i^{(i)} = k_i \quad \textbf{(II.14)}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)} \, 1 \le j \le i-1$$

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

La résolution de ces équations se fait de manière récursive pour $i = 1, 2, \ldots, p$. Quand i atteint la p^{ième} itération, l'ensemble des coefficients est donnés par :

$$a_{j=} a_j^{(p)} \quad \textbf{(II.15)}$$

2.1.6. L'analyse cepstrale

L'analyse cepstrale est appliquée aux coefficients de prédiction linéaire obtenus à l'étape précédente (section 2.1.5).

L'analyse cepstrale se réfère au processus de recherche de cepstre d'une séquence de parole. Cepstre, dont l'orthographe est formée en mélangeant les caractères du mot «spectre», est une représentation dans le domaine temporel d'un signal donné. Cependant, un terme spécial de «quéfrence », qui est également formé en mélangeant l'orthographe du mot «fréquence», est souvent utilisé à la place de "temps". Le cepstre est défini comme étant la transformée de Fourier inverse du logarithme du spectre d'un signal [12]. L'équation suivante donne la définition du cepstre d'un signal,

$$\hat{\boldsymbol{s}}[\boldsymbol{n}] = \frac{\mathbf{1}}{\mathbf{2\pi}} \int_{-\pi}^{\pi} \boldsymbol{ln}[\boldsymbol{S}(\boldsymbol{w})] \boldsymbol{e}^{\boldsymbol{i\omega n}} \boldsymbol{dw} \qquad \textbf{(II.17)}$$

où $S(w)$ est la transformée de Fourier d'un signal, dans notre cas une trame de parole, ŝ[n] est le cepstre.

Dans le domaine de la recherche de la parole, la principale utilisation de l'analyse cepstrale est d'extraire les caractéristiques du conduit vocal à partir d'un spectre vocal. Il peut être démontré qu'une séquence de parole peut être modélisée par la convolution de la réponse impulsionnelle du conduit vocal et l'excitation glottale,

$$\boldsymbol{s}[\boldsymbol{n}] = \boldsymbol{v}[\boldsymbol{n}] * \boldsymbol{u}[\boldsymbol{n}] \qquad \textbf{(II.18)}$$

où $v[n]$ est la réponse impulsionnelle du filtre de conduit vocal. $u[n]$ est l'excitation glottique, qui est habituellement une séquence d'impulsions quasi-périodique de la parole voisée. Dans le domaine fréquentiel, l'équation ci-dessus peut s'exprimer comme :

$$\boldsymbol{S}(\boldsymbol{w}) = \boldsymbol{V}(\boldsymbol{w})\boldsymbol{U}(\boldsymbol{w}) \qquad \textbf{(II.19)}$$

Comme représenté par l'équation II.18, le logarithme à l'intérieur de l'intégrale divise le produit de deux spectres en une sommation. Etant donné que l'intégration est une opération linéaire, le cepstre de la séquence de parole est en fait la somme du cepstre du conduit vocal et l'excitation glottale.

$$\hat{\boldsymbol{s}}[\boldsymbol{n}] = \hat{\boldsymbol{v}}[\boldsymbol{n}] + \hat{\boldsymbol{u}}[\boldsymbol{n}] \qquad \textbf{(II.20)}$$

Où $\hat{s}[n]$, $\hat{v}[n]$ et $\hat{u}[n]$ sont le cepstre de la séquence vocale, la réponse impulsionnelle et l'excitation glottale respectivement.

Le cepstre possède deux propriétés importantes : (1) le cepstre d'une séquence périodique est également une séquence périodique, (2) le cepstre d'une réponse arbitraire impulsionnelle du filtre est une séquence de décomposition, qui tend vers zéro quand la quéfrence n tend vers l'infini. Etant donné que v[n] est la réponse impulsionnelle du filtre de l'appareil vocal et u[n] est une séquence de quasi-périodique, il peut être démontré que la partie basse quéfrence de $\hat{s}[n]$ appartient à $\hat{v}[n]$, tandis que la partie haute quéfrence est principalement constitué de $\hat{u}[n]$. Il s'agit d'une pratique courante de prendre les premiers coefficients de la partie basse quéfrence du cepstre de la parole, car il contient les caractéristiques de l'appareil vocal. Ces coefficients sont appelés coefficients cepstraux.

Habituellement, afin d'obtenir les coefficients cepstraux, il est nécessaire d'évaluer l'équation II.17. Toutefois, les coefficients cepstraux peuvent également être calculés par l'intermédiaire d'un ensemble de procédure récursive [12]. Les coefficients cepstraux obtenus de cette manière sont appelés des coefficients PLP cepstraux. Les équations suivantes montrent la procédure récursive,

$$\hat{v}[n] = \ln(G) \qquad \text{Pour n=0}$$

$$\hat{v}[n] = a_n + \sum_{k=1}^{n-1} \left(\frac{k}{n}\right) \hat{v}[k]\, a_{n-k} \qquad \text{Pour } 1 \leq n \leq p \qquad \textbf{(II.21)}$$

Les coefficients de prédiction linéaire deviennent des coefficients PLP cepstraux. L'ordre des coefficients cepstraux PLP est le même que les coefficients de prédiction linéaire. Il y aura un nombre p de coefficients cepstraux PLP générées pour chaque trame de parole. Ces coefficients sont regroupés pour constituer un vecteur de caractéristiques.

Souvent, l'erreur de prédiction log de la prédiction linéaire est jointe au vecteur de caractéristique. L'erreur de prédiction log étant le logarithme du gain de filtre de tous les pôles au carré,

$$PE_{(ln)} = \ln(G^2) = \ln\left(R[0] - \sum_{k=1}^{P} a_k R[k]\right) \qquad \textbf{(II.22)}$$

où $PE_{(ln)}$ est l'erreur de prédiction log et G est le gain du filtre du modèle de filtre tout pôle de prédiction linéaire.

2.2. Les Coefficients Mel Fréquences Cepstraux (MFCC)

Traduite de l'anglais "Mel Fréquency Cepstral Coefficients (MFCC)", c'est l'une des techniques les plus couramment utilisées dans les systèmes de reconnaissance. Les étapes utilisées dans cette technique peuvent être vues dans la Figure II.5.

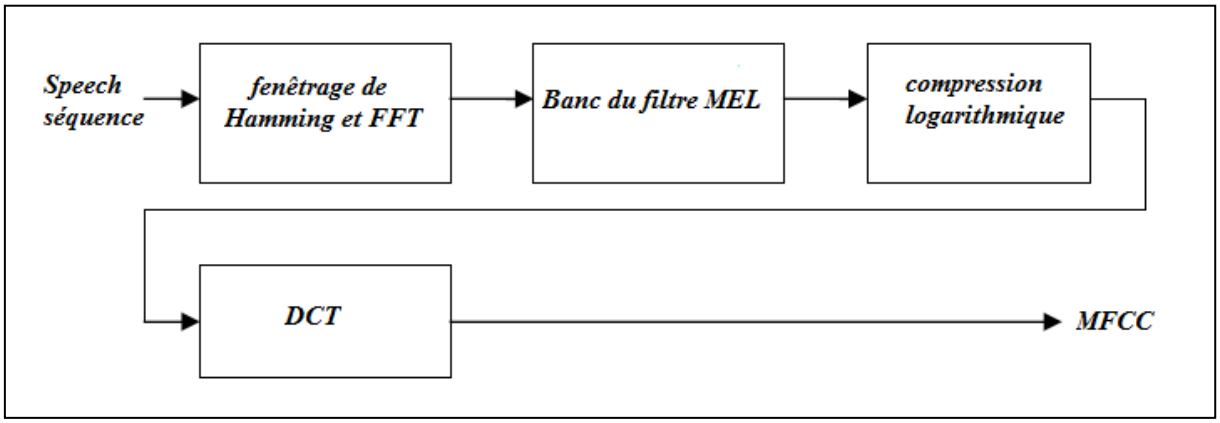


Figure II.5. *Illustration des étapes de l'algorithme MFCC*

2.2.1. Banc du filtre MEL

Le banc de filtres de fréquence Mel est une série de filtres passe-bande triangulaires. Le banc de filtres est basée sur une échelle de fréquence non linéaire appelée échelle Mel. L'échelle de fréquence Mel est une mesure psycho acoustique d'emplacements jugés par des humains. Selon [12], une tonalité de 1000 Hz, avec 40dB au-dessus du seuil de l'auditeur, est définie comme ayant une hauteur de 1000 Mels. Ou-dessous 1000 Hz, l'échelle de Mel est approximativement linéaire à l'échelle de fréquence linéaire. Au-dessus du point de référence de 1000 Hz, les auditeurs ont tendance à percevoir les mêmes incréments de hauteur avec des intervalles plus longs de fréquence. D'où la relation entre l'échelle de Mel et de l'échelle de fréquence linéaire est non linéaire et approximativement logarithmique au-dessus de 1000Hz. L'équation suivante décrit la relation mathématique entre l'échelle de Mel et l'échelle de fréquence linéaire,

$$f_{Mel} = 1127.01\, ln\left(\frac{f_{Hz}}{700} + 1\right) \qquad \textbf{(II.23)}$$

Où f_{Mel} est la fréquence en Mels, et f_{Hz}est la fréquence linéaire en Hz.

Comme mentionné précédemment, le banc de filtres de fréquence Mel se compose d'une série de filtres passe-bande triangulaires. Les filtres sont superposés de manière à ce que la

limite inférieure d'un filtre est situé à la fréquence centrale du filtre précédent, et la limite supérieure est située à la fréquence centrale du filtre suivant. La réponse maximale du filtre, c'est-à-dire le sommet supérieur du filtre triangulaire, est situé à la fréquence centrale du filtre et est normalisé à l'unité.

Les fréquences centrales de la série de filtres sont réparties le long de l'échelle de fréquence Mel. On peut montrer que la fréquence centrale du filtre dans le m^{eme}groupe de filtres peut être trouvée par l'évaluation de l'équation suivante,

$$\boldsymbol{f_{cm(Mel)} = f_{l(Mel)} + \frac{m(f_{H(Mel)} - f_{L(Mel)})}{M+1}} \qquad \textbf{(II.24)}$$

Où $f_{cm(Mel)}$ est la fréquence centrale du m^{eme} filtre dans l'échelle des Mels. $f_{L(Mel)}$ et $f_{H(Mel)}$ sont les extrémités inférieures et supérieures en Mels, respectivement, de la gamme de fréquence couverte par l'ensemble du banc de filtres. Il existe des filtres triangulaires m à l'intérieur de la plage comprise entre $f_{L(Mel)}$ et $f_{H(Mel)}$. La Figure suivante montre les réponses en fréquence d'une série de filtres passe-bande triangulaires sur une échelle de fréquence linéaire.

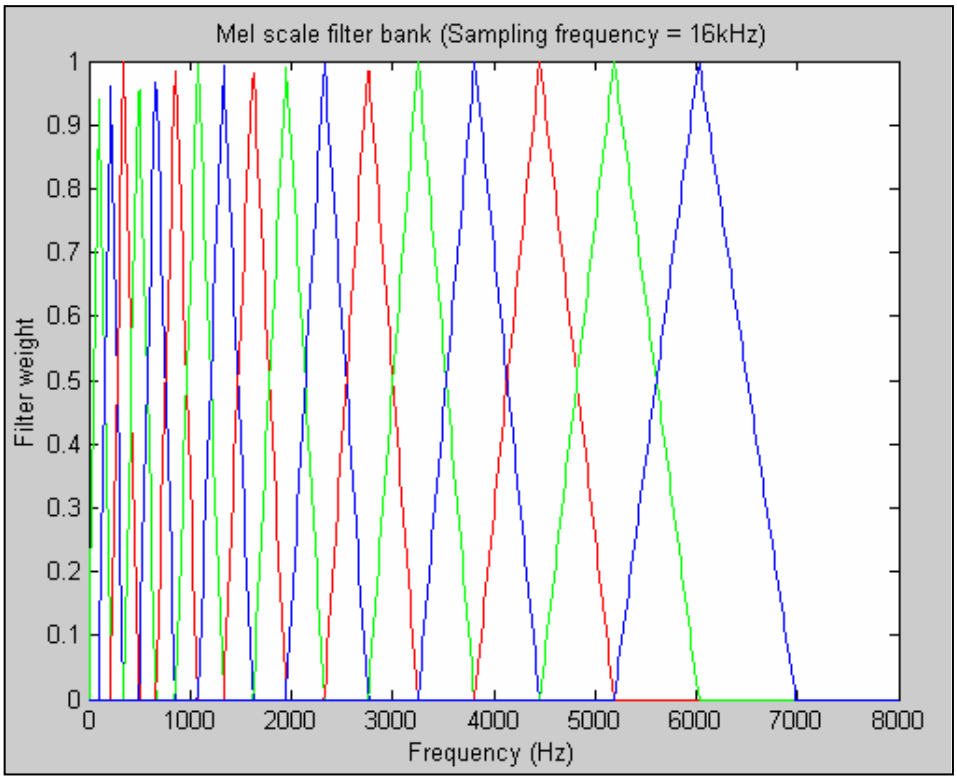


Figure II.6. *Le banc du filtre MEL (Fréquence d'échantillonnage =16 kHz)*

Supposons que $|H_m[k]|$ dénote la réponse d'amplitude en fréquence du m^{eme} filtre, où k est l'indice de fréquence discrète dans le domaine numérique. Le signal de sortie de filtre du m^{eme} filtre, X_m peut être exprimé par l'équation suivante,

$$\boldsymbol{X_m = \sum_{k=o}^{\frac{N}{2}-1} |S[k]|^2 \, |H_m[k]|} \qquad \textbf{(II.25)}$$

Où $S[k]$ est la FFT de la trame de parole fenêtré composée de N points. Notez que le nombre de points de sommation est juste $N/2$. C'est parce que la moitié du spectre FFT est une image miroir de l'autre moitié.

2.2.2. Compression logarithmique

Afin de modéliser le volume sonore perçu d'une intensité de signal donnée, les sorties de filtre sont comprimées par une fonction logarithmique (équation(II.26)).

$$\boldsymbol{X_{m(\ln)} = ln(X_m) 1 \leq m \leq M} \qquad \textbf{(II.26)}$$

Dans l'équation ci-dessus, $X_{m(\ln)}$ est la sortie logarithmique comprimé du m^{eme} filtre.

2.2.3. DCT

L'étape finale de l'algorithme est de décarreler les sorties de filtre. Transformée en cosinus discrète (DCT) est appliquée sur les sorties de filtre et les premiers coefficients sont groupés ensemble en tant que vecteur de caractéristique d'une trame de parole donnée.

Supposons que p est l'ordre du cepstre en échelle Mel. Le vecteur caractéristique est obtenu en considérant les premiers coefficients de p DCT. Mathématiquement, le k^{eme} coefficient MFCC peut être exprimé par la formule suivante.

$$\boldsymbol{MFCC_K = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} X_{m(\ln)} \cos\left(\frac{\pi k(m-k)}{M}\right), 1 \leq k \leq p} \qquad \textbf{(II.27)}$$

Le vecteur de caractéristique statique est souvent accolé soit par une composante d'énergie de log ou un coefficient d'ordre zéro ou les deux.

Le coefficient de MFCC d'ordre zéro correspond à l'ordre $zero^{eme}$ coefficient DCT des sorties de filtre. L'équation suivante est l'expression du coefficient d'ordre zéro de MFCC.

$$\boldsymbol{MFCC_0 = \sqrt{\frac{1}{M}} \sum_{m=1}^{M} X_{m(\ln)}} \qquad \textbf{(II.28)}$$

3. Méthodes de modélisation

Le module de modélisation apprend des modèles d'unités acoustiques ou décode le signal d'entrée à partir d'un jeux de paramètres [13].

Afin de pouvoir effectuer des mesures de similarité, les paramètres extraits doivent être modélisés statistiquement. Il existe plusieurs méthodes de modélisation. La discrimination par la valeur moyenne, l'alignement temporel par programmation dynamique, la représentation par quantification vectorielle ou encore la modélisation par modèles de Markov cachés ne sont que des exemples [14].

3.1. Déformation temporelle dynamique DTW (dynamic time warping)

Cette méthode est la solution la plus simple à mettre en œuvre en reconnaissance de la parole. Elle consiste à effectuer une comparaison dynamique entre un vecteur de référence et un vecteur de test. On cherche à apparier l'ensemble des points du premier vecteur avec le second. On associe un poids à chaque appariement possible. L'utilisation de la programmation dynamique permet de trouver l'appariement minimisant ce coût. Les formes références sont obtenues par extraction des caractéristiques. La forme test peut avoir subi des transformations (rotations, translation, changement de taille). On utilise pour cela une méthode de mesure de similarité. La distance finale entre une forme de référence et une forme de test est calculée comme la somme des distances partielles entre vecteurs de paramètres le long du chemin optimal. Des poids liés à la suppression et à l'insertion d'informations, sont imposés pour calculer le meilleur chemin.

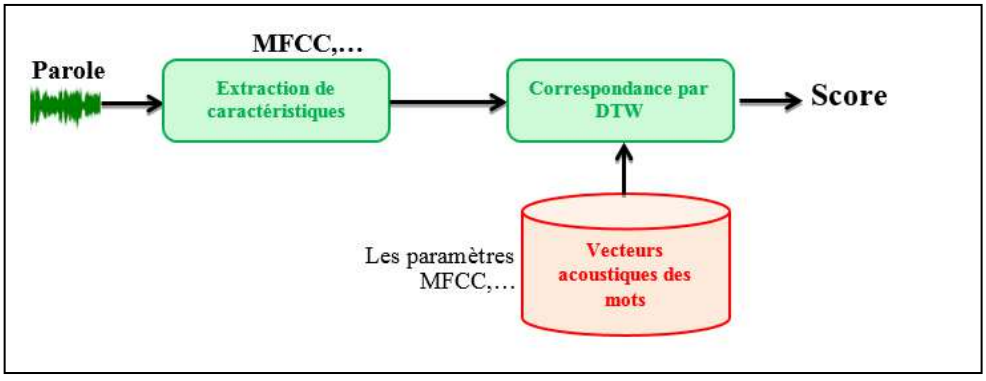


Figure II.7. *Schéma d'un système de reconnaissance basé sur la comparaison dynamique*

Cette méthode a l'avantage d'être simple et de nécessiter que peu de données d'apprentissage. Par contre, elle est limitée par la taille du vocabulaire à reconnaître et à la reconnaissance mono locuteur. [2]

3.2. Modèle de mélange gaussien GMM (Gaussian Mixture Model)

Un modèle de mélange gaussien (usuellement abrégé par l'acronyme anglais GMM pour Gaussian Mixture Model) est un modèle statistique exprimé selon un mélange de densités. Elle sert usuellement à estimer paramétriquement la distribution de variables aléatoires en les modélisant comme une somme de plusieurs gaussiennes (appelées noyaux). Il s'agit alors de déterminer la variance, la moyenne et l'amplitude de chaque gaussienne. Ces paramètres sont optimisés selon un critère de maximum de vraisemblance pour approcher le plus possible la distribution recherchée. Cette procédure se fait le plus souvent itérativement via l'algorithme espérance-maximisation (EM).

Les modèles de mélange gaussien sont réputés reconstruire de manière particulièrement efficace les données manquantes dans un jeu de données expérimentales.

Dans les modèles de mélanges, fréquemment utilisés en classification automatique, on considère qu'un échantillon de données suit, non pas une loi de probabilité usuelle, mais une loi dont la fonction de densité est une densité mélange.

Bien que n'importe quelle loi puisse être utilisée, la plus courante est la loi normale dont la fonction de densité est une gaussienne. On parle alors de mélange gaussien [16].

3.3. Machines à vecteurs de support SVM (Support Vector Machine)

Les SVM sont des classificateurs binaires et statiques. C'est à dire qu'ils permettent de créer une surface de décision entre deux classes définies dans un même espace. Pour cela, ils construisent une frontière de décision par projection des caractéristiques provenant d'un espace d'origine dans un espace de caractéristiques de dimension supérieure (voir infini) dans le but de rendre les classes linéairement séparables.

L'hyperplan choisi est celui qui maximise la marge de séparabilité entre les deux ensembles de données. La sélection de l'hyperplan dans un espace de caractéristiques nécessite d'évaluer un produit scalaire dans cet espace. Ce qui peut être très coûteux en temps

et en complexité si l'espace est de très grande dimension. Heureusement, ce calcul n'est pas obligatoire grâce à une opération mathématique appelé kernel. Le kernel calcule le produit scalaire de deux points dans l'espace de dimension supérieur sans avoir à les projeter. [2]

Pour la reconnaissance des formes (cas de deux classes d'échantillons), les SVM recherchent une surface de décision optimale, déterminée par certains points de l'ensemble d'apprentissage appelés vecteurs support, en projetant les données d'entrée non-linéairement séparables dans un espace de plus grande dimension appelé espace de caractéristiques (Figure 0II.8). Cette surface, qui est dans l'espace des caractéristiques, peut être considérée comme un hyperplan optimal de décision.

Elle est obtenue par la résolution d'un problème de programmation quadratique dépendant des paramètres de régularisation.

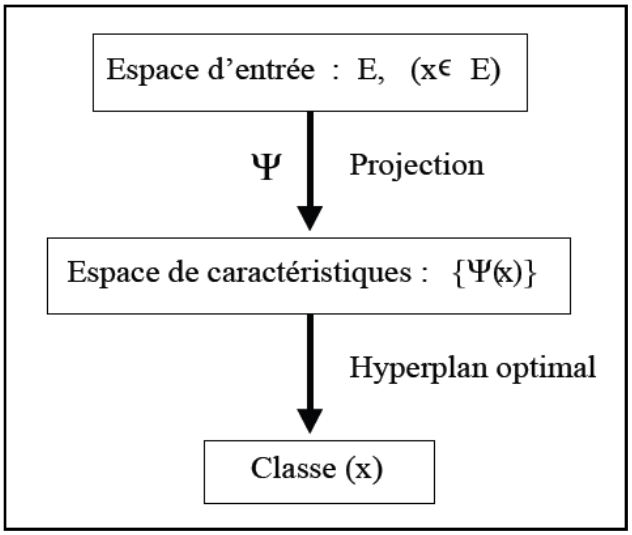


Figure II.8. *Principe des SVM*

Les systèmes d'apprentissage SVM, sont des algorithmes basés sur les trois principes mathématiques suivants [17]:

- Le principe de Fermat (1638) : Les points qui minimisent ou maximisent une fonction dérivable annulent sa dérivée. Ils sont appelés points stationnaires.
- Le principe de Lagrange (1788) : Pour résoudre un problème d'optimisation sous contraintes, il suffit de rechercher un point stationnaire x^0 du Lagrangien L de la fonction f à optimiser

$$\mathbf{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{f}(\mathbf{x}) + \sum_{\mathbf{i}=\mathbf{0}}^{\mathbf{K}} \boldsymbol{\lambda}_{\mathbf{i}} + \mathbf{f}_{\mathbf{i}}(\mathbf{x}) \qquad \textbf{(II.29)}$$

Où les f_i expriment les contraintes et les λ_i sont des constantes appelées coefficients de Lagrange.

- Le principe de Kuhn-Tucker (1951) : Les relations de Kuhn-Tucker peuvent s'appliquer en utilisant les fonctions convexes f et f_i . Il est toujours possible de trouver un point-selle (x^0, λ^*) qui vérifie l'équation suivante :

$$\mathbf{min}_x \, \boldsymbol{L}(\boldsymbol{x}, \boldsymbol{\lambda}^*) = \boldsymbol{L}(\boldsymbol{x}^0, \boldsymbol{\lambda}^*) = \mathbf{max}_{\boldsymbol{\lambda} \geq 0} \, \boldsymbol{L}(\boldsymbol{x}^0, \boldsymbol{\lambda}) \qquad \textbf{(II.30)}$$

3.3.1. Cas des données linéairement séparables

Soit une base d'exemples, $S = \{(x_i, y_i), x_i \in \Re^n; y_i \in \{-1,1\}; i = 1, \dots, m\}$, linéairement séparables.

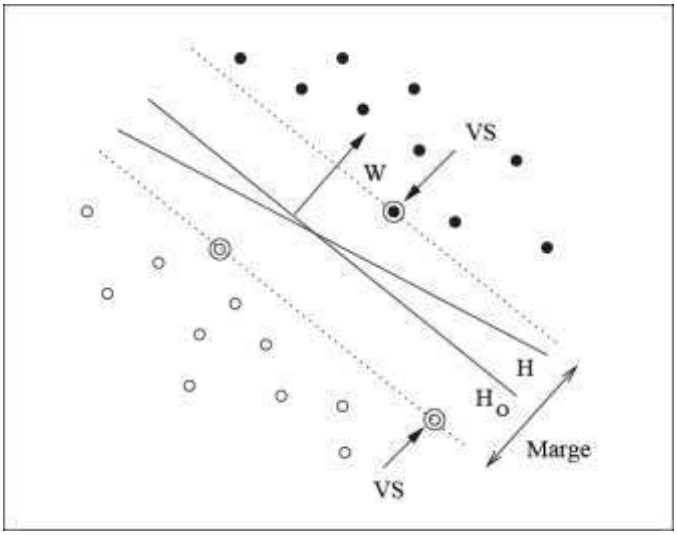


Figure II.9. *Hyperplans séparateurs : H est un hyperplan quelconque,* $\boldsymbol{H_o}$ *est l'hyperplan optimal, VS : sont les Vecteurs Support.*

Soit H : (w, b) un hyperplan séparateur qui est l'un des hyperplans de la Figure II.9 et qui vérifie les conditions suivantes :

$$\boldsymbol{w}.\boldsymbol{x_i} + \boldsymbol{b} \geq \mathbf{1} \quad \boldsymbol{si} \quad \boldsymbol{y_i} = \mathbf{1} \qquad \textbf{(II.31)}$$

$$\boldsymbol{w}.\boldsymbol{x_i} + \boldsymbol{b} \leq -\mathbf{1} \quad \boldsymbol{si} \quad \boldsymbol{y_i} = -\mathbf{1} \qquad \textbf{(II.32)}$$

Ce qui est équivalent à :

$$\boldsymbol{y_i}(\boldsymbol{w}.\boldsymbol{x_i} + \boldsymbol{b}) \geq \mathbf{1} \quad \forall_i \in \{\mathbf{1}, \dots, \boldsymbol{m}\} \qquad \textbf{(II.33)}$$

La distance $d(w, b, x)$ entre un point x et l'hyperplan (w, b) est donnée par :

$$\boldsymbol{d(w,b,x) = \frac{|w.x + b|}{\|w\|}}$$

On appelle **hyperplan optimal**, l'hyperplan séparateur qui se situe à la distance maximale des vecteurs x les plus proches parmi l'ensemble des exemples. Cet hyperplan **maximise la marge**.

L'hyperplan optimal est donné par la maximisation de la marge $M(w,b)$, sous les contraintes de l'équation (II.33). La marge $M(w,b)$ s'écrit sous la forme suivante [17]:

$$\begin{aligned}\mathbf{M(w,b)} &= \mathbf{min_{x_i:y_{i=-1}}} \mathbf{d(w,b,x_i)} + \mathbf{min_{x_i:y_{i=1}}} \mathbf{d(w,b,x_i)} \\ &= \mathbf{min_{x_i:y_{i=-1}}} \frac{\mathbf{|w.x_i + b|}}{\mathbf{\|w\|}} + \mathbf{min_{x_i:y_{i=1}}} \frac{\mathbf{|w.x_i + b|}}{\mathbf{\|w\|}} \\ &= \frac{\mathbf{1}}{\mathbf{\|w\|}} \left[\mathbf{min_{x_i:y_{i=-1}}} \mathbf{|w.x_i + b|} + \mathbf{min_{x_i:y_{i=1}}} \mathbf{|w.x_i + b|}\right] \\ &= \frac{\mathbf{1}}{\mathbf{\|w\|}} \mathbf{[1+1]} = \frac{\mathbf{2}}{\mathbf{\|w\|}}\end{aligned}$$

Plus la marge est grande, plus l'erreur attendue est faible. Maximiser la marge $M(w,b)$, revient à minimiser le carré de la norme $||w||^2 = (w,w)$ du vecteur w, sous les contraintes de l'équation (II.33) [17]:

Ainsi l'hyperplan optimal qui vérifie les conditions de l'équation (II.33) est celui qui minimise la fonction définie par :

$$\boldsymbol{\emptyset(w) = \frac{1}{2} w.w} \qquad \textbf{(II.34)}$$

En appliquant le principe de Kuhn-Tucker, nous sommes menés à la recherche d'un point-selle w^0, b^0, λ^0 du Lagrangien :

$$\boldsymbol{L(w,b,\lambda) = \frac{1}{2} w.w - \sum_{i=1}^{m} \lambda_i [y_i(w.x_i + b) - 1]} \qquad \textbf{(II.35)}$$

Ou $\lambda = (\lambda_1, \dots, \lambda_m)$

Ce point-selle vérifie les conditions :

$$\boldsymbol{\lambda_i^0 [y_i(w^0.x_i + b^0) - 1] = 0, \quad \text{pour tout } i \in \{1, \dots, m\}} \qquad \textbf{(II.36)}$$

Les vecteurs supports sont les vecteurs x_i pour lesquels l'égalité $y_i\ (w^0 . x_i\ +\ b^0\ ) = 1$ est vérifiée. Concrètement, ce sont les vecteurs les plus proches de l'hyperplan optimal. Pour tous les autres exemples qui vérifient l'équation II.36, nous avons donc $\lambda_i^0 = 0$.

Les conditions d'annulation des dérivées partielles du Lagrangien permettent d'écrire les relations que vérifie l'hyperplan optimal, avec λ_i^0 non nuls seulement pour les vecteurs supports. $L(w, b, \lambda)$ Est minimal par rapport à b :

$$\frac{\delta L(w,b,\lambda)}{\delta b} = 0 \Leftrightarrow \sum_{i=1}^{m} \lambda_i^0 \, y_i = 0 \qquad \textbf{(II.37)}$$

$L(w, b, \lambda)$ Est minimal par rapport à w :

$$\frac{\delta L(w,b,\lambda)}{\delta b} = 0 \Leftrightarrow w^0 = \sum_{i=1}^{m} \lambda_i^0 x_i \, y_i = 0 \qquad \textbf{(II.38)}$$

Ces contraintes d'égalité permettent d'exprimer la forme duale du Lagrangien :

$$F(\lambda) = \sum_{i=1}^{m} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{m} \lambda_i \lambda_j y_i y_j (x_i . x_j) \qquad \textbf{(II.39)}$$

Pour trouver le point-selle, il suffit alors de maximiser (et non plus de minimiser) $F(\lambda)$ dans le quadrant positif $\lambda_{\mathrm{i}} \geq 0$, pour tout $i \in \{1, \ldots, m\}$ sous la contrainte :

$$\sum_{i=1}^{m} \lambda_i^0 \, y_i = 0 \qquad \textbf{(II.40)}$$

La résolution de l'équation 3.14, nous donne les solutions suivantes :

$$\lambda_i^0 = 0 \ \textbf{Ou} \ y_i(w^0 . x_i + b^0) = 1 \ \textbf{avec} \ \lambda_i^0 > 0 \qquad \textbf{(II.41)}$$

Pour calculer le biais b^0 nous avons utilisé l'équation :

$$y_i(w^0 . x_i + b^0) = 1 \ \textbf{Avec} \ \lambda_i^0 > 0$$

Pour un vecteur support de la classe +1 :

$$w^0 . vs_{(classe+1)} + b^0 = 1$$

Pour un vecteur support de la classe -1 :

$$w^0 . vs_{(classe-1)} + b^0 = -1$$

D'après ces deux équations nous obtenons la valeur du biais b_0.

$$b^0 = \frac{1}{2}([w^0 . vs_{(classe+1)}] + [w^0 . vs_{(classe-1)}]) \qquad \textbf{(II.42)}$$

Ainsi, la fonction de classification classe(x) s'écrit sous la forme suivante :

$$classe(x) = sign(w^0 . x_i + b^0)$$

$$classe(x) = sign\left[\left(\sum_{x_i \in V \, S} \lambda_i^0 \, y_i x_i\right) . x + b^0\right]$$

$$\boldsymbol{classe(x) = sign\left[\left(\sum_{x_i \in VS} \lambda_i^0 \, y_i . x\right).x + b^0\right]} \quad \textbf{(II.43)}$$

Si la fonction $classe(x)$ est négative alors x appartient à la classe -1 sinon x appartient à la classe +1.

3.3.2. Cas des données non-linéairement séparables

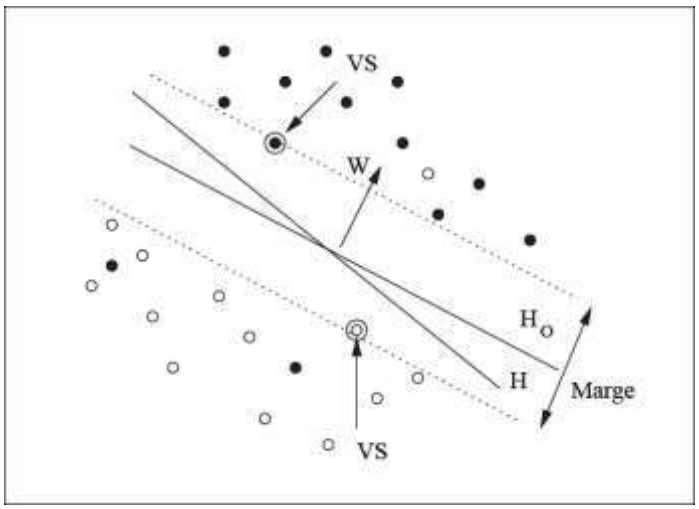


Figure II.10. *Hyperplans séparateurs dans le cas de données non-linéairement séparables : H est un hyperplan quelconque,* $\boldsymbol{H_0}$ *est l'hyperplan optimal, VS : sont les Vecteurs Support.*

Dans ce cas les exemples ne sont pas linéairement séparables (Figure II.10) et l'hyperplan optimal sera celui qui vérifie les contraintes suivantes :

La distance entre l'hyperplan et les vecteurs bien classés doit être maximale.
La distance entre l'hyperplan et les vecteurs mal classés doit être minimale.

Pour résoudre ce problème, on peut reformuler les conditions des équations II.31 et II.32, en introduisant des variables positives ξ_i, pour $i = 1, \ldots, m$. Nous obtenons les inégalités suivantes :

$$\boldsymbol{(w.x_i) + b \geq +1 - \xi_i \;\; \text{Si} \;\; y_i = +1} \quad \textbf{(II.44)}$$

$$\boldsymbol{(w.x_i) + b \leq -1 + \xi_i \;\; \text{Si} \;\; y_i = -1} \quad \textbf{(II.45)}$$

Ainsi, au lieu de chercher le vecteur de poids w^0 qui minimise le carré de la norme (w, w) comme dans le cas des données séparables, on cherche à minimiser la quantité suivante :

$$\boldsymbol{\phi(w, \xi) = \frac{1}{2} w.w + C \sum_{i=1}^{m} \xi_i} \quad \textbf{(II.46)}$$

Où $\xi = (\xi_i, \ldots, \xi_m)$ et C est une constante choisie par l'utilisateur.

La solution λ_i^0 s'obtient de la même manière, par la maximisation du Lagrangien dual.

$$\boldsymbol{max_{\lambda} L(w, b, \lambda)}$$

Sous des contraintes, un peu diffiérentes du cas d'une classification de données linéairement séparables :

$$\sum_{i=1}^{m} \boldsymbol{\lambda_i^0 y_i = 0} \quad \boldsymbol{0 \leq \lambda_i \leq C} \quad \forall_i \in \{1, \ldots, m\}$$

Ainsi, nous obtenons l'équation de l'hyperplan optimal au sens des nouvelles contraintes tenant compte des ξ_i. Les vecteurs support sont toujours les vecteurs exemples les plus proches de cet hyperplan. [17]

3.3.3. Type de noyau

C'est le type de la fonction noyau à utiliser, peut être défini à : Linéaire, polynôme, RBF ou Sigmoïde.

Linéaire : $K(x_i, x_j) = x_i^T x_j$.

polynôme: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$.

fonction de base radiale (RBF): $K(x_i, x_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2\right), \gamma > 0$.

Sigmoïde : $K(x_i, x_j) = tanh(\gamma x_i^T x_j + r)^d$.

Tels que : x_i^T représente la transposé du vecteur contenant les valeurs des attributs de l'ensemble d'apprentissage, et x_j le vecteur des labels (étiquettes). Le $\gamma, r \; et \; d$ sont des paramètres (rentrés pas l'utilisateur, sont les paramètres du noyau). [18 19]

3.3.4. SVMs binaires

Le cas le plus simple est celui où les données d'entrainement viennent uniquement de deux classes différentes (+1 ou -1), on parle alors de classification binaire. L'idée des SVMs est de rechercher un hyperplan (droite dans le cas de deux dimensions) qui sépare le mieux ces deux classes (Figure II.11).

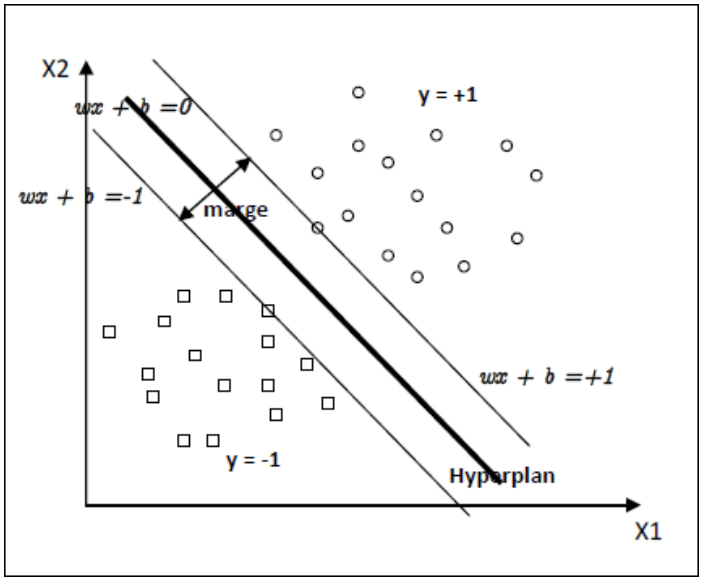


Figure II.11. *SVM binaire*

Si un tel hyperplan existe, c'est-à-dire si les données sont linéairement séparables, on parle d'une machine à vecteur support à marge dure (Hard margin).

3.3.5. SVMs multiclasse

Les machines à vecteur support sont dans leur origine binaire. Cependant, les problèmes du monde réel sont dans la plupart des cas multiclasse, l'exemple le plus simple en est la reconnaissance des caractères optiques (OCR). Dans de tels cas, on ne cherche pas à affecter un nouvel exemple à l'une de deux classes mais à l'une parmi plusieurs, c.-à-d. que la décision n'est plus binaire et un seul hyperplan ne suffit plus.

Les méthodes des machines à vecteur support multiclasse, réduisent le problème multiclasse à une composition de plusieurs hyperplans biclasses permettant de tracer les frontières de décision entre les différentes classes [20]. Ces méthodes décomposent l'ensemble d'exemples en plusieurs sous-ensembles représentant chacun un problème de classification binaire. Pour chaque problème un hyperplan de séparation est déterminé par la méthode SVM binaire. On construit lors de la classification une hiérarchie des hyperplans binaires qui est parcourue de la racine jusqu'à une feuille pour décider de la classe d'un nouvel exemple. On trouve dans la littérature plusieurs méthodes de décomposition :

3.3.5.1. Une-contre-reste (1vsR)

C'est la méthode la plus simple et la plus ancienne. Selon la formulation de Vapnik [20], elle consiste à déterminer pour chaque classe k un hyperplan $H_k(w_k, b_k)$ la séparant de toutes

les autres classes. Cette classe K est considérée comme étant la classe positive (+1) et les autres classes comme étant la classe négative (-1), ce qui résulte, pour un problème à K classes, en K SVM binaires. Un hyperplan H_k est défini pour chaque classe k par la fonction de décision suivante :

$$\mathbf{H_k(x)} = \mathbf{signe}(\langle \mathbf{w_k}, \mathbf{x} \rangle + \boldsymbol{b_k}) = \begin{cases} \mathbf{+1 \ si \ f_k(x) > 0;} \\ \mathbf{0 \quad sinon} \end{cases} \quad \textbf{(II.47)}$$

La valeur retournée de l'hyperplan permet de savoir si x appartient à la classe K ou non. Dans le cas où il n'appartient pas à K ($H_k(x)=0$), nous n'avons aucune information sur l'appartenance de x aux autres classes. Pour le savoir, on présente x à tous les hyperplans, ce qui donne la fonction de décision de l'équation (II.48) suivante :

$$\mathbf{k^* = \underbrace{Arg}_{(1 \le k \le K)} \ Max(H_k(x))} \quad \textbf{(II.48)}$$

Si une seule valeur $H_k(x)$ est égale à 1 et toutes les autres sont égales à 0, on conclut que x appartient à la classe K. Le problème est que l'équation (II.48) peut être vérifiée pour plus d'une classe, ce qui produit des régions d'ambiguïté, et l'exemple x est dit non classifiable. La Figure II.12 représente un cas de séparation de 3 classes.

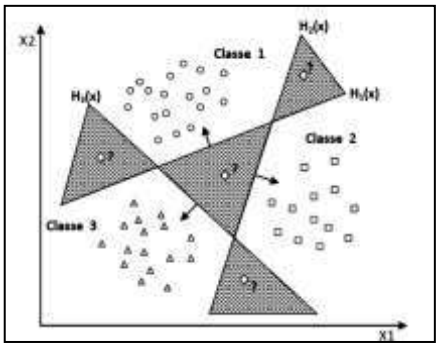


Figure II.12. *Approche une-contre-reste avec des zones d'indécision*

Pour surmonter cette situation, la méthode 1vsR utilise le principe de "le gagnant prend tout" ("winner-takes-all") : la classe K retenue est celle qui maximise de l'équation $f_k(x) = \langle w_k, x \rangle + b_k$ (II.49).

$$\mathbf{k^* = \underbrace{Arg}_{(1 \le k \le K)} \ Max \ (\langle wk, x \rangle + bk)} \quad \textbf{(II.49)}$$

Géométriquement interprétée, tout nouvel exemple x est affecté à la classe dont l'hyperplan est le plus loin de x, parmi les classes ayant $H(x) = 1$ (Figure II.13).

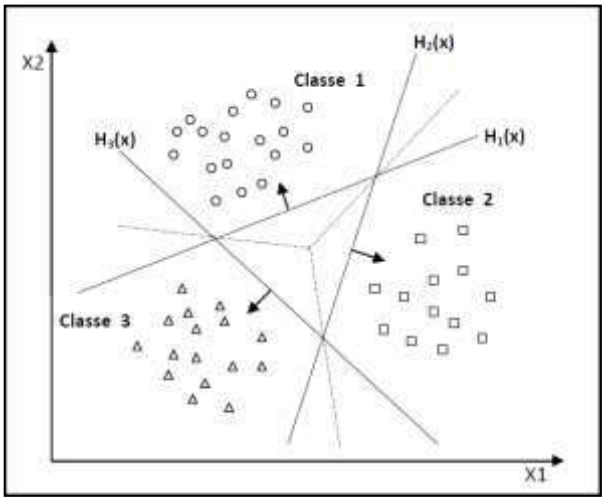


Figure II.13. *Résolution des cas d'indécision dans la méthode 1vsR*

La méthode 1vsR peut être utilisée pour découvrir même les cas de rejet où un exemple n'appartient à aucune des K classes. Pour cela, on prend les deux fonctions de décision les plus élevées, puis on calcule leur différence, si elle est au-dessous d'un certain seuil, l'exemple est rejeté.

Souvent, la méthode 1vsR est critiquée à cause de son asymétrie [20], puisque chaque hyperplan est entrainé sur un nombre d'exemples négatifs beaucoup plus important que le nombre d'exemples positifs. Par exemple dans le cas de l'OCR, le classificateur du caractère 'A' est entrainé sur des exemples positifs représentant 'A' et des exemples négatifs représentant tous les autres caractères. La méthode une contre une suivante est une méthode symétrique qui corrige ce problème.

3.3.5.2. Une-contre-une (1vs1)

Cette méthode, appelée aussi "pairwise", revient à Kner et ses co-auteurs [20] qui l'ont proposée pour les réseaux de neurones. Elle consiste à utiliser un classificateur pour chaque paire de classes. Au lieu d'apprendre K fonctions de décisions, la méthode 1vs1 discrimine chaque classe de chaque autre classe, ainsi $\mathrm{K}\,(\mathrm{K}-1)/_2$ fonctions de décisions sont apprises.

Pour chaque paire de classes (K, s), la méthode 1vs1 définit une fonction de décision binaire $h_{ks}: \Re \rightarrow \{-1, +1\}$. L'affectation d'un nouvel exemple se fait par liste de vote. On teste un exemple par le calcul de sa fonction de décision pour chaque hyperplan. Pour chaque test, on vote pour la classe à laquelle appartient l'exemple (classe gagnante). On définit pour le faire la fonction de décision binaire $H_{ks}(x)$ de l'équation II.50.

$$\mathbf{H_{ks}(x)} = \mathbf{signe(f_{ks}(x))} = \begin{cases} \mathbf{+1\ si\ f_{ks}(x) > 0;} \\ \mathbf{0\quad sinon} \end{cases} \qquad \textbf{(II.50)}$$

Sur la base des $K(K-1)/2$ fonctions de décision binaires, on définit K autres fonctions de décision (équation II.51) :

$$\mathbf{H_k(x)} = \sum_{\mathbf{s=1}}^{\mathbf{m}} \mathbf{H_{ks}(x)} \qquad \textbf{(II.51)}$$

Un nouvel exemple est affecté à la classe la plus votée. La règle de classification d'un nouvel exemple x est donnée par l'équation II.52:

$$\mathbf{k^*} = \underbrace{\mathbf{Arg}}_{\mathbf{(1 \le k \le K)}} \ \mathbf{Max(H_k(x))} \qquad \textbf{(II.52)}$$

Malheureusement, la fonction II.52 peut être vérifiée pour plusieurs classes, ce qui produit des zones d'indécisions. La méthode de vote affecte dans ce cas, un exemple aléatoirement à l'une des classes les plus votées.

La Figure II.14 représente un exemple de classification de trois classes avec la zone d'indécision.

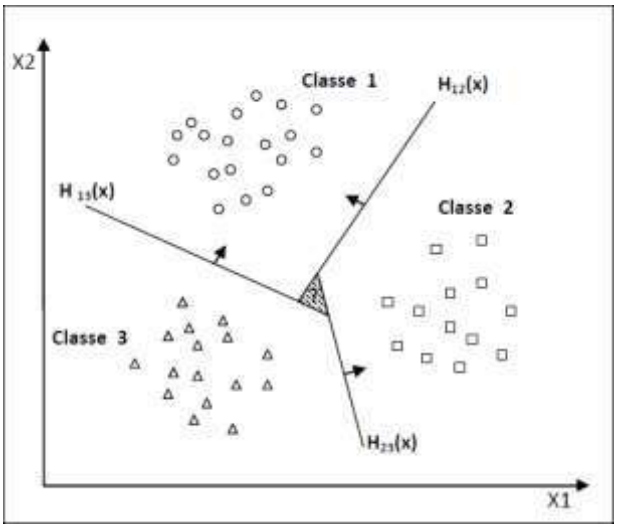


Figure II.14. *Approche une-contre-une*

Bien que La méthode 1vs1 utilise, pour l'entrainement, un nombre plus important d'hyperplans que la méthode 1vsR, elle est souvent plus rapide. Cela est dû, d'une part, au

nombre limité d'exemples utilisés pour entrainer chaque hyperplan, et d'autre part, à la simplicité des problèmes à résoudre. En effet, chaque deux classes prises à part sont moins chevauchées que toutes les classes.

4. Décision

Ce module fait appel à un jeu de données dit « d'apprentissage » dont les instances contiennent une valeur cible (target) également appelée « étiquette de classe » ainsi que plusieurs attributs (attributes) représentant l'ensemble des variables observées, et ce pour produire un modèle permettant de prédire les valeurs cibles d'un autre ensemble dit «de test», en ne fournissant à ce modèle-là que les attributs des données de test. En d'autres termes, si on considère un jeu de données divisé en deux groupes : un groupe pour les exemples connus et un autre pour les exemples non connus, le but des SVM est d'apprendre une fonction qui traduit le comportement des exemples connus pour prédire les cibles des exemples inconnus, ça veut dire, le SVM va faire une comparaison entre les modèles (les exemples connus) et les tests (les exemples non connus) tel que le modèle prédit le test

5. Conclusion

Un système de reconnaissance vocale typique se compose d'un prétraitement et post traitement. Le prétraitement est essentiellement un module d'extraction de caractéristiques, plusieurs méthodes sont utilisées dans ce domaine pour extraire les paramètres pertinents utiles à la reconnaissance. Nous nous avons cites les plus communément utilisées comme la technique MFCC, PLP...etc. Le post traitement est essentiellement un classificateur. Un classificateur peut être un modèle de correspondance (Dynamic Time Warp, DTW) ou un classificateur statistique (Gaussian Mixture Model, GMM) ou des modèles de machines à vecteurs de support (Support Vector Machine, SVM). La fonction du post traitement est de prendre des décisions de reconnaissance précise basée sur les caractéristiques de parole extraite par le prétraitement. Par conséquent, laqualité d'un algorithme frontal joue un rôle important dans les systèmes de reconnaissance de la parole.

CHAPITRE III : Implémentation du système de reconnaissance

1. Introduction

Ce chapitre met l'accent sur la conception d'une application ayant pour objectif de réaliser un système de reconnaissance des marques de téléphones portables à partir des signaux de la parole. Il couvrira tous les aspects de la partie pratique de notre travail de recherche en englobant les divers schémas. Ces derniers, à travers des algorithmes et des méthodes appliqués, illustrent dans la mesure du possible tout ce qui a été présenté dans la partie théorique.

2. Base de données et Protocole d'évaluation

Notre système est sous la forme d'une application de reconnaissance des marques de téléphones portables à partir des signaux de la parole, Ce système a la capacité de juger si l'opération d'identification des téléphones mobiles est juste ou fausse. C'est la raison pour laquelle il était indispensable de préparer une collection des sons pour construire une base de données.

2.1. Construction de la base de données

Dans les expériences, nous avons utilisé une base de données de téléphones portables qui se compose de 15 dispositifs. La collection de marques incluent Nokia, Samsung, Sony Ericsson, LG, Iphone, BlackBerry et Htc. Quatre de Nokia, deux de Samsung, deux de LG, deux de Sony Ericsson, trois de iPhone, un BlackBerry et un Htc. Ceux-ci ont été utilisés dans les expériences. La marque et le modèle des téléphones cellulaires sont résumés dans le Tableau III.1. Nous avons utilisé les paires de téléphones cellulaires avec exactement la même marque et le modèle pour explorer la capacité de discrimination des fonctions et le système de reconnaissance sur les mêmes unités de dispositifs. Deux Nokia C9 et deux iPhone 3GS sont les paires de téléphone cellulaire avec la même marque et le modèle.

Tableau III.1. *La marque et le modèle des téléphones cellulaires*

Marque	Modèle	Nombre
Black Berry	8320	1
Htc	Wildfire S	1
Iphone	3 GS	2
	5	1
LG	405	1
	P 350	1
Nokia	C 5	1
	C 9	2
	N 97	1
Samsung	E 250	1
	MINI 2	1
Sony Ericsson	W 100 I	1
	Experia	1

Pour les sons enregistrés, on a utilisé la base de données TIMIT. La base de données TIMIT se compose de différents d'intervenants dialectes de l'anglais américain. Dans notre cas on a pris 10 personnes où chaque locuteur lit une phrase d'environ 30 secondes. Nous avons choisi une fréquence d'échantillonnage de 8000 Hz et un nombre de codage de 16 bits.

Pour construire le modèle de chaque téléphone portable, nous avons considéré les 10 premières secondes des 30 secondes de chaque personne, on aura 100 secondes. Apres avoir extrait les paramètres ceux-ci seront utilisés pour l'extraction du modèle de téléphone. le reste des enregistrement c'est à dire les 20 secondes qui restent de ces 30 secondes de chaque personnes seront divisés en 5 parties égales pour être utilisés dans la partie test.

Tous ce travail sera appliqué à chaque appareil téléphonique, ce qui veut dire que chaque appareil téléphonique a un signal de parole de formation de 100 seconde et 5 signaux de parole pour les essais (tests), chacun environ 40 secondes.

Comme nous avons cités dans le premier paragraphe que nous avons 10 personnes dont chacun entre eux lis une phrase environs 30 secondes, on enlève les 10 premières secondes des 30 secondes de chaque personnes et puis on additionne tous les 10 secondes enlevés, on aura 100 secondes qui seront utilisés pour l'extraction du model de téléphone, ça veut dire 20 secondes qui restent de ces 30 secondes de chaque personnes seront partageais en 5 parties égaux, à la fin on additionne les 1er parties de chaque personnes ensembles, les 2eme parties ensembles, jusqu'à 5eme parties pour avoir à la fin 5 parties, chaque partie environs 40 secondes et cela pour les utilisés dans le test.

2.2. Protocol d'évaluation

Pour évaluer notre système, on s'est basé sur le calcul du taux de reconnaissance correcte. Ce taux se calcul par le simple rapport entre le nombre de fois où la reconnaissance est correcte sur le nombre total de tests dans une seule expérience.

2.2.1. Collection des échantillons et extraction des paramètres acoustique

Pour toutes ces expériences, nous nous sommes basés sur le logiciel MATLAB pour collecter les données de chaque signal puis lire ce son à partir du disque dur. Nous avons choisi une fréquence d'échantillonnage fe de 8000Hz et un nombre de bits par échantillon de 16 bits.

L'analyse consiste à calculer les paramètres acoustiques du signal d'apprentissage et du signal de test. Nous avons choisi deux types de paramètres pour notre système, qui sont, les coefficients MFCC et PLP.

En plus de ces paramètres, on a considéré aussi les caractéristiques dynamiques, a travers les dérivées première et deuxième.

L'énergie du signal a été aussi considérée dans notre travail, car nous estimant que cette information permet d'améliorer les résultats de reconnaissance.

Tous ces paramètres vont être transmis vers un autre traitement qui est soit, la modélisation par SVM (machine à vecteurs de support), soit le test. Notons que nous avons 15 appareils téléphoniques, chaque appareil a un signal d'apprentissage qui sera incorporé dans le système SVM pour créer un modèle, et les 5 signaux restants vont servir de tests, et cela pour comparer chaque model d'un appareil téléphonique avec tous les tests des téléphones portables, ça veut dire comparer chaque model avec 75 tests.

2.2.2. Reconnaissance des marques de téléphones portables par la méthode (SVM)

Cette technique fait appel à un jeu de données dit « d'apprentissage » dont les instances contiennent une valeur cible (Target) également appelée « étiquette de classe » ainsi que plusieurs attributs (attributs) représentant l'ensemble des variables observées, et ce pour produire un modèle permettant de prédire les valeurs cibles d'un autre ensemble dit «de test», en ne fournissant à ce modèle-là que les attributs des données de test. En d'autres termes, si on considère un jeu de données divisé en deux groupes : un groupe pour les exemples connus

et un autre pour les exemples non connus, le but des SVM est d'apprendre une fonction qui traduit le comportement des exemples connus pour prédire les cibles des exemples inconnus.

Nous allons utiliser la machine à vecteurs de support (SVM) classificateur dans nos expériences de reconnaissance. La SVM est un classificateur puissant qui a été utilisé dans de nombreuses applications de reconnaissance des formes. La zone la plus difficile de l'application de SVM classificateur est les applications vocales. Leur avantage principal est leur performance remarquable en généralisation avec des données de grandes dimensions. On leur reproche cependant d'être lents en convergence et gourmands en mémoire dès qu'il s'agit de traiter des jeux de données volumineux. Etant donné que le signal de parole est générée par la division du signal en trames courtes se chevauchant, à partir de caractéristiques extraites d'un signal de parole est un ensemble de vecteurs, plutôt que d'un seul vecteur. Ainsi, l'approche du noyau de séquence qui associe l'ensemble des vecteurs de caractéristiques en un seul vecteur caractéristique dimensionnel est utilisée dans les applications vocales de SVM. Dans notre travail nous avons l'outil open source LIBSVM très utilisée en reconnaissance [21].

2.2.3. Fonctionnalités de la LIBSVM

- ✓ Normalisation des données : svm-scale

La normalisation des données est préconisée dans la plupart des traitements d'analyse, la LIBSVM appuie cette pratique et propose une fonctionnalité pour la normalisation : svm-scale, elle recommande une normalisation linéaire utilisant les intervalles [-1; +1] ou [0; 1], en utilisant bien sûr le même intervalle pour le jeu de données d'apprentissage et de test. Pour faire appel à cette commande il faut taper en ligne de commande :

svm-scale -y lower upper

Où **lower** et **upper** constituent les bornes de l'intervalle à utiliser pour la normalisation, généralement [0,1] ou [-1,1].

- ✓ Apprentissage : svm-train

La commande svm-train permet de lancer le processus d'apprentissage sur les données, elle prend en entrée un format de données spécifique, décrit plus loin, et génère un fichier de modèle. La commande pour utiliser cette fonctionnalité est :

svmtrain [options] training_set_file [model_file]

Où **training_set_file** constitue le fichier d'apprentissage et **model_file** le nom du modèle à générer.

- ✓ Prédiction : svm-predict

Génère les classes prédites ou les valeurs prédites (en cas de régression), en se basant sur un modèle entrainé, la commande à utiliser est la suivante :

svmpredict test_file model_file output_file

Où **test_file** est fichier de test, ou les données que l'on souhaite prédire, à noter qu'il doit également être sous le format LIBSVM et **model_file** le nom du modèle à générer.

Après prédiction, la fonction svm-predict compare les résultats de la prédiction avec les labels du fichier de test, pour en déduire la justesse de la prédiction et par la même occasion le taux d'erreur commis.

3. Expériences et résultats

Le système de reconnaissance est basé sur deux types de paramètres à savoir les coefficients MFCC et PLP. Le Tableau III.2 montre les taux de reconnaissance pour différents cas.

Tableau III.2. *Effet de l'énergie sur le taux d'identification (p=12).*

caractéristique	Taux d'identification (%)
MFCC	60
MFCC + E	93.33
PLP	26.67
PLP + E	45.33

Les résultats montrent la supériorité des paramètres MFCC sur les paramètres PLP. Elles montrent aussi l'effet positif apporté par l'ajout de l'énergie, qui joue un rôle important dans l'opération d'amélioration le taux de reconnaissance que ce soit pour MFCC ou PLP.

Tableau III.3. *Effet des caractéristiques dynamiques et l'augmentation du nombre de paramètres sur le taux d'identification.*

caractéristique	Taux d'identification (%) p=12	Taux d'identification (%) p=14	Taux d'identification (%) p=16	Taux d'identification (%) p=22
MFCC+E	93.33	93.33	93.33	98.67
MFCC+E+Δ	93.33	93.33	94.67	100
MFCC+E+Δ+Δ^2	93.33	94.67	94.67	100
PLP+E	45.33	53.33	56	89.33
PLP+E+Δ	60	72.00	82.67	100
PLP+E+Δ+Δ^2	33.33	77.33	85.33	100

Pour la chaine de données de D=12, la technique MFCC manifeste des résultats meilleures que ceux du type PLP, dont ce dernier a besoin de plus de données (Δ) pour donner un résultat de 60%. Pour la technique MFCC même avec l'ajout d'énergie et de la dériver le résultat reste à 93.33%.

Pour les chaines de paramètre de D=14 et16, les résultats de la technique MFCC restent les mêmes, dont l'ajout d'énergie et de la dérive cause une légère amélioration de 1.34%. Par contre l'augmentation de la chaine des paramètres fait un impact considérable sur la fonction PLP, comme montré dans Tableau III.3 des résultats ou il y a une augmentation avec l'ajout de l'énergie et Δ, et aussi une augmentation avec l'ajout de $(\Delta + \Delta^2)$ entre D=12 et 14. D'autre part, l'ajout d'énergie et de la dériver pour D= 14 et 16 avec PLP, améliore les résultats d'une façon brusque dont le taux d'amélioration après l'ajout de ces paramètres.

Pour la chaine de D=22, la fonction MFCC donne un résultat de 98.67% avec l'ajout de l'énergie, par contre le résultat est de 100% avec la présence de la dérivée Δ. Pour la technique PLP, le résultat est parfaitement amélioré (89.33%), ces résultats deviennent de 100% après l'ajout de la dérivée Δ, donc une amélioration d'environ 11%.

D'après les tableaux III.3, les systèmes basés sur la technique MFCC ont atteint une performance supérieure à 90%.

Apres l'analyse des résultats, Nous sommes arrivés aux points suivants :

- La fonction MFCC donnede meilleure performance par rapport à la fonction PLP.

- L'augmentation de paramètre D permet aux fonctions MFCC et PLP de donner des bons résultats et de plus de précision.
- L'introduction des dérivés première et seconde (Δ et Δ^2) aide à améliorer les résultats surtout pour la fonction PLP dont la différence après l'ajout des deux dernier est très considérable.

4. Conclusion

Dans ce chapitre nous avons décrit notre système en expliquant comment on a créé notre base de données, ensuite nous avons cité l'ensemble des étapes appliquées pour la conception de notre application.

Nous avons comparé les différentes caractéristiques acoustiques de reconnaissance de téléphone cellulaire source à l'aide des signaux de parole enregistrées. Il a été brièvement montré que le signal de parole transmet des informations sur le périphérique source dans des formes convolutifs, multiplicative et additifs dans le temps, la fréquence et les domaines Cepstraux, respectivement. Ainsi, tout type de caractéristiques extraites à partir du signal de parole contient des informations sur la source dans une certaine forme. Les paramètres MFCC et PLP ont été comparés. Dans les comparaisons, on a constaté que les paramètres MFCC basé surpassent l'autre type de caractéristiques acoustiques. C'est ainsi que les MFCCs donnent de meilleures performances. Enfin, l'ajout de la 1ère dérivées d'ordre des vecteurs de caractéristiques, fonctions Δ, améliore la précision d'identification alors que cette observation ne tient pas pour les fonctions Δ^2.

Conclusion générale

Le travail que nous avons effectué consiste à concevoir un système de reconnaissance de téléphone cellulaire à l'aide des signaux de parole enregistrées. Pour réaliser notre système, nous avons commencé par donner les concepts généraux sur la reconnaissance basée sur la parole, ceci nous a permet de citer les méthodes les plus utilisées dans l'analyse de la parole et dans le domaine de reconnaissance basée sur la parole. nous avons ensuite donné une étude plus détaillée sur la représentation acoustique du signal de parole ainsi que la reconnaissance. C'est ainsi qu'on a définit les coefficients MFCC (Mel Frequency Cepstrum Coefficients) et PLP (Perceptual Liner Prédiction) utilisés dans la majorité des systèmes de reconnaissance actuelle. De plus, nous avons décrit l'approche statistique pour la reconnaissance fondée sur l'utilisation des modèles de SVM (support vector machine).

Nous avons consacré le dernier chapitre à la conception et la mise en œuvre de notre projet. Nous avons vus les différents mécanismes et outils utilisés pour la réalisation de notre système. Des expériences ont été menées pour comparer les caractéristiques des coefficients MFCC et PLP. Nous avons aussi fait des comparaisons de différentes caractéristiques dans un même type de paramètres. Aussi, nous avons examiné l'effet de l'énergie, des caractéristiques dynamiques (1^{ers} et $2^{èmes}$ ordre dérivatives des paramètres, Δ et Δ^2) et le nombre de paramètres utilisés.

Les différentes expériences réalisées nous ont permet de conclure que l'utilisation des paramètres MFCC ou PLP dans le système de reconnaissance basé sur la modélisation SVM permettent d'identifier avec efficacité les téléphones cellulaires en utilisant la bonne configuration. Ainsi, l'ajout de l'énergie, et l'augmentation du nombre de paramètre a un rôle important dans l'amélioration de taux de reconnaissance.

Notons aussi que les paramètres MFCC présentent de meilleurs résultats par rapport au paramètres PLP.

Références

[1] Z. HAMAIZIA, M. BEDDA, « *ANALYSE ET SYNTHESE DE LA PAROLE ARABE*», Courrier du Savoir – N°12, Octobre 2011, pp.71-74. Université Mohamed Khider – Biskra, Algérie, 2011.

[2] OULD RAMOUL Amina, YAHIA HELALI Nouha, «*CONTRIBUTION A LA CONCEPTION D'UN CORRECTEUR AUTOMATIQUE DES REGLES DE RECITATION DU CORAN « TAJWID » (APPLICATION A L'IDGHAM) »*, Mémoire fin d'études 2013-2014.

[3] http://fr.wikipedia.org/wiki/Reconnaissance_automatique_de_la_parole. Consulté le 17 juillet 2014

[4] Mr. Abdelghani HARRAG, « *Extraction des donnees d'une base: Application à l'extraction des traits du locuteur »*, Ministère de l'Enseignement Supérieur et de la Recherche Scientifique. Université Ferhat ABBAS – Sétif. UFAS – Algérie, THESE présentée à la faculté de technologie, 26/06/2011.

[5] Cemal Hanilçi, Figen Ertas, *« Optimizing Acoustic Features for Source Cell-Phone Recognition Using Speech Signals »*, Department of Electronic Engineering, Uludag University, 16059, Bursa, Turkey, 17 June 2013.

[6] G Fant, *« Acoustic theory of speech production »*, s-Gavenhage, Mouton, 1960.

[7] René Boite, Hervé Bourlard, Thierry Doutoit, Joël Hancq, et Henri Leich, *«Traitement de la parole »*, Nouvelle édition, entièrement revue et augmentée, ISNB 2-880746-388-5, 2000. Presses polytechniques et universitaires romandes, CH _ 1015 Lausanne. [Livre électronique]. Disponible.

[8] P. Pravinkumar and B, M Wasfy *«Speaker Verification/Recognition and the Importance of Selective Feature Extraction: Review»*, *Proc. IEEE*, 2001.

[9] Tounsi bilal, *«Inférence d'identité dans le domaine forensique en utilisant un système de reconnaissance automatique du locuteur adapté au dialecte Algérien»*, Institut National de Formation en Informatique (I.N.I) Oued-Smar Alger, 2007-2008.

[10] http://fr.wikipedia.org/wiki/Cepstre#MFCC Consulté le 17 juillet 2014.

[11] Loïc BARRAULT, *« Diagnostic pour la combinaison de systèmes de reconnaissance automatique de la parole »*, THÈSE présentée pour obtenir le grade de Docteur en Sciences, SPÉCIALITÉ Informatique, l'Université d'Avignon et des Pays de Vaucluse, École Doctorale 166 I2S «Mathématiques et Informatique» Laboratoire d'Informatique (EA 931), 16 oct. 2009.

[12] Octavian Cheng, Waleed Abdulla, Zoran Salcic, *« Performance Evaluation of Front-end Processing for Speech Recognition Systems»*, School of Engineering Report No. 621, Electrical and Computer Engineering Department, School of Engineering, The University of Auckland, 2005.

[13] Haton 1991 J-P. Haton, J-M. Pierrel, G. Perennou, J.Calen, J-L.Gauvain, *«Reconnaissance automatique de la parole»*, ISBN 2-04-018827-9, Bordas, Paris, 1991.

[14] LÊ Viet Bac, *«Reconnaissance automatique de la parole pour des langues peu dotées»*, Thèse préparée au sein du laboratoire de Communication Langagière et Interaction Personne-Système, Fédération IMAG – Université Joseph Fourier – Grenoble I, le 1er juin 2006.

[15] Philipp Zimmermann, *«Analyse de l'Influence des Conditions d'Enregistrement dans la Reconnaissance Automatique de Locuteurs en Sciences Forensiques »*, Séminaire de Sciences Forensiques, Ecole des Sciences Criminelles – Institut de Police Scientifique – BCH – UNIL – 1015 Dorigny/Lausanne, Le 10 juin 2005.

[16] http://fr.wikipedia.org/wiki/Mod%C3%A8le_de_m%C3%A9langes_gaussiens Consulté le le 9 aout 2014.

[17] Yassine BEN AYED, *«Détection de mots clés dans un flux de parole»*, THESE l'obtention du Doctorat de l'Ecole Nationale Supérieure des Télécommunications (Spécialité : Signal et Images), présentée et soutenue publiquement le 23 décembre 2003.

[18] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, *« A Practical Guide to Support Vector Classi_cation »*, Department of Computer Science National Taiwan University, Taipei 106, Taiwan. Initial version: 2003 Last updated: April 15, 2010.

[19] Melle Imane EL HASSANI, *« SVR avec Boosting pour la prévision à long terme»*, Projet de Fin d'études, École Polytechnique de l'Université de Tours - Département Informatique, Année scolaire 2011 -2012.

[20] Abdelhamid DJEFFAL, *«Utilisation des méthodes Support Vector Machine (SVM) dans l'analyse des bases de données»*, Thèse présentée pour l'obtention du diplôme de Docteur en science Spécialité Informatique, Université Mohamed Khider – Biskra, Année Universitaire 2011/2012.

[21] Chang, Chih-Chung; Lin, Chih-Jen (2011). "LIBSVM: A library for support vector machines". *ACM Transactions on Intelligent Systems and Technology* **2** (3)