N° d'ordre :
N° de série :

République Algérienne Démocratique etPopulaire
Ministère de l'Enseignement Superieur et de la Recherche Scientifique

UNIVERSITÉ D'EL-OUED

FACULTÉ DESSCIENCES ET DE TECHNOLOGIE

Mémoire de fin d'étude

LICENCEACADEMIQUE

Domaine: Mathématiques et Informatique
Filière: Informatique
Spécialité: Informatique fondamental

Présenté par:
BERCHAOUA Salah Eddine

Thème

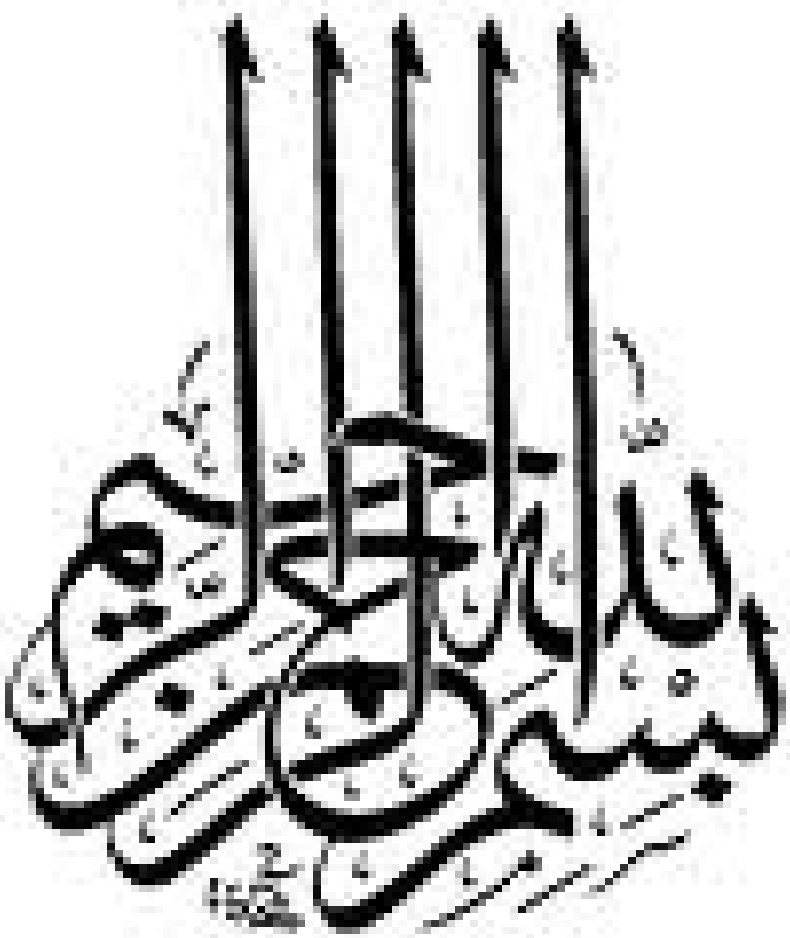
Reconnaissance de la parole arabe par les supports vecteurs machines (SVM)

Soutenu le …juin 2013

Devant le jury composé de:

Mr. X	MC(B) Univ. El Oued	Président
Mr. Y	A (B) Univ. El Oued	examinateur
Mr. ZAIZ Faouzi	MA(B) Univ. El Oued	Rapporteur

Année universitaire 2012 – 2013



Remerciements

A Dieu, le tout puissant, nous rendons grâce pour nous avoir donné santé, patience, volonté et surtout raison.

En premier lieu, nous tiendrons à remercier notre encadreur Mr ZAIZ Faouzi A qui nous a aidé et conseillé durant cette année.

Nos remerciements vont également à tous les Enseignants de la faculté d'université d'EL-Oued Institute des Sciences ET Technologies département d'informatique pour avoir accepté de juger ce travail.

Enfin, nous remercions tous ceux que nous accompagnés ou que nous ait croisés durant ces années, tous ceux qui nous en soutenu, encouragé et donné l'envie de mener à terme ce travail.

BERCHAOUA Salah Eddine

Dédicace

Je dédie ce modeste travail :

A ma mère pour sa tendresse et mon père pour sa patience et encouragement

A mes très chers frères, Fars, Mohamed, Eze Eddine

Et tous les membres de familles BERCHAOUA, ATIA, ZAIZ

Et à tous ceux que j'aime,

Et toutes mes amies.

BERCHAOUA Salah Eddine

Sommaire

Chapitre 1: Reconnaissance de la parole

Chapitre 2: Support Vector Machines

Chapitre 3: Conception et implémentation du système

Chapitre 4: Réalisation et test de résultats

Liste des figures

Chapitre 1: Reconnaissance de la parole

Chapitre 2: Support Vector Machine

Chapitre 3: Conception et implémentation du système

Chapitre 4:Réalisation et test de résultats

Liste des tableaux

Chapitre 1: Reconnaissance de la parole

Chapitre 3: Conception et implémentation du système

Chapitre 4:Réalisation et test de résultats

Résumé

Si l'homme a la faculté de comprendre un message vocal provenant d'un locuteur quelconque, dans des environnements souvent perturbés, quelques soient son mode d'élocution, la syntaxe et le vocabulaire utilisés, la machine est-elle capable d'en faire autant ? Une solution peut-elle répondre en globalité à ces difficultés ? Le problème de la reconnaissance vocale est un sujet d'actualité et pour l'instant, seules les solutions partielles sont aptes à répondre aux différentes tâches que la machine doit effectuer.

Ce document est destiné à la conception et à la réalisation d'un système de saisie à l'aide des commandes vocales basée sur l'apprentissage par SVM (support vector machines) l'une des méthodes d'apprentissage inspirée de la théorie de statistique de l'apprentissage de Vladimir Vapnik. C'est une méthode de classification binaire par apprentissage supervisé qui fut introduite par Vapnik en 1995. Cette méthode se base sur la recherche d'un hyperplan séparateur entre les classes dans la phase d'apprentissage et l'utilisation d'une fonction de décision dans la phase de décision.

Une commande vocale issus d'un locuteur passe par une succession d'opérations (Acquisition, Segmentation et extraction de vecteur acoustique, Classification, exécution ou calcul et synthèse de résultat) afin qu'elle soit interprétée et exécutée. Le signal acoustique est premièrement numérisé, ensuite soumis à la méthode LPC pour extraire un vecteur caractéristique. Ce dernier est comparé ensuite par la méthode SVM à d'autres vecteurs pour trouver un ressemblant dans une base de sons. Une fois la classe trouvée la commande est décodée et exécutée.

Mots clé

RAP : Traitement de la parole, Reconnaissance de la parole.
LPC : Codage prédictif linéaire.
SVM : Support Vector Machine.

Introduction générale

Aujourd'hui, l'impact des systèmes de RAP est encore minime dans la vie courante, et la commande des ordinateurs ne s'effectue toujours pas par la voix, malgré les promesses de fabricants de logiciel ou de matériel informatique (Microsoft, Apple). L'annonce de la commercialisation du système de dictée vocale d'IBM pour les ordinateurs PC en 1994 a suscité de l'intérêt, mais aussi des réserves quant aux performances actuelles du système. Pourtant les progrès réalisés depuis 25 ans en RAP sont très importants, grâce à un grand nombre de recherches traitant du problème sous tous ses aspects. Les limitations de la capacité des systèmes de reconnaissance, imposées à l'origine par la complexité de la tâche, sont progressivement repoussées, et des systèmes efficaces pour des applications spécialisées sont maintenant disponibles et commercialisés.

Notre étude s'intègre dans le cadre du développement d'un système de dictée vocale indépendant du locuteur (logiciel de saisie des notes des étudiants par dicter initialiser les nombres de l'inscription). La modélisation acoustique par les méthodes les plus performantes de l'état de l'art reste insuffisante; cette faiblesse est un facteur limitant des systèmes de RAP. Nous cherchons à améliorer la qualité de la modélisation acoustique, en intégrant certains traitements adaptés à un processus de décodage de la parole continue.

Le chapitre 1 présente les méthodes classiques employées en reconnaissance de la parole. Les difficultés rencontrées pour la mise au point des systèmes de RAP proviennent de la variabilité du signal de parole et de la continuité du processus de production.

Parmi les méthodes développées, l'approche statistique SVM (Support Vector Machine) semble la plus efficace. En introduisant un nombre d'exemples présentant des échantillons pour chaque classe plus les étiquettes de chaque classe nous pouvons définir un hyperplan séparant chaque classe de l'autre, ce qui est présenté en chapitre 2.

Le chapitre 3 présente la conception du système où en définissant les différents modules du système leur architecture générale, ensuite nous illustrant notre implémentation ainsi que la validation de notre système abordé dans le chapitre 4.

Finalement, nous terminons notre mémoire par une conclusion et les perspectives de notre projet.

Chapitre 1

Reconnaissance de la parole

- **Introduction**
- **Généralité**
- **Traitement de la parole**
- **La reconnaissance de parole**
- **Conclusion**

Introduction

Dans ce chapitre, nous allons parler sur les principales techniques associées au prétraitement du signal de la parole, on proposant un état de l'art de la reconnaissance vocale et suivant le processus de génération de la parole jusqu'à sa reconnaissance. On y trouvera les domaines d'application, le mécanisme de production de la parole et les paramètres qui la caractérisent, les principes des techniques dominantes d'analyse du signal.

L'information portée par le signal de parole peut être analysée de bien des façons. On en distingue généralement plusieurs niveaux de description non exclusifs : acoustique, phonétique, phonologique, morphologique, syntaxique,

1. Généralités

1.1. Panorama sur la reconnaissance de la parole

1.1.1. Historique

Une évolution rapide :

Les premiers travaux qui se relient directement à la reconnaissance automatique de la parole furent ceux de **J.Dreyfus-Graf**, en Suisse puis en France.

1949 : Visualisation sur un oscilloscope du signal de parole filtré dans six bandes de fréquence différentes.

1952 : reconnaissance des 10 chiffres par un dispositif électronique câblé.

1956 : Système de distinction des voyelles pour différents locuteurs et première " machine à écrire phonétiquement ", dix syllabes par locuteur.

1960 : Utilisation des méthodes numériques.

1960 - 1970 : L'ordinateur recherche automatiquement des sons spécifiques et les met en mémoire pour référence ultérieure.

1968 : reconnaissance de mots isolés par des systèmes implantés sur gros ordinateurs (jusqu'à 500 mots)

1970 : Utilisation des niveaux syntaxiques et sémantiques.

1972 : premier appareil commercialisé de reconnaissance de mots.

1990 : premières véritables applications de dialogue oral homme-machine.

1994 : IBM lance son premier système de reconnaissance vocale sur PC.

1.1.2. Problèmes rencontrés durant ces années

1.1.2.1. Continuité

Contrairement au langage écrit, où les mots sont séparés par des "blancs" dans les textes manuscrits ou par des espaces dans les textes dactylographiés, les séparateurs, symbolisés par les silences entre les mots, sont parfois très difficiles à repérer **[4].**

1.1.2.2. Variabilité

Elle provient de la position d'un phonème par rapport aux autres (coarticulation), des locuteurs aux timbres différents : homme, femme, enfant et à leur mode d'élocution : voix chantée, criée, enrouée, sous stress,....Elle est due aussi à la qualité du moyen d'acquisition et du bruit environnemental **[4].**

1.1.2.3. Reconnaissance des informations en fonction de la tâche à accomplir

La reconnaissance vocale peut s'effectuer sur les sons eux-mêmes, sur la structure syntaxique d'une phrase (dictée), sur la signification d'une phrase (robots) ou sur l'identité du locuteur et son état émotionnel (joyeux, en colère,...) **[4]**.

1.1.2.4. Depuis 1970

Les difficultés rencontrées durant ces débuts ont amené les scientifiques à classifier, puis à déterminer des axes de recherches suivant le tableau 1.1 **[4]**.

		Type d'élocution	
		Mots isolés	**Parole continue**
Approche	**Globale**	Reconnaissance de mots (petits vocabulaires) Systèmes existants	Localisation de mots Dans les phrases
	Analytique	Reconnaissance de mots (grands vocabulaires)	Localisation de mots Reconnaissance et compréhension de phrases

Tableau 1.1 : Les problèmes de la reconnaissance de la parole.

1.1.2.5. L'approche globale

Ce domaine de recherche concerne la reconnaissance, après une phase d'apprentissage, de quelques mots isolés pour un même locuteur. Elle se concrétisa en 1972 par l'industrialisation du VIP100, puis du VNC par la société Threshold Technologie (30 mots reconnus avec un taux proche de 100%) **[4]**.
La fin de cette décennie fut marquée en France par Martine Kempf et son "Katalavox".

1.1.2.6. L'approche analytique

C'est une voie de recherche fondamentale qui concerne la reconnaissance et la compréhension de la parole continue, multi locuteur, à grand vocabulaire et langage peu contraint.

Cette méthode, basée sur l'identification d'éléments phonétiques, engendra ces années là un recours massif aux traitements du type intelligence artificielle pour pallier aux erreurs de décodage des phonèmes.

Trois systèmes issus du projet "ARPA/SUR "virent le jour aux USA.

La recherche française aussi active produisit les systèmes "Myrtille 1 et 2" au C.R.I.N., Keal au C.N.E.T. de Lannion, Esope au L.I.M.S.I. à Orsay et Arial II au C.E.R.F.I.A. de Toulouse.

On remarque déjà, à la fin de ces années, l'importance prise par la modélisation "Markovienne du langage **[4]**.

1.1.2.7. Le mécanisme de la parole

L'appareil phonatoire humain (Figure 1.1) peut être assimilé, et est même souvent modélisé comme un système composé d'une source et d'un filtre. La source est un élément qui vibre soit dans un mode harmonique, soit dans un mode "aléatoire" quand il y a une constriction au niveau des cordes vocales et donc un écoulement turbulent de l'air. Le filtre résulte du conduit vocal qui est formé d'une cavité résonante complexe **[1,4]**.

1.1.3. Les résonateurs

Les cordes vocales sont les éléments vibreurs ; et comme une anche d'un instrument de musique, elles possèdent la particularité de produire, en plus de leur fréquence fondamentale, un spectre riche en harmoniques.

Mais un élément vibreur, placés devant une cavité résonante, produira alors un son dont les fréquences seront filtrées par la bande passante du résonateur.

Les ordres de grandeur des fréquences fondamentales sont de 120Hz pour les hommes, 250Hz pour les femmes et de 450Hz pour les enfants **[1,4]**.

1.1.4. L'appareil phonatoire

Le résonateur de l'appareil phonatoire est composé de quatre cavités principales en “série (Figure.1.1): le Pharynx ou arrière gorge, les deux cavités buccales délimitées par la langue et que l'on simplifiera à une seule et l'ajutage labiale situé entre les dents et les lèvres. La cavité nasale, en “parallèle” sur l'ensemble série précédent, vient compléter ce résonateur.

La source de ce résonateur est en fait décomposable en deux émissions distinctes et d'origines différentes. Les cordes vocales, en fournissant un spectre riche en harmoniques, produisent les sons voisés. Le bruit d'écoulement de l'air en provenance des poumons, dont le spectre est similaire à un bruit blanc, crée les sons non voisés.

Les sons et donc la parole naissent de l'excitation d'un résonateur et sont formés par les ouvertures et les volumes de ce dernier qui varient très rapidement.

L'observation spectrale du conduit vocal laisse apparaître des pics de résonance, appelés formants. Les affaiblissements constatés dans le spectre, nommés anti-formants, sont introduits par les sons nasalisés **[1]**.

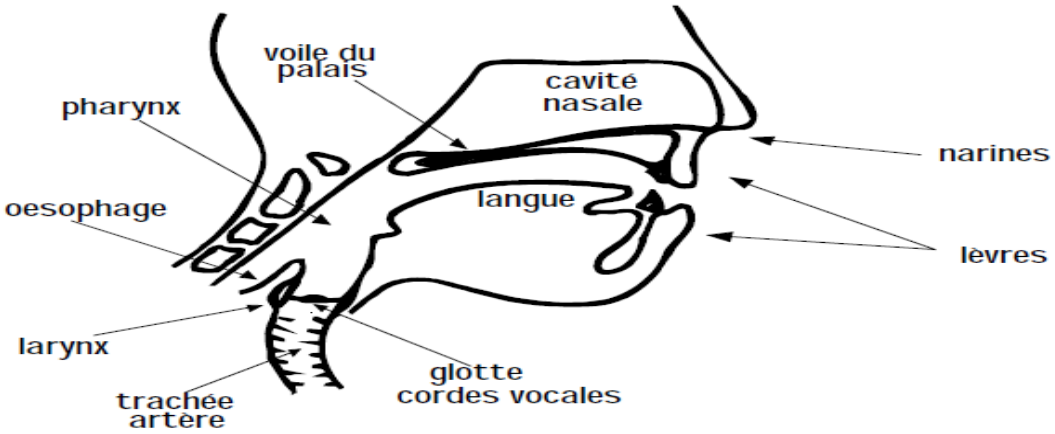


Figure 1.1 L’appareil phonatoire.

1.2. L'information vocale

Le signal de la parole véhicule plusieurs types d'informations, tels que le fondamental, la prosodie, le timbre et les phonèmes. Par conséquent, ceci impose, aux systèmes de reconnaissance vocale, de n'extraire que l'information nécessaire à son application, les phonèmes pour les machines de dictée par exemple.

La parole est surtout contenue dans les deux premiers formants, mais l'information proprement dite provient des transitions for mantiques.

En général, on considère que la plage de fréquence d'un signal de parole se situe dans la bande de 100Hz-5KHz (300Hz-3.4KHz pour la téléphonie) **[4]**.

1.3. L'appareil auditif

1.3.1. Echelles des hauteurs

1.3.1.1. L'échelle des Mels

Après 500Hz, l'oreille perçoit moins d'une octave pour un doublement de la fréquence. Des expériences psycho acoustiques ont alors permis d'établir la loi qui relie la fréquence et la hauteur perçue : l'échelle des Mels où le « Mel » est une unité représentative de la hauteur perçue d'un son (Figure 1.2) **[4]**.

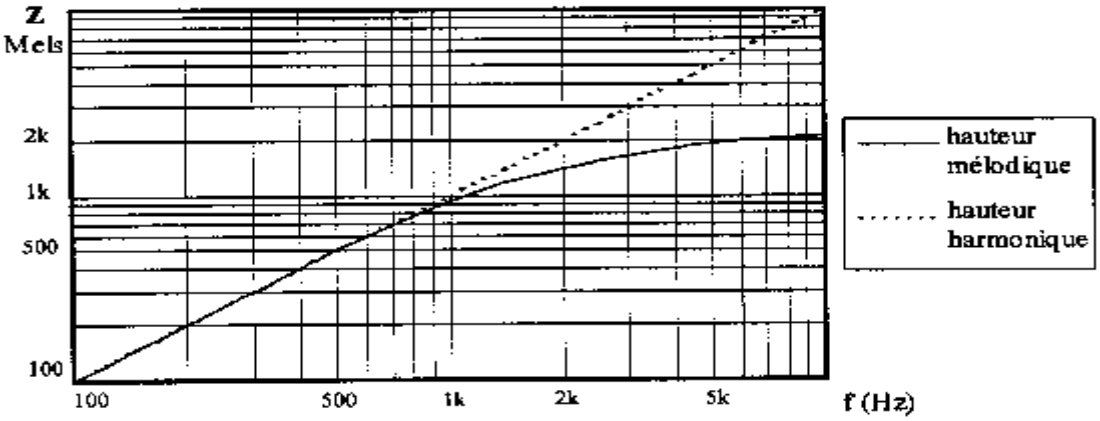


Figure 1.2 : L'octave d'un son de 2000Hz (1800 Mels) sonnera l'octave supérieure à 4600Hz (3600 Mels) au lieu de 4000Hz.

1.3.1.2. L'échelle de Bark

Le système auditif se comporte comme un banc de filtres dont les bandes, appelées “bandes critiques”, se chevauchent et dont les fréquences centrales s'échelonnent continûment. Cette bande critique correspond à l'écartement en fréquence nécessaire pour que deux harmoniques soient discriminées dans un son complexe périodique.

Remarque : Les échelles Mel ou de Bark sont approchées par un banc de 15 à 24 filtres triangulaires espacés linéairement jusqu'à 1KHz, puis espacés logarithmique ment jusqu'aux fréquences maximum **[4]**.

1.4. Les méthodes de reconnaissance vocale

1.4.1. Technologie analogique : Le spectrographe

Le spectrographe de la parole est un appareil inventé voilà plus d'un demi-siècle et commercialisé plus tard sous le nom de Sonagraphe. Historiquement, ce premier outil d'analyse pour les phonéticiens (Figure1.3) était composé d'un banc de filtres analysant les différentes fréquences successivement.

Une autre technique de cet appareil est basée sur le filtrage hétérodyne : on fait défiler le signal vocal, modulé en amplitude par une sinusoïde variable en fréquence, sous un filtre fixe. On recueille alors l'énergie pour chaque incrément de fréquence. Le signal évoluant dans le temps, on obtient alors une représentation graphique à deux dimensions (fréquence temps), nommée "sonagramme" et dont l'intensité est représentée par une échelle de gris **[4].**

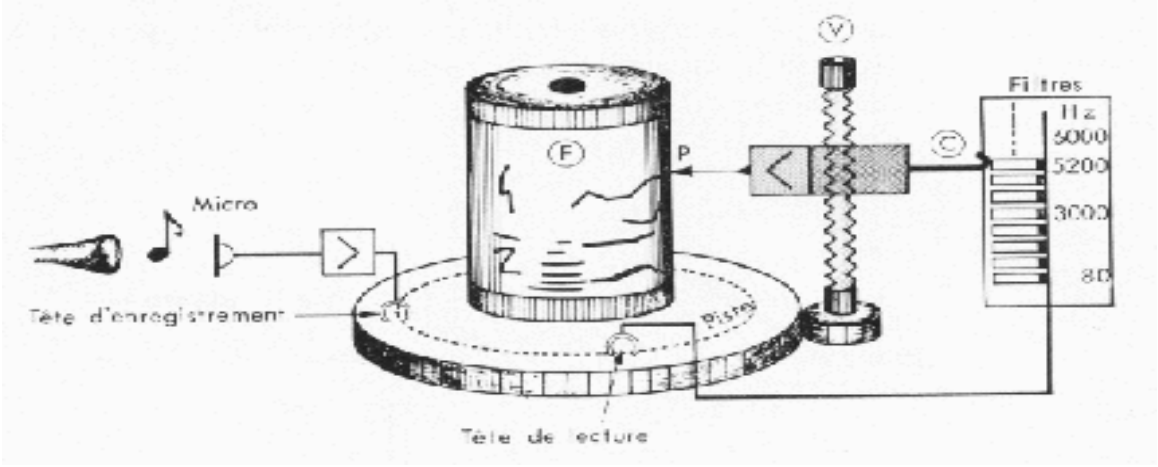


Figure 1.3 : Le spectrographe de la parole

1.4.2. Technologie numérique, introduction

Les systèmes de reconnaissance vocale numériques sont caractérisés par :

- le prétraitement qui comprend l'acquisition du signal de la parole et l'extraction des paramètres,
- l'apprentissage du vocabulaire et la comparaison aux références,
- le traitement des résultats en fonction de l'application finale.

Ces trois fonctions sont réalisées suivant deux approches : l'approche globale et l'approche analytique **[4]**.

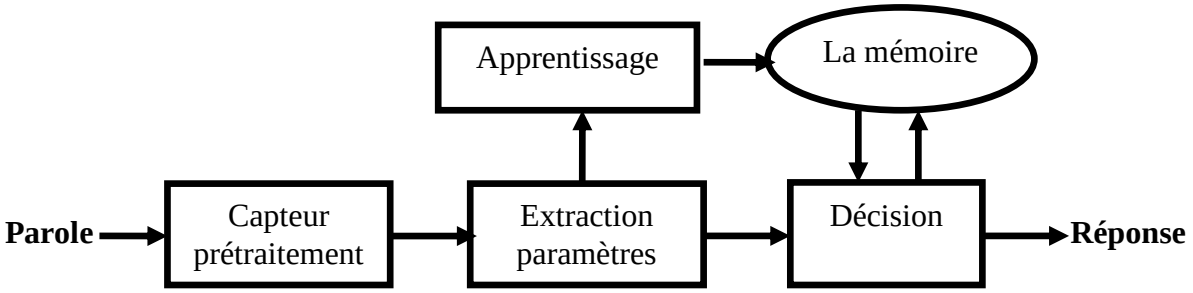


Figure 1.4: système de la reconnaissance vocale numérique.

1.4.3. La reconnaissance globale

Dans cette approche globale, dite aussi acoustique, on considère le message, mot ou groupe de mots, comme une forme insécable en lui attribuant une classe d'appartenance : le mot ou la

phrase sont donc les unités de base du décodage et ne sont définis qu'à partir de paramètres purement acoustiques numérisés **[4]**.

La reconnaissance globale comprend deux phases distinctes :

La phase d'apprentissage pendant laquelle un ou plusieurs locuteurs prononcent une ou plusieurs fois chacun des mots de l'application prévue. Ces prononciations sont toutes prétraitées puis conservées telles quelles ou bien moyennées dans un dictionnaire de références en tant que “ images acoustiques ”.

Puis la phase de reconnaissance où le signal à reconnaître subit le même prétraitement que la phase précédente. Il est ensuite comparé aux références contenues dans le dictionnaire. Le calcul d'une « distance » et sa comparaison à un seuil permet ou non de retenir la ou les références les plus proches.

Mais les différences de prononciations et les variations de débit d'élocution, parfois importantes et non linéaires imposent l'utilisation d'algorithmes de comparaison tels que la comparaison dynamique ou les chaînes de Markov.

C'est une méthode bien adaptée aux applications mono locuteur, à faible vocabulaire et plutôt à mots isolés.

1.4.4. La reconnaissance analytique

Par cette approche, appelée aussi analyse **phonétique,** on considère la segmentation du message en **constituants élémentaires** tels que les phonèmes, les di phonèmes ou les tris phonèmes (Figure 1.5). En effet, ces éléments présentent l'avantage d'être en nombre réduit : 37 phonèmes permettent de décrire le français parlé et une analyse statistique réalisée au LIMSI a montré, qu'à partir d'un répertoire de 627 di phonèmes, il était possible de reconstituer n'importe quelle phrase en français.

Quant aux tris phonèmes, ou triplet phonétique, il est constitué d'un phonème et de ses transitions antérieures et postérieures. Ils sont bien sûr en plus grand nombre, mais ils ont l'avantage de prendre en compte la coarticulation des phonèmes.
Le caractère continu du signal vocal complique beaucoup la reconnaissance de la parole : aucun indice acoustique ne permet de localiser les frontières de mots. Ce problème est abordé, après la phase de prétraitement, d'une part par un décodage acoustique phonétique (DAP) permettant la transcription de la phrase sous forme d'une suite d'éléments phonétique du langage, et d'autre part par un traitement linguistique faisant appel à diverses sources d'informations (lexicales, syntaxiques, sémantiques) permettant la reconnaissance des mots.
Ce sont donc des systèmes à architectures logicielles complexes à plusieurs sources de connaissances qui pallient aux problèmes de reconnaissances des phrases **[4]**.

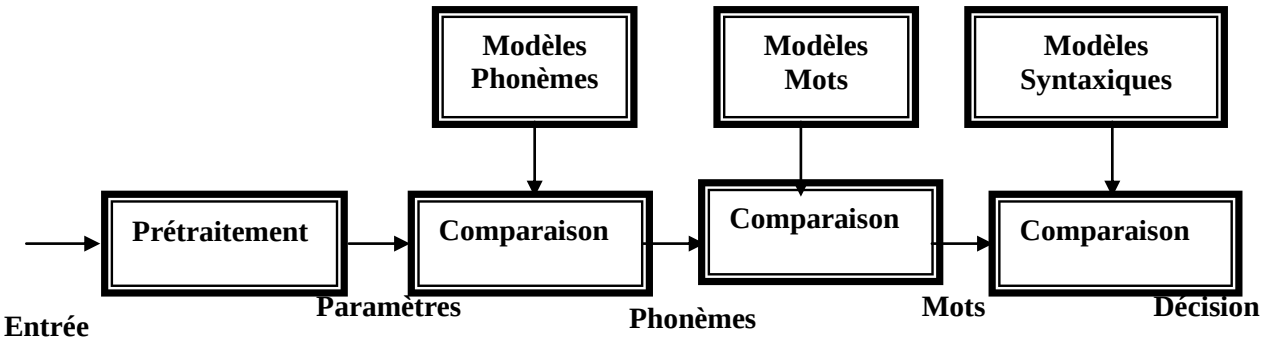


Figure 1. 5 : Schéma synoptique d'un système de reconnaissance de parole selon une approche comparaison.

2. Traitement de la parole

2.1. Introduction

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage (c'est-à-dire du traitement de données symboliques), cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications. L'importance particulière du traitement de la parole dans ce cadre plus général s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine **[1].**

2.2. Le niveau acoustique

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La phonétique acoustique étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone (lui-même associé à un préamplificateur).

De nos jours, le signal électrique résultant est le plus souvent numérisé. Il peut alors être soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les traits acoustiques : sa fréquence fondamentale, son énergie, et son spectre. Chaque trait acoustique est lui-même intimement lié à une grandeur perceptuelle : pitch, intensité, et timbre.

L'opération de numérisation, schématisée à (la figure 1.6), requiert successivement : un filtrage de garde, un échantillonnage, et une quantification.

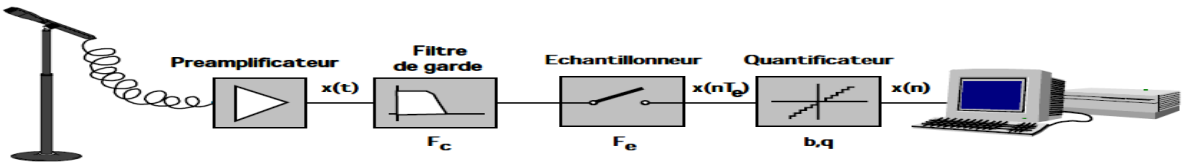


Figure 1.6: Enregistrement numérique d'un signal acoustique.

La fréquence de coupure du filtre de garde, la fréquence d'échantillonnage, le nombre de bits et le pas de quantification sont respectivement notés fc, fe, b, et q **[3]**.

2.2.1. Audiogramme

L'échantillonnage transforme le signal à temps continu x (t) en signal à temps discret X (nTe) défini aux instants d'échantillonnage, multiples, entiers de la période d'échantillonnage Te, celle-ci est elle-même l'inverse de la fréquence d'échantillonnage fe.

Pour ce qui concerne le signal vocal, le choix de fe résulte d'un compromis. Son spectre peut s'étendre jusque 12 kHz **[3]**.La figure 1.7 représente l'évolution temporelle, ou audiogramme du signal vocal pour les mots "بسم الله الرحمان الرحيم".

Figure 1.7 : Audiogramme de signaux de parole.

2.2.2. Transformée de Fourier à court terme

La transformée de Fourier à court terme est obtenue en extrayant de l'audiogramme une trentaine de ms de signal vocal, en pondérant ces échantillons par une fenêtre de pondération (souvent une fenêtre de Hamming) et en effectuant un transformée de Fourier sur ces échantillons.(La figure 1.9) illustre la transformée de Fourier d'une tranche voisée et celle d'une tranche non voisée. Les parties voisées du signal apparaissant sous la forme de successions de pics spectraux marqués, dont les fréquences centrales sont multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière. La forme générale de ces spectres, appelée **enveloppe spectrale**, présente elle-même des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal et sont appelés **formants** et **anti-formants [2]**.

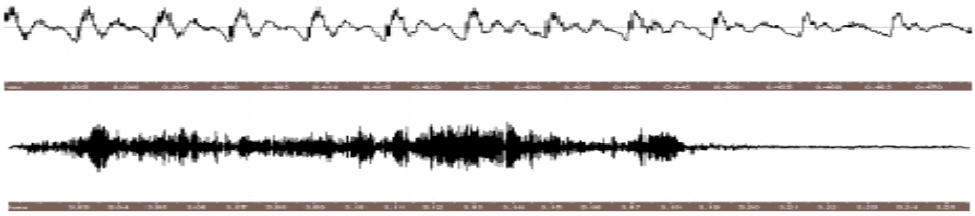


Figure 1.8: Exemples de son voisé (haut) et non voisé (bas) **[1]**.

L'évolution temporelle de leur fréquence centrale et de leur largeur de bande détermine le timbre du son. Il apparaît en pratique que l'enveloppe spectrale des sons voisés est de type passe bas, avec environ un formant par kHz de bande passante, et dont seuls les trois ou quatre premiers contribuent de façon importante au timbre. Par contre, les sons non-voisés présentent souvent une accentuation vers les hautes fréquences **[2]**

2.2.3. Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un **spectrogramme**. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. On parle de spectrogramme **à large bande** ou **à bande étroite** selon la durée de la fenêtre de pondération. Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms), ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales **[3]**.

2.2.4. Fréquence fondamentale

Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la fréquence fondamentale ou pitch. La figure 1.10 donne l'évolution temporelle de la fréquence fondamentale de la phrase « les techniques de traitement de la parole ». On constate qu'à l'intérieur des zones voisées la fréquence fondamentale évolue lentement dans le temps.
Elle s'étend approximativement de 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes, et de 200 à 600 Hz chez les enfants **[1]**.

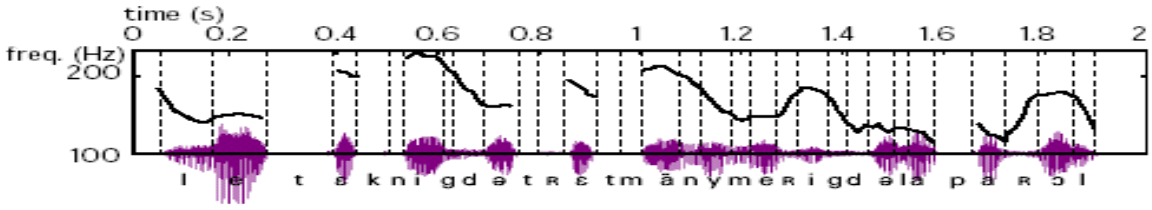


Figure 1.9 : Evolution de la fréquence de vibration des cordes vocales.

2.3. Généralités sur le signal vocal

La parole transmet la pensée, considérée comme information, à travers le canal acoustique par l'intermédiaire de sons articulés différenciés. Cette information transmise est de trois dimensions, vu qu'elle représente trois composantes d'informations transmises, dont la plus importante est l'information linguistique qui correspond souvent à la signification d'une articulation. La deuxième composante étant l'information sociolinguistique (région de l'interlocuteur et sa classe sociale). La dernière composante correspond à l'information personnelle de l'interlocuteur (son identité, sa qualité de voix et ses habitudes d'articulation).Ces trois composantes d'information sont combinées en un seul signal de parole.

Le signal vocal transmet simultanément deux types de messages : un message sémantique convoyé par la parole, expression verbale de la pensée, et un message esthétique perceptible au travers des qualités esthétiques de la voix (timbre, intonation, débit,.. etc.) **[5].**

2.3.1. Caractéristique d'un signal vocale

En plus de ses caractéristiques, en tant qu'onde longitudinale (l'oscillation a la même direction que la propagation), qui sont la fréquence, la longueur d'onde, et la vitesse de propagation qui dépend du milieu matériel de propagation. Le son a par conséquent d'autres caractéristiques, qui sont :

2.3.1.1. La hauteur

C'est la qualité qui fait distinguer un son gras d'un son aigu .La hauteur d'un son est liée à la fréquence des vibrations de la source sonore .Les sons aigus sont dus aux mouvements vibratoires de fréquence élevée, à la différence des sons graves qui sont dus aux mouvements de basse fréquence. **[3].**

2.3.1.2. L'intensité

C'est la qualité qui fait distinguer un son fort d'un faible. L'intensité est liée à la pression de l'air en amont du larynx, qui fait varier l'amplitude des vibrations sonores **[3]**.

2.3.1.3. Le timbre

Le timbre est l'ensemble des caractéristiques qui permettent de différencier une voix.Il provient en particulier de la résonance dans la poitrine, la gorge la cavité buccale et le nez sont les amplitudes relatives des harmoniques du fondamental qui déterminent le timbre du son.

Les éléments physiques du timbre comprennent :

- Les relations entre les parties du spectre, harmoniques ou non.
- les bruits existant dans le son (qui n'ont pas de fréquence particulière, mais dont l'énergie est limitée à une ou plusieurs bandes de fréquence).
- L'évolution dynamique globale du son.
- L'évolution dynamique de chacun des éléments les uns par rapport aux autres [20].

2.3.1.4. Fréquence et amplitude

La répétition d'une forme d'onde périodique est appelée un cycle, et la fréquence fondamentale de la forme d'onde est le nombre de cycles qui se produit par seconde.
Lorsque la longueur du cycle appelée longueur d'onde ou période augmente, la fréquence en cycles par seconde diminue et vice versa.
Nous substituons Hz pour 'cycle par seconde' en conformité avec la terminologie standard de l'acoustique (Hz est une abréviation de Hertz) d'après le nom de l'acousticien allemand Heinrich HERTZ) **[3]**.

2.3.1.5. Le théorème de l'échantillonnage

Définit la relation entre le taux d'échantillonnage et la largeur de bande du signal transmis.
Il fut énoncé par Harold NYQUIST (1928) comme suit :
« Pour toute déformation donnée du signal reçu, le domaine de fréquence transmis doit être augmenté en proportion directe avec la vitesse du signal. La conclusion est que la largeur de fréquence est directement proportionnelle à la vitesse. » Le point essentiel du théorème de l'échantillonnage peut être établi précisément comme ceci :
«Afin d'être capable de reconstruire un signal, la fréquence d'échantillonnage doit être le double de la fréquence du signal échantillonné ».
En raison de sa contribution à la théorie de l'échantillonnage, la plus haute fréquence qui puisse être produite dans un système audionumérique. C'est-à-dire la moitié du taux d'échantillonnage est appelée 'la fréquence de NYQUIST'. Dans les applications musicales, la fréquence de NYQUIST est en général dans le domaine supérieur à celui de l'écoute humaine, au dessus de 20 KHZ. Ainsi la fréquence d'échantillonnage peut être spécifié comme étant au moins le double : au dessus de 40 KHZ **[5]**.

2.3.1.6. Fréquence d'échantillonnage idéale

La question de savoir quelle fréquence d'échantillonnage est idéale pour l'enregistrement et la reproduction musicale de haute qualité est un débat encore en cours.
L'une des raisons est que la théorie mathématique et la pratique des ingénieurs rentrent souvent en conflit :

Les horloges des convertisseurs ne sont pas stables, leurs voltages ne sont pas linéaires, les filtres introduisent de la distorsion de phase et ainsi de suite. Une autre des raisons est que beaucoup de personnes entendent des informations 'on emploie alors le terme 'ambiance' 'dans la région située autour de la 'limite' humaine d'écoute de 20KHZ**.**

Dans les applications d'échantillonnage et de déplacement des hauteurs, le manque de hauteur libre nécessite un filtrage passe bas des échantillons avant que ceux-ci ne soient déplacés vers le haut. Il est clair que des enregistrements à un taux d'échantillonnage élevé sont préférables d'un point de vue artistique, bien qu'ils posent des problèmes pratiques de stockage et la nécessité d'avoir des systèmes de reproduction de haute qualité afin que cet effort en vaille la peine **[3].**

2.3.1.7. Quantification

L'échantillonnage à intervalles de temps discrets dont nous avons parlé dans les parties précédentes, constitue l'une des différences majeures entre les signaux analogiques et les signaux numériques.
Une autre différence est la quantification ou résolution d'amplitude discrète. Les valeurs du signal échantillonné ne peuvent pas prendre n'importe quelle valeur. Ceci en raison du fait que les nombres numériques ne peuvent être représentés qu'à l'intérieur d'un certain domaine et avec une certaine exactitude, qui varie selon le matériel utilisé. Les implications de ceci sont un facteur important de la qualité audionumérique. **[2].**

2.3.1.8. Définition du bruit

On appelle bruit tout phénomène perturbateur (interférence, bruit de fond, etc.) gênant la perception ou l'interprétation d'un signal, ceci par analogie avec les puissances acoustiques de même nom (Figure 1.11) **[3].**

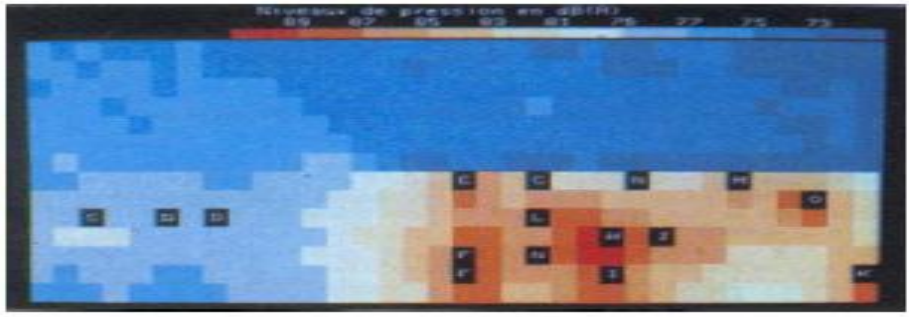


Figure 1.10 : Exemple de carte prévisionnelle de niveaux sonores du plus bruyant (rouge) au plus silencieux (bleu foncé).

2.3.2. Conversion Analogique Numérique

Le son atteint les oreilles de l'auditeur après avoir été transmis par l'air depuis sa source. Les auditeurs entendent des sons car la pression de l'air change légèrement dans leurs oreilles. Si la pression varie selon un modèle répétitif, nous disons que le son a une forme d'onde périodique s'il n'y a pas de modèle discernable, on parle de bruit. Entre ces deux extrêmes se trouve le vaste domaine des sons quasi périodique et quasi bruit eux **[5].**

Un appareil 'le convertisseur Analogique Numérique 'se charge de convertir les tensions en chaînes de nombres binaires à chaque période de l'horloge d'échantillonnage. Les nombres binaires sont stockés sur un support d'enregistrement numérique sorte de mémoire (Figure 1.12)**.**

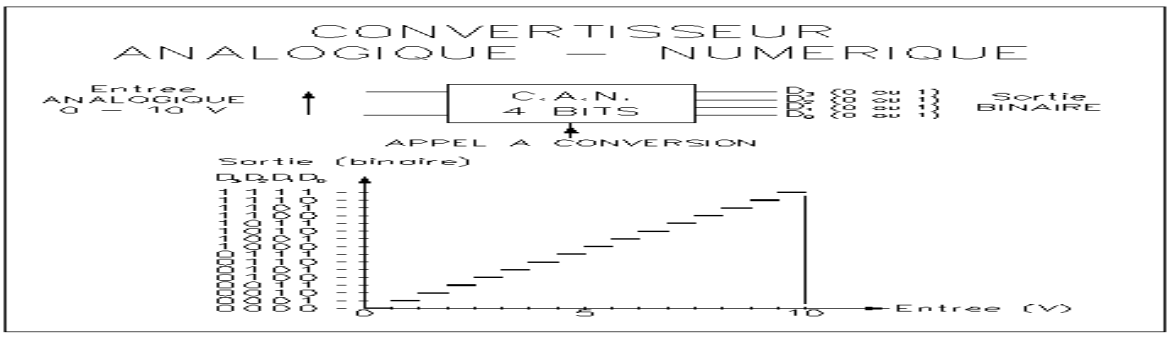


Figure 1.11: Conversion analogique numérique.

2.3.3. Méthodes d'analyse du signal vocal

L'arsenal des méthodes d'analyse et de traitement du signal est considérable. Nous présentons les méthodes générales couramment utilisées pour l'analyse du signal de parole, puis les méthodes utilisées pour l'extraction des paramètres.

Les méthodes d'analyse du signal de la parole peuvent être divisées en deux grandes classes, les méthodes paramétriques et les méthodes non paramétriques **[6].**

2.3.3.1. Les méthodes temporelles

Le signal de parole jouissant de quelques propriétés, exploitables à partir de la représentation temporelle. Les méthodes d'analyses temporelles se basent essentiellement sur la mesure du max, du min, du nombre des passages par zéro,de la fonction d'auto corrélation,du calcul de l'énergie et autres.

Du fait que le signal vocal est considéré comme étant un signal quasi-stationnaire, le traitement doit se faire sur des tranches de 5 à 30ms.

2.3.3.2. Les méthodes d'analyse spectral

Les propriétés spectrales du signal vocal présentent un intérêt majeur pour la perception auditive. L'analyse spectrale est une technique qui met en évidence les caractéristiques fréquentielles des signaux. Elle est souvent utilisée dans les techniques d'analyse synthèse du signal vocal, notamment dans l'analyse par banc de filtre **[6].**

2.3.3.3. Les méthodes non paramétriques

Ces méthodes sont basées principalement sur le calcul de la transformée de Fourrier, soit sur le signal directe, soit sur sa fonction d'auto corrélation. Le calcul de la TF permet l'obtention de la densité spectrale de puissance, qui nous mène ainsi l'extraction des paramètres nécessaires à l'analyse et la synthèse du signal vocal **[6].**

2.3.3.4. Les méthodes paramétriques

Les techniques basées sur l'analyse spectrale présentent quelques limitations, liées à l'hypothèse que le signal est nul au-delà de la fenêtre d'analyse. Pour remédier à ce problème, des méthodes paramétriques sont apparues. Parmi ces méthodes ont trouve les méthodes dites autorégressive **[6].**

2.3.3.5. Le codage Prédictif Linéaire (LPC)

C'est une méthode de type essentiellement temporel qui permet de calculer des coefficients appelés coefficients de la prédiction linéaire.

2.4. Méthodes d'extraction des paramètres

Ces méthodes consistent à extraire les paramètres essentiels qui caractérisent généralement le signal de parole à savoir l'énergie, la fréquence fondamentale et les formants.

2.4.1. Extraction de la fréquence fondamentale (pitch)

L'extraction de la fréquence fondamentale (ou pitch) est comme son nom l'indique fondamentale. Les variations de la fondamentale pour un locuteur donné constituent ce qu'on

appelle la prosodie. Celle-ci influe considérablement sur l'oreille humaine pour permettre la différentiation entre locuteurs et ainsi la reconnaissance du locuteur **[6].**

2.4.2. Extraction des formants

Le début d'utilisation des méthodes d'extraction des formants remonte à 1934. Ces formants sont les résonances du conduit vocal considéré comme un filtre et correspondant aux pôles de la fonction de transfert de ce dernier. Ce sont des paramètres privilégiés dans l'étude et l'analyse de la parole, ils apparaissent plus clairement pour les sons voisés **[6].**

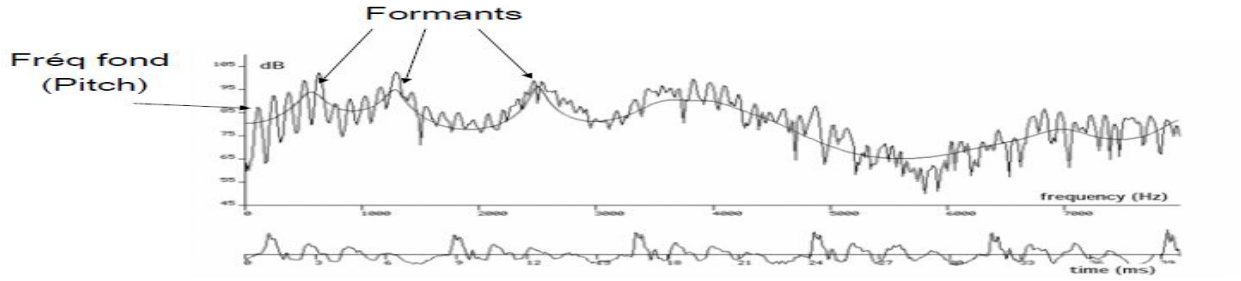


Figure 1.12 : Exemple d'extraction des paramètres.

2.5. Conclusion

Les méthodes illustrées si dessus sont destinées à l'analyse et l'extraction des paramètres à savoir l'énergie, la fréquence fondamentale et les formants. L'extraction de ces paramètres revêt une importance remarquable, notamment en synthèse et en reconnaissance de la parole. Les méthodes précédentes et beaucoup d'autres s'identifient à leurs avantages et leurs inconvénients propres. La qualité d'une méthode peut se chiffrer par les critères suivants :

- La précision de la méthode et sa fiabilité dans la détection du fondamentale et les formants pour différents locuteurs et pour larges gammes de sons.
- Le temps de calcul et les possibilités de câblage en temps réel.
- Le coût de l'implantation (câblage) relativement à son usage et à ses utilisateurs.
- La méthode que nous allons utiliser pour extraction de caractéristiques d'un signal vocal est Le codage prédictif linéaire LPC, car elle a prouvé son efficacité au niveau de l'identification et la classification de locuteurs **[5].**

3. Reconnaissance de la parole

3.1. Introduction

Notre Dieu tout puissant, nous a délégué un organisme biologique très complexe et très développé, notre espèce humaine est l'unique à être privilégiée de « la pensé », le message représentant cette dernière, en générale, peut prendre trois aspects, l'aspect écrit, l'aspect signé et celui verbal, en prenant le dernier aspect, la forme la plus simple qui le concrétise est **la parole.**

L'expression répandue 'ce ne sont que des paroles' banalise ce terme, cependant nous apercevons son véritable poids du point de vue phénomène à étudier, seulement chez les chercheurs de ce domaine, car la parole pour eux est un phénomène très complexe, non seulement en tenant compte de la difficulté du mécanisme interne qui la génère et celui qui la transmet, mais aussi de l'entrave de l'organisme qui la reconnaît.

Sans sa reconnaissance, la parole n'a aucun sens, car elle est produite pour être reconnue dans le but de transmettre une pensé précise afin de satisfaire un certain besoin.

Par l'esprit scientifique, de savoir, de développement et de création, les chercheurs optent pour dépasser la compréhension de la communication entre Hommes, en se dirigeant vers un nouvel horizon qui englobe la reconnaissance automatique de la parole.

La reconnaissance automatique de la parole est la manière évoluée pour établir un dialogue artificiel « Homme-Machine », dans le but d'adapter une machine à un vocabulaire limité, qui traduit un besoin issu d'un locuteur.

Le domaine de cette application peut atteindre plusieurs domaines tels que : le pilotage d'avion, la composition du numéro téléphonique du correspondant, faire acquérir des informations à un PC, différentes aides à des handicapés …etc.

Ce domaine fait l'objectif des chercheurs depuis de longues années, par conséquent un bon nombre de méthodes sont incorporées, telles que les méthodes globales, analytiques, probabilistes et encore les méthodes connexionnistes qui sont adoptées depuis les années quarante **[6].**

3.2. Définition

La reconnaissance automatique de la parole est l'un des deux domaines du traitement automatique de la parole, l'autre étant la synthèse vocale.

La reconnaissance automatique de la parole permet à la machine de comprendre et de traiter des informations fournies oralement par un utilisateur humain. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots mais aussi, plus récemment, de phonèmes (unité sonore minimale). En revanche, le système de synthèse de la parole permet de reproduire d'une manière sonore un texte qui lui est soumis, comme un humain le ferait.

Ces deux domaines et notamment la reconnaissance vocale, font appel aux connaissances de plusieurs sciences : l'anatomie (les fonctions de l'appareil phonatoire et de l'oreille), les signaux émis par la parole, la phonétique, le traitement du signal, la linguistique, l'informatique, l'intelligence artificielle et les statistiques. Il faut bien distinguer ces deux mondes : un système de synthèse vocale peut très bien fonctionner sans qu'un module de reconnaissance n'y soit rattaché. Evidemment le contraire est également tout à fait possible. Par contre, dans certains

domaines bien précis, l'un ne va pas sans l'autre. Il est bien entendu que l'étude se portant sur la reconnaissance automatique de la parole, l'autre aspect du traitement de la parole.

Le traitement automatique de la parole ouvre des perspectives de nouveaux comptes tenus de la différence considérable existant entre la commande manuelle et vocale.

L'utilisation du langage naturel dans le dialogue personne/machine met la technologie à la portée de tous et entraîne sa vulgarisation, en réduisant les contraintes de l'usage des claviers, souris et codes de commandes à maîtriser. En simplifiant le protocole de dialogue personne/machine, le traitement automatique de la parole vise donc aussi un gain de productivité puisque c'est la machine qui s'adapte à l'homme pour communiquer, et non l'inverse. De plus, il rend possible l'utilisation simultanée des yeux ou des mains à une autre tâche.
Il permet d'humaniser les systèmes informatiques de gestion de l'information, en axant leur conception sur les utilisateurs **[3].**

Figure 1.13 : Processus de reconnaissance de la parole.

3.3. Principe de fonctionnement

3.3.1. Problématique

Pour bien appréhender le problème de la reconnaissance automatique de la parole, il est bon d'en comprendre les différents niveaux de complexité et les différents facteurs qui en font un problème difficile. Le système doit-il être optimisé pour un unique locuteur ou est-il destiné à devoir se confronter à plusieurs utilisateurs ?

On peut aisément comprendre que les systèmes dépendants d'un seul locuteur sont plus faciles à développer et sont caractérisés par de meilleurs taux de reconnaissance que les systèmes indépendants du locuteur étant donné que la variabilité du signal de parole est plus limitée.

Cette dépendance au locuteur est cependant acquise au prix d'un entraînement spécifique à chaque utilisateur. Ceci n'est néanmoins pas toujours possible. Par exemple, dans le cas d'applications téléphoniques, on comprend bien que les systèmes puissent être utilisés par n'importe qui et donc être indépendants du locuteur. Bien que la méthodologie de base reste la même, cette indépendance au locuteur est obtenue par l'acquisition de nombreux locuteurs (couvrant si possible les différents dialectes) qui sont utilisés simultanément pour l'entraînement de modèles susceptibles d'en extraire toutes les caractéristiques majeures. Une solution intermédiaire parfois utilisée consiste à développer des systèmes capables de s'adapter rapidement (de façon supervisée ou non) au nouveau locuteur. Le système est-il robuste ?

Autrement dit, le système est-il capable de fonctionner proprement dans des conditions difficiles? En effet, de nombreuses variables pouvant affecter significativement les performances des systèmes de reconnaissance ont été identifiées :

Bruits d'environnement (dans une rue, un bistrot, etc.…).

- Déformation de la voix par l'environnement (réverbérations, échos, etc.…).
- Qualité du matériel utilisé (micro, carte son, etc.…).
- Bande passante fréquentielle limitée (fréquence limitée d'une ligne téléphonique).
- Elocution inhabituelle ou altérée (stress, émotions, fatigue, etc.…) **[6].**

3.3.2. Fonctionnement

Le problème de la reconnaissance automatique de la parole consiste à extraire l'information contenue dans un signal de parole, typiquement par échantillonnage du signal électrique obtenu à la sortie d'un microphone, afin qu'il puisse être comparé à des modèles sous forme numérique. Parmi plusieurs techniques de reconnaissance, il y en a deux qui sont majoritairement utilisées afin de parvenir à résoudre ce problème : la comparaison à des exemples et la comparaison d'unités de parole **[6].**

3.3.2.1. Reconnaissance par comparaison à des exemples

Les premiers succès en reconnaissance vocale ont été obtenus dans les années 70 à l'aide d'un paradigme de reconnaissance de mots. L'idée, très simple dans son principe, consiste à faire prononcer un ou plusieurs exemples de chacun des mots susceptibles d'être reconnus, et à les enregistrer sous forme de vecteurs acoustiques (représentation numérique du signal sonore).

Puisque cette suite de vecteurs acoustiques caractérise complètement l'évolution de l'enveloppe spectrale du signal enregistré, on peut dire qu'elle correspond à l'enregistrement d'un spectrogramme.
L'étape de reconnaissance proprement dite consiste alors à analyser le signal inconnu sous la forme d'une suite de vecteurs acoustiques similaires, et à comparer la suite inconnue à chacune des suites des exemples préalablement enregistrés. Le mot «reconnu» sera alors celui dont la suite de vecteurs acoustiques «spectrogramme» colle le mieux à celle du mot inconnu. Il s'agit en quelque sorte de voir dans quelle mesure les spectrogrammes se superposent **[6].**

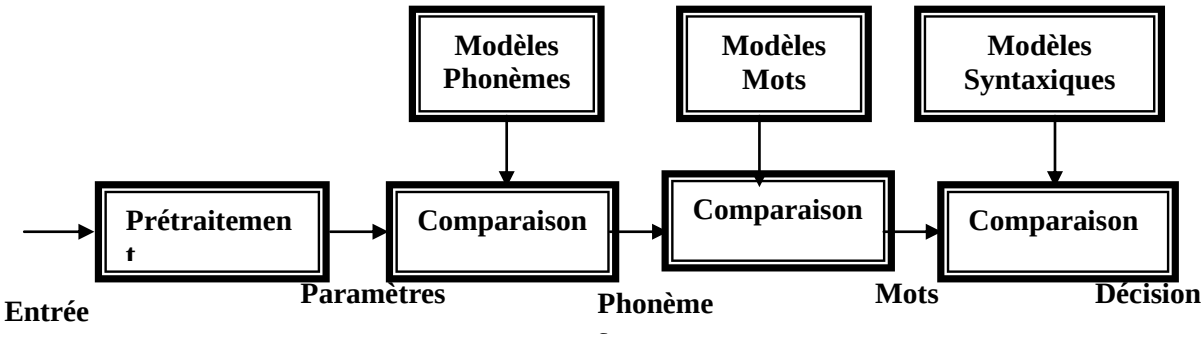


Figure 1.14 : Schéma synoptique d'un système de reconnaissance de parole selon une approche comparaison.

3.3.2.2. Reconnaissance par modélisation d'unités de parole

La plupart des systèmes de reconnaissance de la parole sont de nos jours basés sur ce mode là. Dés que l'on cherche à concevoir un système réellement multi locuteur, à plus grand vocabulaire et s'adaptant facilement à une application, il devient nécessaire de mener la reconnaissance sur base d'unités de parole de plus petite taille, que l'on appelle phonèmes.

En effet, la parole est constituée d'une suite de sons élémentaires : «a», «é», «ss». Ils sont produits par la vibration des cordes vocales. Ces sons mis bout à bout composent des mots. On ne se contente plus alors d'exemples de ces unités, mais on cherche plutôt à en déduire un modèle (un modèle par unité), qui sera applicable pour n'importe quelle voix. Il apparaît ainsi dans de nombreuses publications que l'on peut décomposer la reconnaissance de la parole en quatre modules **[7].**

Un module d'acquisition et de modélisation du signal qui transforme le signal de parole en une séquence de vecteurs acoustiques. Pour être utilisable par un ordinateur, un signal doit tout d'abord être numérisé. Cette opération tend à transformer un phénomène temporel analogique, le signal sonore dans notre cas, en une suite d'éléments discrets, les échantillons. Ceux-ci sont obtenus avec une carte spécialisée courante de nos jours dans les ordinateurs depuis l'avènement du multimédia. La numérisation sonore repose sur deux paramètres : la quantification et la fréquence d'échantillonnage.

La quantification définit le nombre de bits sur lesquels on veut réaliser la numérisation. Elle permet de mesurer l'amplitude de l'onde sonore à chaque pas de l'échantillonnage. Le choix de la fréquence d'échantillonnage est aussi déterminant pour la définition de la bande passante représentée dans le signal numérisé.

Un module acoustique qui peut produire une ou plusieurs hypothèses phonétiques pour chaque segment de parole (par exemple de 10 ms, pour chaque vecteur acoustique), associées en général à une probabilité. Ce générateur d'hypothèses est généralement basé sur des modèles statistiques de phonèmes, qui sont entraînés sur une grande quantité de données de parole (par exemple, enregistrement de nombreuses phrases) contenant plusieurs fois les différentes unités de parole dans plusieurs contextes différents. Ces modèles statistiques sont le plus souvent constitués de lois statistiques paramétriques dont on ajuste les paramètres pour «coller» au mieux aux données ou de réseaux de neurones artificiels.

Un module lexical dans le cadre de la reconnaissance de la parole continue, même si le système acoustique est basé sur des phonèmes, il faut obtenir, pour chaque entrée du dictionnaire phonétique, un modèle qui lui est propre. Un tel module lexical embarque en général des modèles des mots de la langue (les modèles de base étant de simples dictionnaires phonétiques, les plus complexes sont de véritables automates probabilistes, capables d'associer une probabilité à chaque prononciation possible d'un mot). A l'issue de ce module, il peut donc y avoir plusieurs hypothèses de mots qui ne pourront être départagées que par les contraintes syntaxiques.

Un module syntaxique qui interagit avec un système d'alignement temporel pour forcer la reconnaissance à intégrer des contraintes syntaxiques, voire sémantiques. Les connaissances syntaxiques sont généralement formalisées dans un modèle de la langue, qui associe une probabilité à toute suite de mots présents dans le lexique. Ainsi le système est capable de choisir entre plusieurs mots selon le contexte de la phrase ou du texte en cours, et de son modèle lexical.On peut ajouter à cela un module de filtrage pouvant corriger le signal après l'acquisition afin de retirer les distorsions ou les bruits provenant du matériel ou de l'environnement du locuteur. Ce module est aussi appelé «traitement du canal de transmission». Du fait de sa complexité et du peu d'amélioration qu'il apporte, ce module n'est pas toujours intégré aux systèmes. Cependant la recherche de meilleurs traitements du canal de transmission sera sûrement nécessaire à l'amélioration des systèmes de reconnaissance vocale.

3.4. Reconnaissance de petits vocabulaires

Ça concerne la reconnaissance de mots isolés, multi locuteurs dans des conditions difficiles, par exemple : reconnaissance de chiffres à travers le réseau téléphonique **[6].**

3.5. Reconnaissance de petits vocabulaires de mots isolés

La reconnaissance de mots isolés, le plus souvent mono locuteur, pour des vocabulaires de quelques dizaines jusqu'à quelques centaines de mots est un problème assez bien résolu. Les premiers systèmes commerciaux de cette catégorie sont apparus il y a un peu plus de vingt ans **[3].**

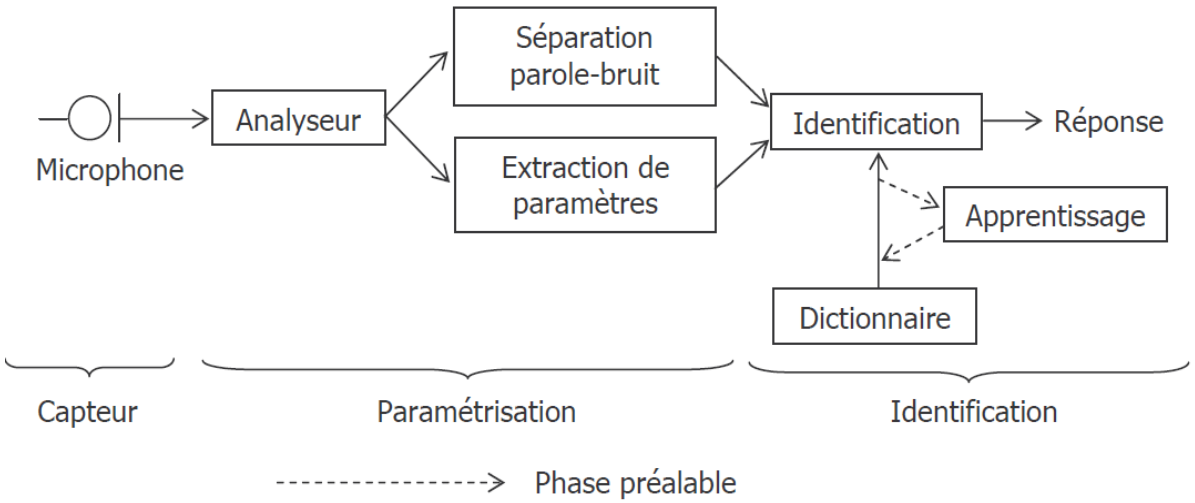


Figure 1.15 : Système de reconnaissance de mots isolés **[7]**.

3.6. Reconnaissance de grands vocabulaires

Par exemple par IBM, Kurzeil et Dragon systèmes mono locuteur en particulier pour des tâches de dictée de textes dans des domaines d'application fixés .Des systèmes de ce type sont présentés, fondés sur une modélisation stochastique de la parole, méthode actuellement la plus

performante. Dans cette catégorie apparaissent aussi des systèmes utilisant une reconnaissance phonétique des mots, c'est notamment le cas d'un produit de speech system **[3].**

Microsoft, en passant par Apple et IBM, de nombreux industriels travaillent sur des projets de reconnaissance vocale, généralement en complément d'une activité de recherche sur la synthèse de la parole, le tout s'insérant dans des projets plus généraux d'interface Homme Machine.

Il faudra attendre encore plus longtemps avant que la machine remplace purement et simplement la secrétaire dactylo pour la saisie de textes sur ordinateur. Les systèmes de reconnaissance vocale actuels sont encore bien trop grossiers pour comprendre toutes les finesses qui peuvent se glisser dans la **syntaxe** et dans les intonations de la langue parlée en continu et non plus sous la forme de mots clés ou de petites phrases sommaires **[3].**

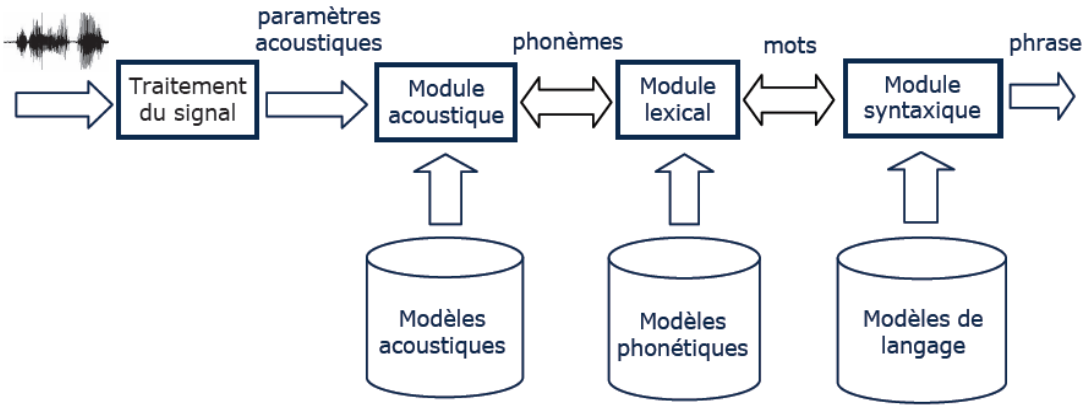


Figure 1.16 : Système de reconnaissance de grands vocabulaires.

3.7. Caractéristiques du système de reconnaissance de la parole

On peut caractériser un tel système par **:**

3.7.1. Le mode de fonctionnement

La reconnaissance de la parole pour la voix d'un seul locuteur est déjà en elle-même un problème élémentaire, en raison de la variabilité intra locuteur, inhérent au processus humain de production. Cette variation est liée aux différences du débit d'élocution, à l'émotion, au stress, aux rhumes, … etc **[3].**

3.7.2. Le mode d'élocution

Si le locuteur marque une pause après chaque mot de l'énoncé, la complexité du problème est réduite, puisque les frontières du mot sont alors disponibles (contrairement au cas de la parole continue).
Dans ce cas on parle des mots isolés. Les phrases et les mots enchaînés présentent d'autres genres de modes d'élocution **[3].**

3.7.3. La taille du vocabulaire

Il est plus facile de travailler sur un vocabulaire très étendu (quelques milliers ou dizaines de milliers de mots) que sur vocabulaire très restreint (quelques dizaines de mots).

La taille du vocabulaire est cependant un paramètre insuffisant, un ensemble de mots très différents les uns des autres étant plus faciles à traiter que des mots proches phonétiquement.

3.7.4. Le langage

La prise en compte de la syntaxe du langage produit par l'utilisateur sera plus facile pour un langage ' rigide', très contraint, que si toute la souplesse de la langue naturelle palée peut être rencontrée.

3.7.5. Le mode de décodage de l'information

Ce mode de décodage de l'information est lié aux types d'approches utilisées pour le traitement du signal :

3.7.5.1. Approche analytique

Cette approche utilise tout d'abord une segmentation à priori du signal en unités de tailles phonétiques, puis chacun des segments est identifié en comparant les mesures acoustiques à des formes de reconnaissance **[5].**

3.7.5.2. Approche globale

Approche globale ou pragmatique consiste à faire totalement abstraction des phonèmes linguistiques pour ne retenir que l'aspect acoustique de la parole.

Elle applique des hypothèses simplificatrices (au risque de dénaturer la communication vocale) au problème de façon à le rendre plus abordable **[5].**

3.7.6. L'environnement

Indépendamment de ce qui précède, l'environnement acoustique et les conditions de prise du son, constituent un facteur important : la présence du bruit, même stationnaire, dégrade en général fortement les performances des systèmes de reconnaissance. De plus si ce bruit est intense, il induit une augmentation de la variabilité chez le locuteur **[3].**

3.8. Reconnaissance de la parole continue

Tout d'abord, qu'est ce que la parole continue ? C'est un discours, des phrases où les mots s'enchaînent sans moyen de séparer, contrairement aux mots isolés. Le but de cette partie n'est pas de rentrer dans les détails de la programmation d'un logiciel de reconnaissance de la parole

continue, cela serait trop long et fastidieux. On va donc présenter les " ficelles " de la reconnaissance de la parole continue de manière très générale.

Les objectifs de cette partie étant donc éclaircis, on peut entamer la réflexion autour de la reconnaissance de la parole continue. Pourquoi, après tout, s'évertuer à attribuer à une machine de telles capacités ? Est-ce par pure fantaisie que les auteurs de science-fiction inventent des dialogues entre un héros et sa machine ? Non, ceci relève d'un besoin qui pourrait se résumer à une chose : la recherche d'un confort et d'amélioration de l'interaction de l'homme avec la machine. Les avantages d'un tel progrès sont simples à imaginer **[6].**

3.9. Quelques applications

De façon générale, le choix d'une application doit faire l'objet d'une étude attentive, fondée sur un ensemble de critères objectifs. En particulier, il est important d'examiner si la voix apporte véritablement un accroissement des performances ou un meilleur confort d'utilisation. Par ailleurs, il ne faut pas trop attendre de la commande vocale mais la considérer, en tout état de cause, comme un moyen complémentaire parmi d'autres moyens d'interaction Homme-Machine plus traditionnels**.** Bien entendu, à chaque type d'application correspondent des critères de performance différents. Ainsi, pour des applications en reconnaissance de la parole, on jugera la qualité d'une application sur les quatre critères principaux suivants :

- Le débit du flux de parole correctement reconnu. Si le locuteur prononce les mots séparément avec de petites pauses (environ 200 ms) entre chaque mot, on parlera de reconnaissance par mots isolés, sinon ce sera de la reconnaissance de parole continue.
- La taille du vocabulaire correctement reconnu. Ce vocabulaire variera de quelques mots (la cabine téléphonique à entrée vocale) à plusieurs milliers de mots (la machine à écrire à entrée vocale).
- Les contraintes imposées par le système sur l'environnement de fonctionnement : acceptation de bruits de fond et parasites divers. Des critères de qualité positifs dans certaines applications peuvent être négatifs dans d'autres : l'indifférence au locuteur est recherchée pour une cabine téléphonique à numérotation vocale alors qu'au contraire c'est la capacité de discrimination entre locuteurs qui déterminera la qualité d'une serrure à commande vocale.
- Les contraintes imposées par le système sur l'utilisateur : est-il unique ou multiple, doit-il s'astreindre à une phase d'apprentissage préalable ? **[7].**

3.9.1. Services vocaux

Les serveurs passifs (sans reconnaissance vocale) existent depuis de nombreuses années tels que l'horloge parlante, la météo, les résultats des courses, du loto, etc.… Mais lorsque la quantité d'information est importante, il devient nécessaire pour l'utilisateur de pouvoir sélectionner ce qu'il veut entendre. Dans des cas simples la sélection de touches «multifréquences» (DTMF) peut suffire. Mais pour des applications plus complexes des systèmes sont en voie de développement pour que l'utilisateur puisse naviguer sur un serveur vocal en prononçant les

mots de contrôle de l'application. Ces services pourront s'étendre à tout un ensemble de domaines : la réservation de place d'avion, de train, de théâtre, de chambre d'hôtel, les déclarations de sinistre à l'assureur, les consultations et transactions bancaires, les opérations boursières, la facturation automatique des appels à distance, etc.…**[3].**

3.9.2. Contrôle de qualité, saisie des données

Dans de nombreux environnements de travail la possibilité de décharger le travailleur, grâce à une interface vocale, apporte un gain incontestable de liberté et de rapidité de mouvement. Pendant qu'il observe un processus complexe, il peut par exemple décrire des informations visuelles. Il a aussi la possibilité de commander à distance un automate évoluant en milieu hostile (apesanteur, sous-marin, industrie pétrolière) **[5].**

3.9.3. Avionique

A bord des avions, les tâches étant de plus en plus complexes et le tableau de bord de plus en plus réduit, la parole permet au pilote d'avoir à sa disposition un moyen supplémentaire d'interaction avec la machine, sans cependant gêner l'accomplissement des tâches courantes qui requièrent de sa part toute son attention visuelle. Les autorités canadiennes ont été les précurseurs des techniques vocales dans l'avionique. Ainsi, l'Institut de Recherche Aérospatiale (IRA) a effectué des travaux de recherche sur la technologie vocale depuis la fin des années 70.
Actuellement, un système de reconnaissance de la parole est à bord du dernier des avions de l'armée française, le Rafale, capable de traiter un vocabulaire de près de 300 mots **[5].**

3.9.4. Formation

Les enfants, mais aussi les adultes, sont attirés par des jeux doués de parole (poupées qui parlent, jeux de société, jeux vidéo, jeux éducatifs). L'enseignement assisté par ordinateur et notamment les laboratoires de langue commence à intégrer de plus en plus de possibilités vocales, et évolue vers une interactivité plus grande : les systèmes d'aide à l'apprentissage des langues étrangères, permettant d'acquérir une prononciation correcte, une maîtrise du vocabulaire et de la syntaxe, ne peuvent que bénéficier des technologies vocales qui leur confèrent en outre un aspect ludique.
Du côté des applications proprement dites, la société Auralog a été précurseur avec ses logiciels d'apprentissage de langues (TeLL me More, Atout Clic Anglais). Ainsi, grâce à la technologie avancée de la reconnaissance vocale, l'utilisateur engage un véritable dialogue avec son PC. Suivant son niveau, l'apprenant paramètre la reconnaissance vocale pour la rendre plus tolérante ou plus exigeante quant à la qualité de sa prononciation. L'utilisateur s'entraîne à prononcer une phrase ou un mot et obtient un score lui permettant d'évaluer la qualité de son accent, de sa prononciation et de son intonation **[7].**

3.9.5. Aide aux handicapés

Différents programmes européens ont permis de mieux cerner les différents types d'handicap dont souffre la population, ainsi que le nombre de personnes concernées. On dénombre actuellement en Europe 12 millions de malvoyants dont 1 million de non-voyants, 81 millions de

malentendants, dont 1 million de non entendants, environ 30 millions de personnes ayant un handicap moteur des membres supérieurs et 50 millions ayant un handicap des membres inférieurs. L'intérêt des technologies vocales apparaît évident dans la mesure où celles-ci permettent aux personnes handicapées de retrouver une certaine autonomie et de bénéficier d'une meilleure insertion dans leur environnement tant professionnel que familial, la parole se substituant au sens défaillant.
Ainsi beaucoup de systèmes existent pour cela, tel que le contrôle de fauteuil roulant, le contrôle de fonctions secondaires dans la voiture, le contrôle d'appareil électrique à la maison, le contrôle de l'ordinateur, …etc.**[7].**

3.9.6. Dictée vocale

L'orientation actuelle des logiciels tend de plus en plus à offrir un contrôle total de l'environnement permettant de se passer du clavier et de la souris pour utiliser l'ordinateur. Les nouveaux systèmes d'exploitation couplés aux logiciels à venir devraient enfin permettre d'offrir un ordinateur fonctionnant réellement «sans les mains» **[7].**

3.9.7. Relation avec la télécommunication

Dans le secteur de la téléphonie, les grandes sociétés de télécommunication ont engagé une course à l'innovation. Ainsi, il suffit de dire le nom du correspondant désiré dans le récepteur, à condition de l'avoir préalablement encodé, pour obtenir la communication souhaitée. Ceci peut-être très utile pour téléphoner depuis une voiture.

L'information au public est aussi un domaine concerné par la numérisation de la parole. Dans les gares ou les aéroports, par exemple, on pourra bientôt voir des bornes interactives qui remplaceront les agents préposés aux renseignements. Pour connaître l'horaire d'un train, il suffira de demander de vive voix à la machine où on veut aller et quand, et elle répondra dans la langue de notre choix, avant de nous souhaiter un agréable voyage.

Plus précisément, aujourd'hui, deux gammes de services dominent le marché des services de Télécommunication à commande vocale : ce sont les services à opérateurs partiellement automatisés et les services de répertoires vocaux, évoluant progressivement vers des services plus complets d'assistants téléphoniques **[3].**

3.9.8. Et aussi …

On peut aussi citer les modules de reconnaissance vocale embarqués, comme dans les téléphones mobiles ou les assistants numériques ainsi que les futures possibilités en cours de développement chez les fabricants automobile avec le contrôle de différents éléments de la voiture grand public : autoradio, climatisation, navigation de bord. On peut même de nos jours surfer sur Internet grâce à des commandes vocales, c'est ce que propose la société Interactive Speech. L'utilisation de la reconnaissance automatique de la parole devient courante et devrait très bientôt apparaître dans la plupart des domaines d'activités et la plupart des applications futures **[3].**

Conclusion

Au terme de ce bilan rapide sur la reconnaissance vocale, on a pu constater que ce domaine est particulièrement vaste et qu'il n'existe pas de produit miracle capable de répondre à toutes les applications. Le bruit, par exemple, non traité par ce document, reste un frein à la généralisation des systèmes de reconnaissance.

La reconnaissance vocale reste un compromis entre la taille du vocabulaire, ses possibilités multi locuteur, son encombrement physique, sa rapidité, temps d'apprentissage, et...

La puissance des outils de calcul actuels et les capacités d'intégration des systèmes ont provoqué un regain d'intérêt depuis ces dernières années chez les industriels. En effet, ces derniers voient dans la reconnaissance vocale, « le plus commercial »", permettant de faire la différence avec la concurrence.

Chapitre 2

Support Vector Machine

- **Introduction**
- **Méthodes de classification**
- **Apprentissage statistique et SVM**
- **SVM principe de fonctionnement général**
- **Fondements mathématiques**
- **SVMs et analyse des bases de données**
- **Les domaines d'application**
- **Conclusion**

Introduction

Parmi les méthodes à noyaux, inspirées de la théorie statistique de l'apprentissage de Vladimir Vapnik, les SVM constituent la forme la plus connue. SVM est une méthode de classification binaire par apprentissage supervisé, elle fut introduite par Vapnik en 1995. Cette méthode est donc une alternative récente pour la classification. Cette méthode repose sur l'existence d'un classificateur linéaire dans un espace approprié. Puisque c'est un problème de classification à deux classes, cette méthode fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle. Elle est basée sur l'utilisation de fonctions dites noyau (kernel) qui permettent une séparation optimale des données. Dans la présentation des principes de fonctionnements, nous schématiserons les données par des « points » dans un plan.

La notion d'apprentissage étant importante, nous allons commencer par effectuer un rappel. L'apprentissage par induction permet d'arriver à des conclusions par l'examen d'exemples particuliers. Il se divise en apprentissage supervisé et non supervisé. Le cas qui concerne les SVM est l'apprentissage supervisé. Les exemples particuliers sont représentés par un ensemble de couples d'entrée/sortie. Le but est d'apprendre une fonction qui correspond aux exemples vus et qui prédit les sorties pour les entrées qui n'ont pas encore été vues. Les entrées peuvent être des descriptions d'objets et les sorties la classe des objets donnés en entrée.

1. Méthodes de classification

Parmi les méthodes de classification, deux sont particulièrement classiques dans ce domaine : la méthode des plus proches voisins (*k-ppv*) et les arbres de décision et une méthode moderne appeler SVM.

1.1. K-ppv

La méthode des K plus proches voisins (k-ppv) est une méthode particulièrement élémentaire. Comme son nom l'indique elle consiste à rechercher dans la base d'apprentissage les k individus qui sont les plus proches d'une nouvelle donnée, et la règle de décision consiste à faire un vote majoritaire sur les classes de ces *k-ppv*. On peut noter qu'il est préférable de prendre k impair pour ne pas avoir de problèmes d'égalité lors de la prise de décision. La méthode repose donc sur un critère de similarité qu'il faut définir a priori pour comparer les données. Le seul paramètre à régler est alors le nombre de voisins à considérer. Cette méthode peut paraître élémentaire mais dans de nombreux cas réels elle s'avère efficace, et même plus performante que des modèles plus complexes. Elle peut par conséquent constituer une bonne référence pour quantifier les performances de classification d'autres méthodes **[12]**.

1.2. Arbres de décision

Les arbres de décision représentent une méthode très efficace d'apprentissage supervisé. Il s'agit de partitionner un ensemble de données en des groupes les plus homogènes possible du point de vue de la variable à prédire. On prend en entrée un ensemble de données classées, et on fournit

en sortie un arbre qui ressemble beaucoup à un diagramme d'orientation où chaque nœud final (feuille) représente une décision (une classe) et chaque nœud non final (interne) représente un test. Chaque feuille représente la décision d'appartenance à une classe des données vérifiant tous les tests du chemin menant de la racine à cette feuille [21].
Les arbres de décisions sont très répandus, à cause de la simplicité de lecture de leurs résultats et leur traitement naturels des cas multi classe. Néanmoins, ils posent beaucoup de problèmes tel que :

- La difficulté de manipulation des attributs numériques.
- L'espace nécessaire pour leur déduction.

1.3. Machines à vecteurs de support (SVM)

L'algorithme des machines à vecteurs de support a été développé dans les années 90 par Vapnik. Il a initialement été développé comme un algorithme de classification binaire supervisée. Il s'avère particulièrement efficace de par le fait qu'il peut traiter des problèmes mettant en jeu de grands nombres de descripteurs, qu'il assure une solution unique (pas de problèmes de minimum local comme pour les réseaux de neurones) et il a fourni de bons résultats sur des problèmes réels L'algorithme sous sa forme initiale revient à chercher une frontière de décision linéaire entre deux classes, mais ce modèle peut considérablement être enrichi en se projetant dans un autre espace permettant d'augmenter la séparabilité des données. On peut alors appliquer le même algorithme dans ce nouvel espace, ce qui se traduit par une frontière de décision non linéaire dans l'espace initial.

Nous reviendrons à cette méthode dans ce chapitre puisque cette approche fait précisément l'objet du sujet **[12]**.

2. Apprentissage statistique et SVM

2.1. Objectif de l'apprentissage statistique

Effectuer une classification consiste à déterminer une règle de décision capable, à partir d'observations externes, d'assigner un objet à une classe parmi plusieurs. Le cas le plus simple consiste à discriminer deux classes. D'une manière plus formelle, la classification bi-classe revient à estimer une fonction $f : x \rightarrow \{+1,-1\}$ à partir d'un ensemble d'apprentissage constitué de couples (xi, yi), qu'on suppose i.i.d. suivant une distribution de probabilité P(x, y) inconnue, tels que

$$(xi, yi) \in X \times Y \text{ où } i = 1,\ldots, N \text{ x et } Y = \{+1,-1\},$$

de sorte à ce que f classifie correctement des exemples inconnus(xt, yt). Par exemple, on peut assigner xt à la classe (+1) si $f(xt) \geq 0$, et à la classe (-1) sinon. Les exemples inconnus sont supposés suivre la même distribution de probabilité P(x, y) que ceux de l'ensemble d'apprentissage. La meilleure fonction f est celle obtenue en minimisant le risque :

$$R[f] = \int L\,[f(x), y]\, d\,P(x, y). \qquad (2.1)$$

Où L désigne une fonction de coût, comme par exemple :

$$L\,[f(x),y] = (f(x)-y)^2$$

Malheureusement, le risque (2.1) ne peut être directement minimisé dans la mesure où la distribution de probabilité sous-jacente P(x, y) est inconnue. Aussi, on va chercher une fonction de décision proche de celle optimale à partir de dont on dispose, c'est-à-dire l'ensemble d'apprentissage et la classe de fonctions F est à laquelle la solution f appartient. Pour ce faire, on approxime le minimum du risque théorique par le minimum du risque empirique qui s'écrit :

$$R_{emp}[f]=\frac{1}{Nx}\sum_{i=1}^{Nx} L\,[f(x_i),y_i]\ . \qquad (2.2)$$

Il est possible de donner des conditions au classifieur pour qu'asymptotiquement (si $Nx\rightarrow \infty$), le risque empirique (2.2) converge vers le risque (2.1). Cependant, si on dispose de peu d'exemples pour faire l'apprentissage (i.e Nx petit), on s'expose au risque de sur-apprentissage (Figure 2.1). Pour éviter le sur-apprentissage, on peut restreindre la complexité de la classe F à laquelle appartient *f*. Intuitivement, une fonction de décision simple (la classe la plus simple se constituant des fonctions linéaires) capable de discriminer correctement les données est préférable à une fonction complexe. Pour cela, on introduit un terme de régularisation pour limiter la complexité des fonctions de F.

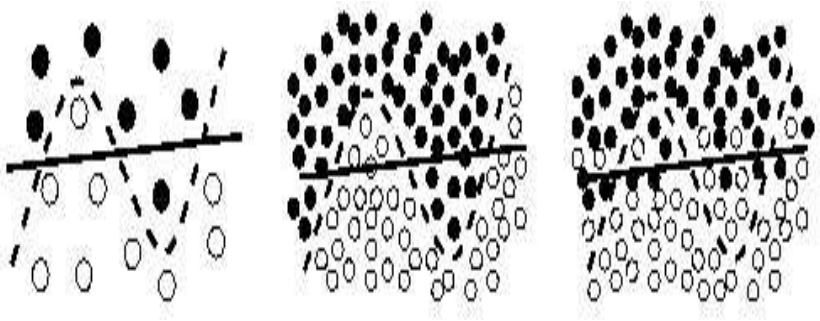


Figure 2.1 : Illustration du problème de sur apprentissage.

Etant donné un petit ensemble d'apprentissage (schéma de gauche), deux frontières de discrimination (représentées par les lignes continue et discontinue) sont possibles. La ligne discontinue est plus complexe mais minimise davantage le risque empirique. Seul un ensemble d'exemples plus grand permet de déterminer la meilleure des deux frontières de décision. S'il s'agit de la ligne discontinue, alors la ligne continue n'est pas suffisamment discriminante (schéma du milieu) ; s'il s'agit de la ligne continue, alors la ligne discontinue ne convient pas et caractérise un sur apprentissage (schéma de droite) **[12]**.

2.2. Théorie de Vapnik-Chervonenkis

Une manière de contrôler la complexité d'une classe de fonctions est donnée par la théorie de Vapnik-Chervonenkis (VC) et le principe de minimisation du risque structurel. Ici, le concept de

complexité de la fonction de décision f s'exprime par la dimension de VC (notée h) de la classe de fonctions F à laquelle appartient f . Grossièrement, la dimension de VC mesure combien d'échantillons de l'ensemble d'apprentissage peuvent être séparés par toutes les classifications possibles issues des fonctions de la classe **[8]**.

Considérons une famille imbriquée de classes de fonctions

$$F_1 \subset F_2 \subset \ldots \subset F_k ;$$

avec une dimension de VC non-décroissante, et $f_1 \ldots f_k$ les fonctions minimisant le risque empirique dans chacune de ces classes.

La minimisation du risque structurel consiste à choisir la classe F_i (et la fonction $f_{i)}$ de sorte à ce qu'une borne supérieure de l'erreur de généralisation puisse être minimisée (grâce, par exemple, au théorème suivant) **[8]**.

Théorème 1 : Soient h la dimension de VC de la classe de fonctions F, Remp[f] le risque empirique défini par (2.2) avec la fonction perte 0/1 (i.e. L [f(xi),yi] =H(-yf(x)) Où H désigne la fonction de Heaviside). Pour tout $\delta>0$ et $f \in F$, l'inégalité bornant le risque

$$R[f] = R_{emp}[f] + \sqrt{\frac{h(\ln \frac{2Nx}{h} + 1) - \ln(\frac{\delta}{4})}{Nx}} \qquad (2,3)$$

est vraie avec une probabilité de moins $(1- \delta)$pour $Nx>h$ *[12]*.

Cette borne n'est qu'un exemple et des formulations du même type ont été démontrées pour d'autres fonctions perte et d'autre mesures de complexité. Le but recherché ici est de minimiser l'erreur de généralisation R[f] en obtenant un faible risque empirique Remp [f] tout en gardant la plus petite classe de fonctions possible.

L'inégalité (2.3) fait apparaître deux cas extrêmes:

- une très petite classe de fonctions (par exemple F_1) fait décroître rapidement le terme de complexité (celui en racine carrée), mais le risque empirique demeure grand,
- une très grande classe de fonctions (par exemple F_k) implique un risque empirique petit, mais le terme de complexité explose.

La meilleure classe de fonctions est généralement intermédiaire entre la plus petite et la plus grande, puisque l'on cherche une fonction qui explique au mieux les données tout en préservant un faible risque empirique (Figure 2.2) **[12].**

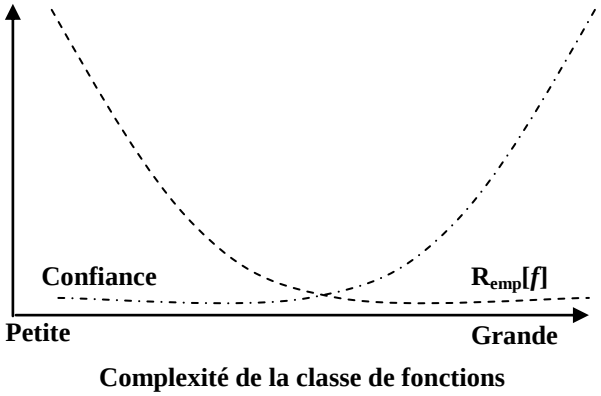


Figure 2.2 : Illustration de l'inégalité (2.3).

La courbe croissante, appelée confiance, correspond à la borne supérieure du terme de complexité. Les comportements du terme de complexité et de l'erreur empirique sont clairement opposés. On recherche donc le meilleur compromis entre complexité et erreur empirique **[12]**.

2.3. Marge et dimension de VC

Supposons pour l'instant que les échantillons de l'ensemble d'apprentissage sont séparables par un hyperplan (Figure 2.3), i.e on choisit des fonctions de décision de la forme :

$f(x)=\langle w, x\rangle +b.$ (2.4)

La marge est la distance minimale entre les échantillons de l'ensemble d'apprentissage et la frontière de décision.

Il a été montré que pour la classe des hyperplans, la dimension de VC peut être bornée en fonction de la marge. La marge peut à son tour être mesurée grâce au vecteur poids *w*: puisque nous supposons que les échantillons sont séparables, on peut redéfinir *w* et *b* de sorte à ce que les échantillons *x* les plus proches de l'hyperplan satisfassent $|\langle w, x\rangle +b|=1$.

Considérons maintenant deux échantillons x_1 et x_2 de classes différentes telles qu'on ait $\langle w, x_1\rangle +b = +1$et $\langle w, x_2\rangle +b = -1$. La marge γ correspond alors à la distance entre x_1 et x_2 mesurée perpendiculairement à l'hyperplan :

$$\gamma=\langle w / ||w||, x_1 - x_2\rangle =2/||w|| ;$$

Les résultats liant la dimension de VC de la classe des hyperplans de séparation à la marge et à la longueur du vecteur poids *w* sont respectivement donnés par les inégalités suivantes :

Où R est le rayon de la plus petite boule englobant les données. Ainsi, en bornant la marge de la classe de fonction, on peut contrôler sa dimension de VC **[8,12]**.

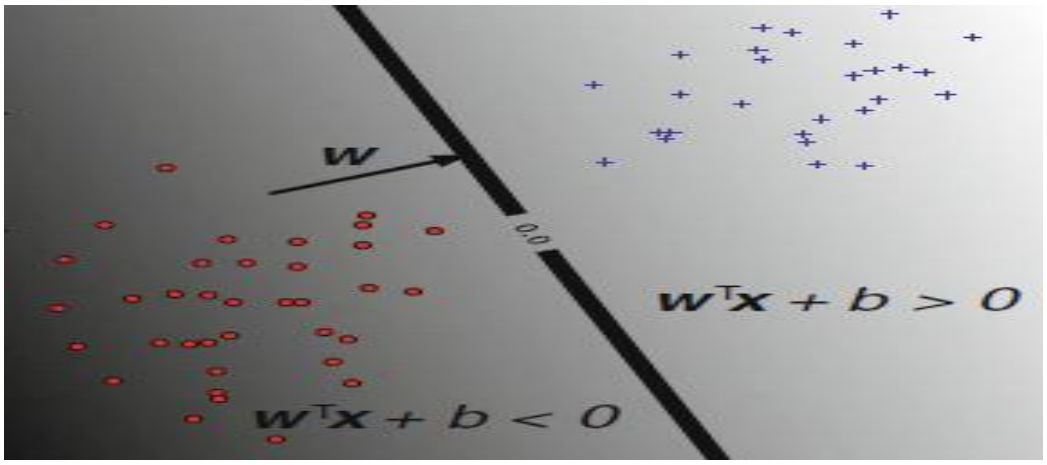


Figure2.3 : Classifieur linéaire et marge.

Un classifieur linéaire est défini par un vecteur normal à l'hyperplan w et un biais b:la frontière de décision est {x| $\langle w, x \rangle$ +b=0}(ligne continue).Chacun des deux sous-espaces séparés par l'hyperplan correspond à une classe, i.e. f(x)=signe($\langle w, x \rangle$ +b). La marge du classifieur linéaire est la distance minimale entre les échantillons de l'ensemble d'apprentissage et la frontière de décision. Sur le schéma, il s'agit de la distance entre la ligne continue et les lignes discontinues **[12]**.

3. SVM principe de fonctionnement général

3.1. Notions de base: Hyperplan, marge et support vecteur

Pour deux classes d'exemples donnés, le but de SVM est de trouver un classificateur qui va séparer les données et maximiser la distance entre ces deux classes. Avec SVM, ce classificateur est un classificateur linéaire appelé hyperplan.

Dans le schéma qui suit, on détermine un hyperplan qui sépare les deux ensembles de points **[16]**.

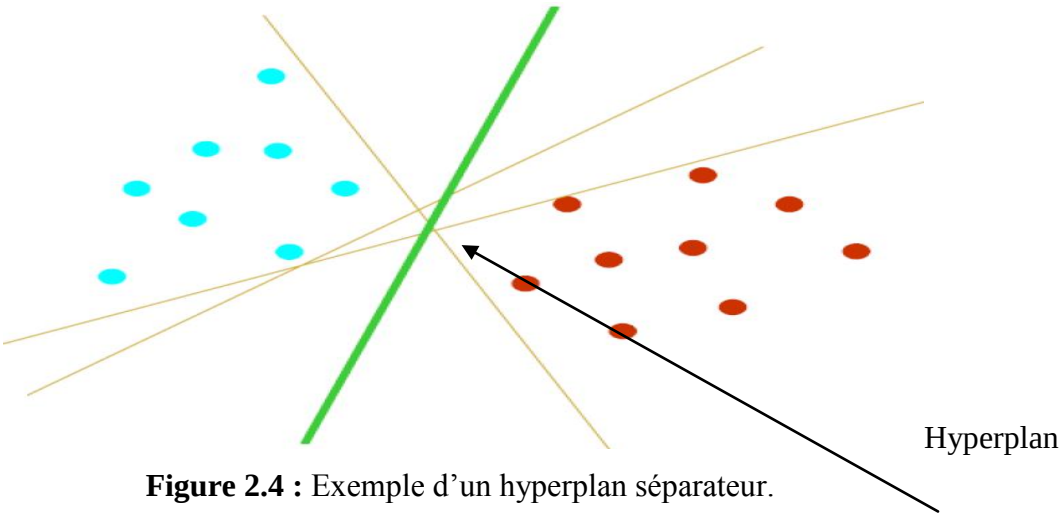


Figure 2.4 : Exemple d'un hyperplan séparateur.

Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support.

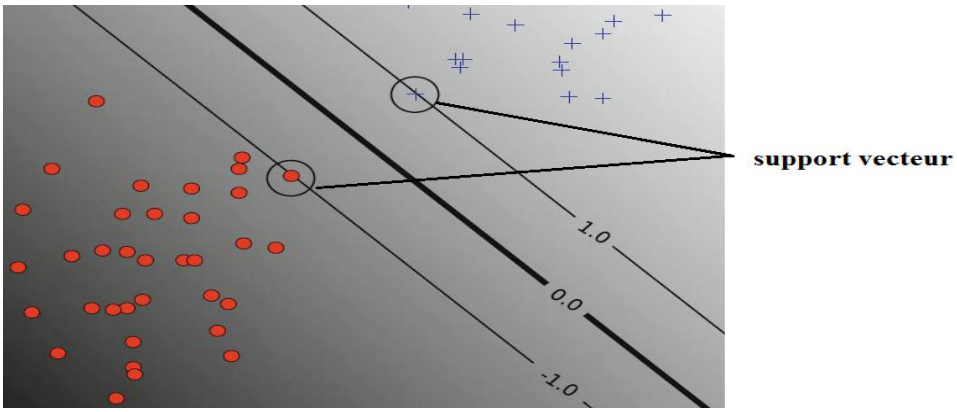


Figure 2.5 : Exemple de vecteurs de support.

Il est évident qu'il existe une multitude d'hyperplan valide mais la propriété remarquable des SVM est que cet hyperplan doit être optimal. Nous allons donc en plus chercher parmi les hyperplans valides, celui qui passe « au milieu » des points des deux classes d'exemples. Intuitivement, cela revient à chercher l'hyperplan le « plus sûr ». En effet, supposons qu'un exemple n'ait pas été décrit parfaitement, une petite variation ne modifiera pas sa classification si sa distance à l'hyperplan est grande. Formellement, cela revient à chercher un hyperplan dont la distance minimale aux exemples d'apprentissage est maximale **[11]**.

On appelle cette distance « marge » entre l'hyperplan et les exemples. L'hyperplan séparateur optimal est celui qui maximise la marge. Comme on cherche à maximiser cette marge, on parlera de séparateurs à vaste marge.

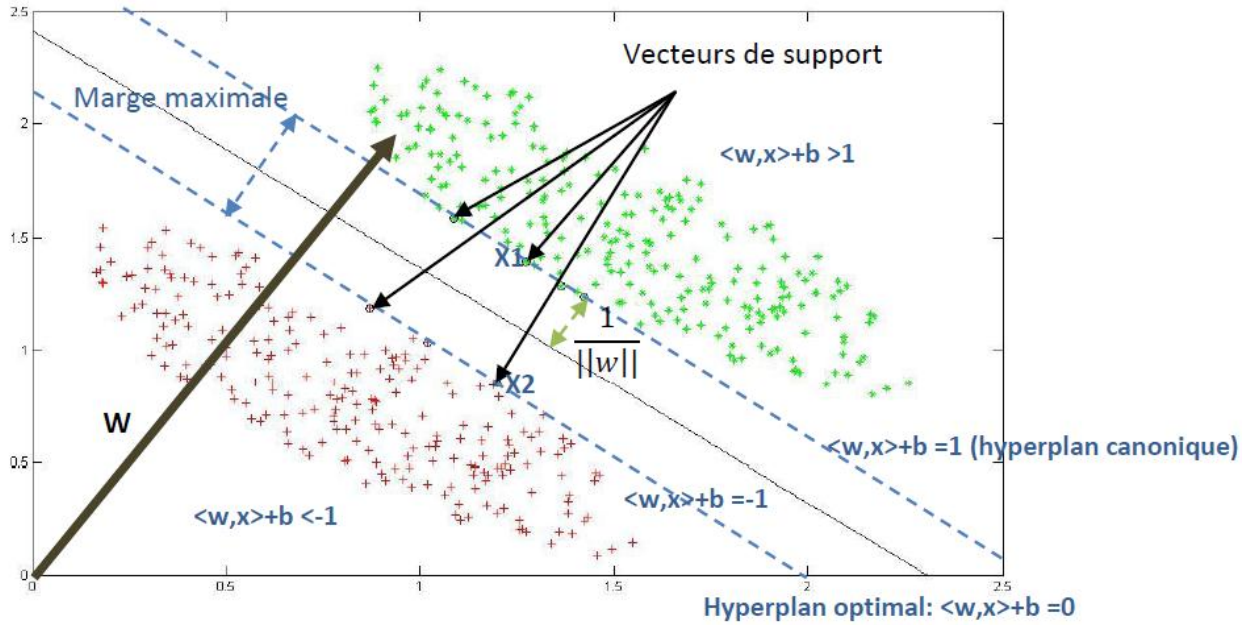


Figure 2.6 : Exemple de marge maximal (hyperplan optimal)

3.2. Pourquoi maximiser la marge ?

Intuitivement, le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple. De plus, si l'on trouve le classificateur qui se comporte le mieux vis-à-vis des données d'apprentissage, il est clair qu'il sera aussi celui qui permettra au mieux de classer les nouveaux exemples. Dans le schéma qui suit, la partie droite nous montre qu'avec un hyperplan optimal, un nouvel exemple reste bien classé alors qu'il tombe dans la marge. On constate sur la partie gauche qu'avec une plus petite marge, l'exemple se voit mal classé **[16]**.

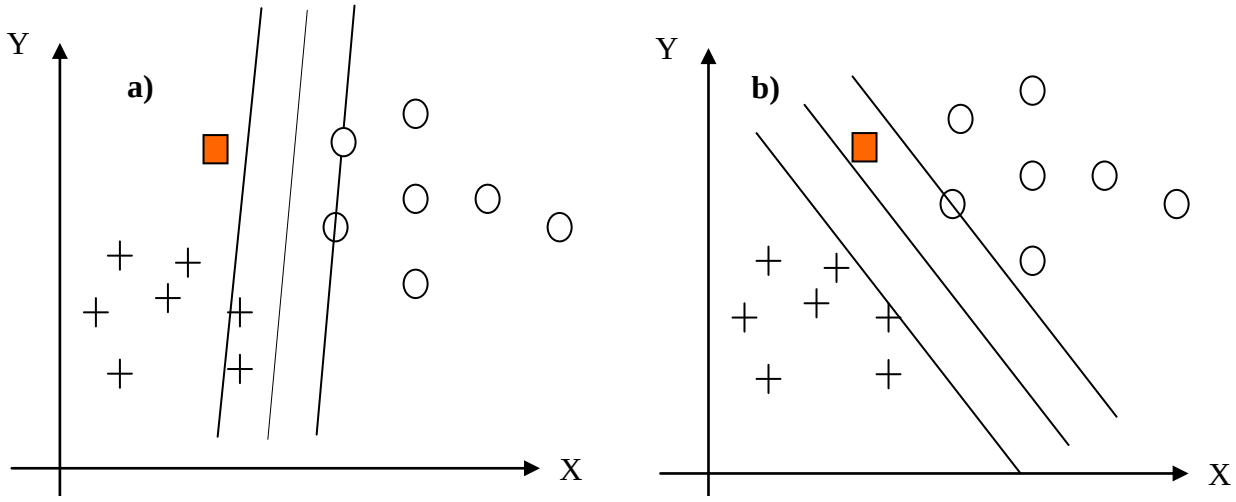


Figure 2.7 :a) Hyperplan avec faible marge **b)** Meilleur hyperplan séparateur **[16]**.

En général, la classification d'un nouvel exemple inconnu est donnée par sa position par rapport à l'hyperplan optimal. Dans le schéma suivant, le nouvel élément sera classé dans la catégorie des « + ».

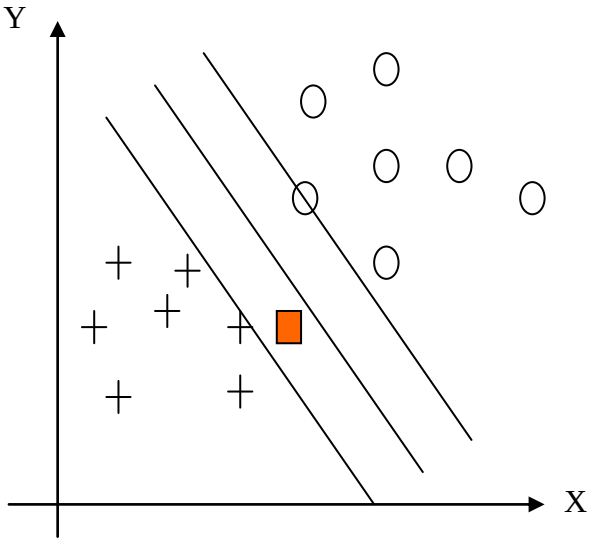


Figure 2.8 : Exemple de classification d'un nouvel élément **[16]**.

3.3. Linéarité et non-linéarité

Parmi les modèles des SVM, on constate les cas linéairement séparable et les cas non linéairement séparable. Les premiers sont les plus simples de SVM car ils permettent de trouver facilement le classificateur linéaire. Dans la plupart des problèmes réels il n'y a pas de séparation linéaire possible entre les données, le classificateur de marge maximale ne peut pas être utilisé car il fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables **[16]**.

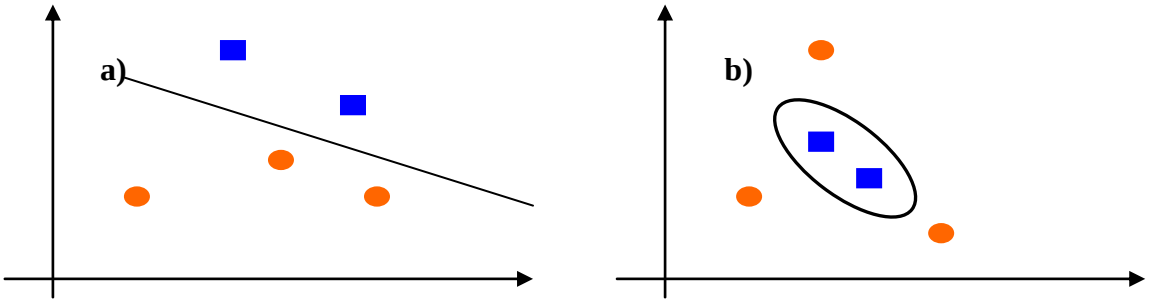


Figure 2.9 : a) Cas linéairement séparable **b)** Cas non linéairement séparable **[16]**.

3.4. Cas non linéaire

Pour surmonter les inconvénients des cas non linéairement séparable, l'idée des SVM est de changer l'espace des données. La transformation non linéaire des données peut permettre une séparation linéaire des exemples dans un nouvel espace. On va donc avoir un changement de dimension. Cette nouvelle dimension est appelé « espace de re-description ».En effet, intuitivement, plus la dimension de l'espace de re-description est grande, plus la probabilité de pouvoir trouver un hyperplan séparateur entre les exemples est élevée. Ceci est illustré par le schéma suivant :

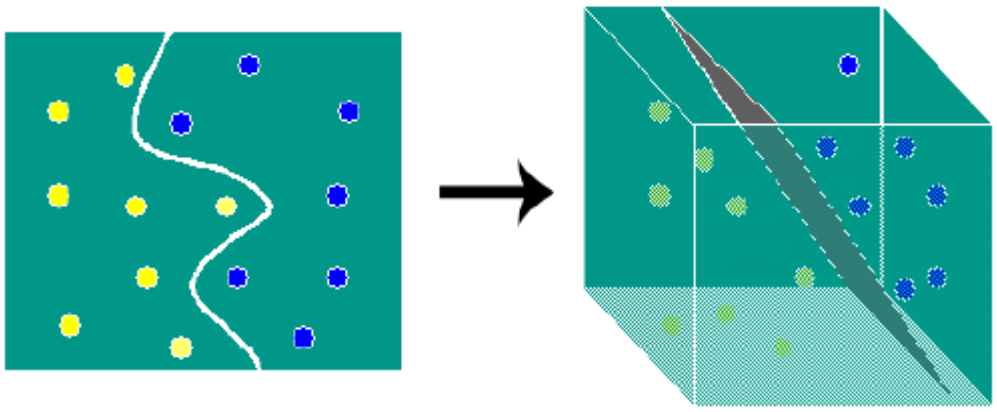


Figure 2.10 : Exemple de changement de l'espace de données **[16]**.

On a donc une transformation d'un problème de séparation non linéaire dans l'espace de représentation en un problème de séparation linéaire dans un espace de re-description de plus grande dimension. Cette transformation non linéaire est réalisée via une fonction noyau. En pratique, quelques familles de fonctions noyau paramétrables sont connues et il revient à l'utilisateur de SVM d'effectuer des tests pour déterminer celle qui convient le mieux pour son application. On peut citer les exemples de noyaux suivants : polynomiale, gaussien, sigmoïde et laplacien **[16]**.

4. Fondements mathématiques

Nous allons détailler dans les paragraphes ci-dessous les principes mathématiques sur lesquels repose SVM.

4.1. Problème d'apprentissage

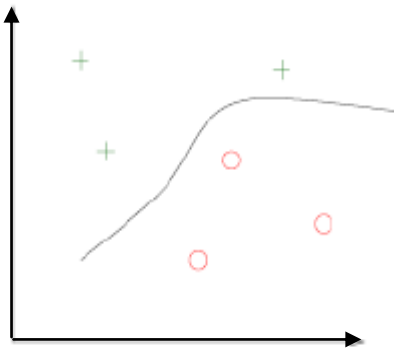
On s'intéresse à un phénomène f (éventuellement non déterministe) qui, à partir d'un certain jeu d'entrées x, produit une sortie y = f(x).

Le but est de retrouver cette fonction f à partir de la seule observation d'un certain nombre de couples entrée-sortie {(xi; yi) : i = 1,…, n} afin de « prédire » d'autres évènements.

On considère un couple (X, Y) de variables aléatoires à valeurs dans X x Y.
Seul le cas : Y = {-1, 1} (classification) nous intéresse ici (on peut facilement étendre au cas : card(Y) = m > 2 et au cas $Y = \Re$. La distribution jointe de (X, Y) est inconnue.

Sachant qu'on observe un échantillon S = {(X1, Y1),..., (Xn, Yn)} de n copies indépendantes de (X, Y), on veut: construire une fonction h : $X \rightarrow Y$ telle que P (h(X) != Y) soit minimale **[14]**.

Illustration :
Trouver une frontière de décision qui sépare l'espace en deux régions (pas forcément connexes).

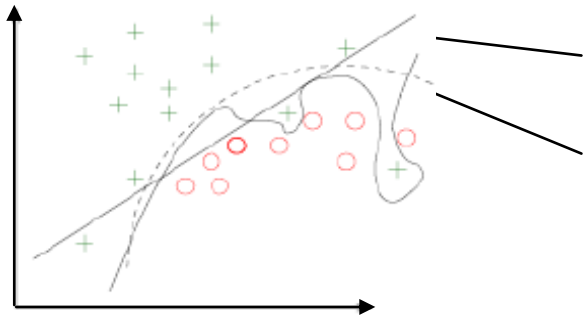


Connaissant h, on peut en déduire la classification des nouveaux points c'est à dire trouver une frontière de décision.

Le problème est de trouver une frontière assez éloignée des points de différentes classes. C'est ce qui constituera l'un des problèmes majeurs de classification grâce aux SVMs [14].

Figure 2.11 : Illustration du problème détermination de frontière assez éloignée des points de différentes classes **[14]**.

Sur et sous- apprentissage :



Si les données sont générées par un modèle quadratique :
Le modèle linéaire est en situation de sousapprentissage
Le modèle de haut degré est en situation de surapprentissage (apprentissage par coeur)

Il faut donc trouver un compromis entre adéquation aux données et complexité pour pouvoir généraliser.

Figure 2.12 : Illustration des sous et sur apprentissage **[14]**.

4.2. Classification à valeurs réelles

Plutôt que de construire directement h : $\boldsymbol{X} \rightarrow$ {-1, 1}, on construit :
f : $X \rightarrow R$ (ensemble des réels).La classe est donnée par le signe de f ;
h = signe (f).

L'erreur se calcule avec P (h(X) != Y) = P (Yf(X) ≤0). Ceci donne une certaine idée de la confiance dans la classification. Idéalement, |Yf(X)| est proportionnel à P (Y|X).Yf(X) représente la marge de f en (X, Y). Le but à atteindre est la construction de f et donc h. Nous allons voir comment y parvenir **[16]**.

4.2.1. Transformation des entrées

Il est peut être nécessaire de transformer les entrées dans le but de les traiter plus facilement. X est un espace quelconque d'objets. On transforme les entrées en vecteurs dans un espace F (feature space) par une fonction: $\Phi : X \rightarrow F$; $\boldsymbol{F}$ n'est pas nécessairement de dimension finie mais dispose d'un produit scalaire (espace de Hilbert). L'espace de Hilbert est une généralisation de l'espace euclidien qui peut avoir un nombre infini de dimensions. La non linéarité est traitée dans cette transformation, on peut donc choisir une séparation linéaire (on verra plus loin comment on arrive à ramener un problème non linéaire en un problème linéaire classique) **[16]**.

Dès lors, il s'agit de choisir l'hyperplan optimal qui classifie correctement les données (Lorsque c'est possible) et qui se trouve le plus loin possible de tous les points à classer.

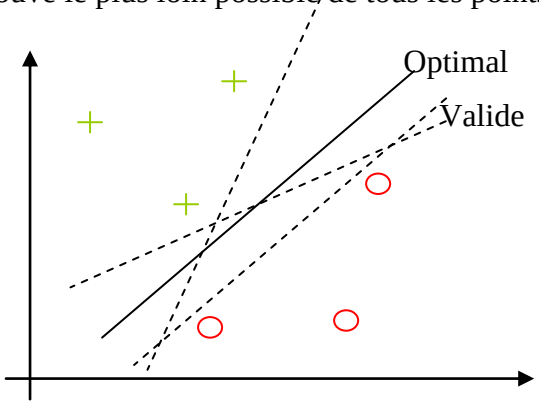


Figure 2.13 : Exemple de recherche d'un hyperplan optimal **[14]**.

Mais l'hyperplan séparateur choisi devra avoir une marge maximale.

4.2.2. Maximisation de la marge

La marge est la distance du point le plus proche à l'hyperplan.

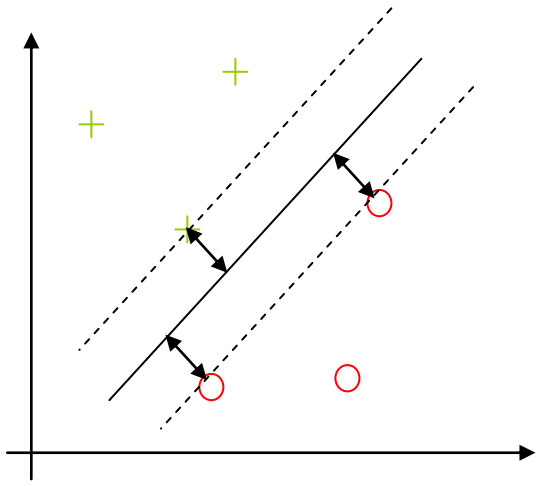


Figure 2.14 : Illustration de la relation entre marge, points de vecteurs de support et hyperplan optimal **[9]**.

4.3. Temps de calcul et convergence

4.3.1. Complexité

Nous allons évaluer la complexité (temps de calcul) de l'algorithme SVM. Elle ne dépend que du nombre des entrées à classer (d) et du nombre de données d'apprentissage (n).

On montre que cette complexité est polynomiale en n.

$$dn^2 \leq \text{Complexité} \leq dn^3$$

Taille de la matrice hessienne = n²
En effet, on doit au moins parcourir tous les éléments de la matrice ainsi que toutes les entrées. Pour un très grand nombre de données d'apprentissage, le temps de calcul explose. C'est pourquoi les SVMs sont pratiques pour des « petits » problèmes de classification **[16]**.

4.3.2. Pourquoi SVM marche?

Les noyaux précédents qui sont les plus utilisés, remplissent les conditions de mercer (facile à vérifier une fois qu'on a le noyau).

Normalement, la classe (le nombre) des hyperplans de R^d est de dH = d + 1. Mais la classe des hyperplans de marge 1/||w|| tels que ||w||² ≤ c est bornée par : dH ≤ Min (R² c, d) + 1 Où R est le rayon de la plus petite sphère englobant l'échantillon d'apprentissage S Donc dH peut être beaucoup plus petit que la dimension d de l'espace d'entrée X ; il est donc toujours possible d'en trouver un c'est la raison pour laquelle **[11]**.

5. SVMs et analyse des bases de données

5.1. Introduction

Les machines à vecteur support présentées au chapitre précédent sont des outils très puissants pour plusieurs tâches d'analyse des bases de données, et qui peuvent faire face à des problèmes difficiles à résoudre par les méthodes classiques d'analyse statistiques et de classification.

Les SVMs viennent même d'être intégrées dans des systèmes de gestion de bases de données tels qu'Oracle, qui présente à partir de sa version 10 g une intégration complète des SVMs.

En effet, l'analyse des bases de données dans les différentes étapes du processus du data mining peut profiter de la robustesse des SVMs pour améliorer ses performances.

5.2. Entrepôt de données

Les bases de données analysées sont généralement rassemblées dans des entrepôts de données Data warehouse. Un entrepôt de données est un environnement structuré conçu pour stocker et analyser toutes les parties significatives d'un ensemble de données [21].

Les données sont physiquement et logiquement transformées de plusieurs applications sources dans une structure commerciale maintenue est la mise à jour pour une longue période. Un entrepôt de données est généralement organisé autour d'un sujet majeur dans une entreprise tel que le client, le vendeur, le produit ou l'activité, ce qui affecte directement la conception et l'implémentation des données dans l'entrepôt de données. Les données de l'application source qui ne sont pas utilisées dans l'analyse pour atteindre l'objectif sont exclues de l'entrepôt de données. La figure 2.16 présente une architecture typique d'un entrepôt de données.

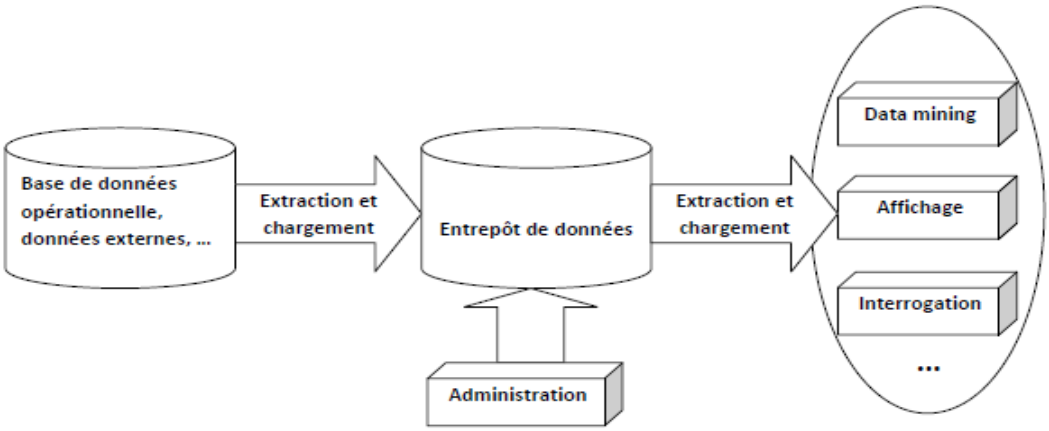


Figure 2.15 : Architecture d'un entrepôt de données.

L'analyse des bases de données se fait généralement pour la découverte des informations qui se cachent dans les grandes quantités de données. Cependant la manipulation des bases de données peut être confrontée dans plusieurs phases du chemin vers la découverte des informations.

En effet, les bases de données peuvent être elles-mêmes des sources naturelles de données telles que dans le cas des systèmes d'information dans une banque ou un supermarché.

Les bases de données peuvent être construites après la phase d'extraction des caractéristiques d'un autre type de données. Après leur prétraitement, les données extraites, sont enregistrés dans un entrepôt de données sous forme de bases de données.

Après la phase d'extraction des connaissances, les informations extraites peuvent être enregistrée sous forme de bases de données.

Les étapes dans lesquelles les bases de données nécessitent d'être analysées sont les étapes de d'acquisition et d'extraction des connaissances [21].

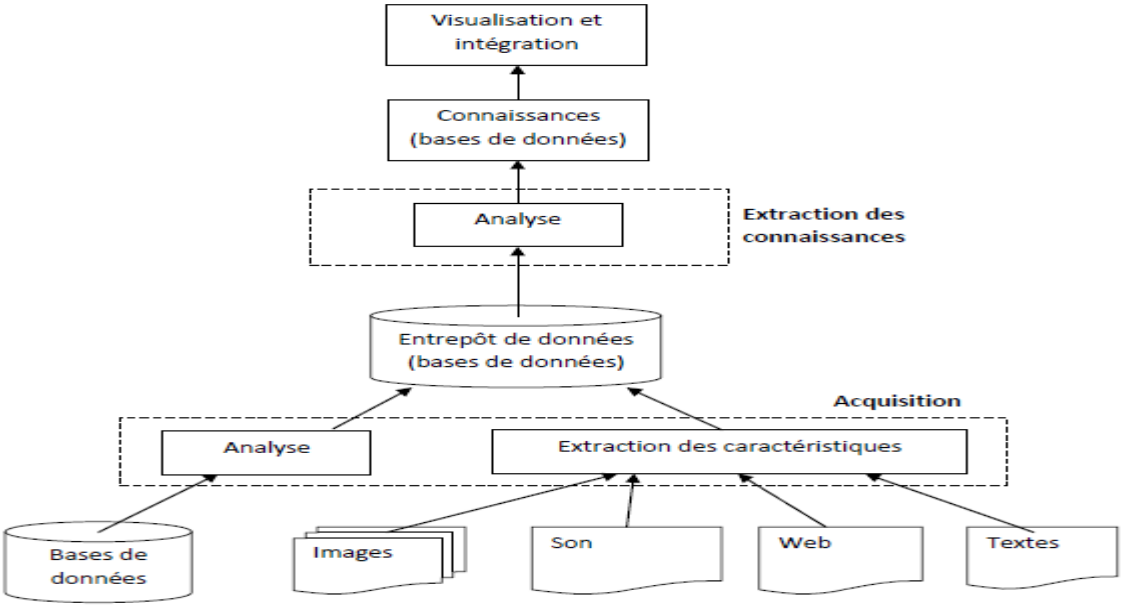


Figure 2.16 : Analyse des BDDs dans le processus de data mining.

6. Les domaines d'applications

SVM est une méthode de classification qui montre de bonnes performances dans la résolution de problèmes variés. Cette méthode a montré son efficacité dans de nombreux domaines d'applications tels que le traitement d'image, la catégorisation de textes ou le diagnostiques médicales et ce même sur des ensembles de données de très grandes dimensions La réalisation d'un programme d'apprentissage par SVM se ramène à résoudre un problème d'optimisation impliquant un système de résolution dans un espace de dimension conséquente. L'utilisation de ces programmes revient surtout à sélectionner une bonne famille de fonctions noyau et à régler les paramètres de ces fonctions. Ces choix sont le plus souvent faits par une technique de validation croisée, dans laquelle on estime la performance du système en la mesurant sur des exemples n'ayant pas été utilisés en cours d'apprentissage. L'idée est de chercher les paramètres permettant d'obtenir la performance maximale. Si la mise en oeuvre d'un algorithme de SVM est en général peu coûteuse en temps, il faut cependant compter que la recherche des meilleurs paramètres peut requérir des phases de test assez longues **[16]**.

7. Avantages et inconvénients

Avantage:

- Absence d'optimum local.
- contrôle explicite du compromis entre la complexité du classifier et l'erreur.
- Possibilité d'utilisation de structure de données comme les chaines de caractères et arbres comme des entrées.
- traitement des données a grandes dimensions.

Inconvénients :

- Demande des données négatives & positives en même temps.
- Besoin d'une bonne fonction Kernel.
- Problèmes de stabilité des calculs dans la résolution de certain programme quadratique a contraintes.

Conclusion

Dans ce chapitre, nous avons tenté de présenter de manière simple et complète le concept de système d'apprentissage introduit par Vladimir Vapnik, les « Support Vecteur Machine » Nous avons donné une vision générale et une vision purement mathématiques des SVM. Cette méthode de classification est basée sur la recherche d'un hyperplan qui permet de séparer au mieux des ensembles de données. Nous avons exposé les cas linéairement séparable et les cas non linéairement séparables qui nécessitent l'utilisation de fonction noyau (kernel) pour changer d'espace. Cette méthode est applicable pour des tâches de classification à deux classes, mais il existe des extensions pour la classification multi classe.

Nous sommes ensuite intéressés aux différents domaines d'application. Il existe des extensions que nous n'avons pas présentées, parmi lesquelles l'utilisation des SVM pour des tâches de régression, c'est-à-dire de prédiction d'une variable continue en fonction d'autres variables, comme c'est le cas par exemple dans la prédiction de consommation électrique en fonction de la période de l'année, de la température, etc. Le champ d'application des SVM est donc large et représente une méthode de classification intéressante.

Chapitre 3

Conception et implémentation du système

- **Introduction**
- **Différents étapes du système**
- **Conclusion**

Introduction

Dans le chapitre précédent nous avons présentés la méthode de classification binaire SVM, inspirées de la théorie statistique de l'apprentissage de Vladimir Vapnik introduite en 1995. Dans ce chapitre nous allons présentés une conception par affinement successif du système en donnant son architecture générale, puis nous détaillons en étudiant séparément chacun de ses composants.

1. Différentes étapes du système

L'objectif de notre système est la réalisation d'un système vocale (saisir les note d'étudiants à partir les nombres d'inscrit) avec ce nom (XLManager) pour ce faire, on utilise un ensemble de commandes vocales où chaque commande passe par une succession d'opérations : acquisition, prétraitement, segmentation et extraction des vecteurs acoustiques, apprentissage et classification, post-traitement et finalement modifier le fichier Excel xls.

Le système peut être vus ou décomposé en modules (composants):

- Acquisition.
- Prétraitement.
- Segmentation parole/ silence (méthode structurelle).
- Extraction des caractéristiques.
- Apprentissage et classification.
- Post-traitement.

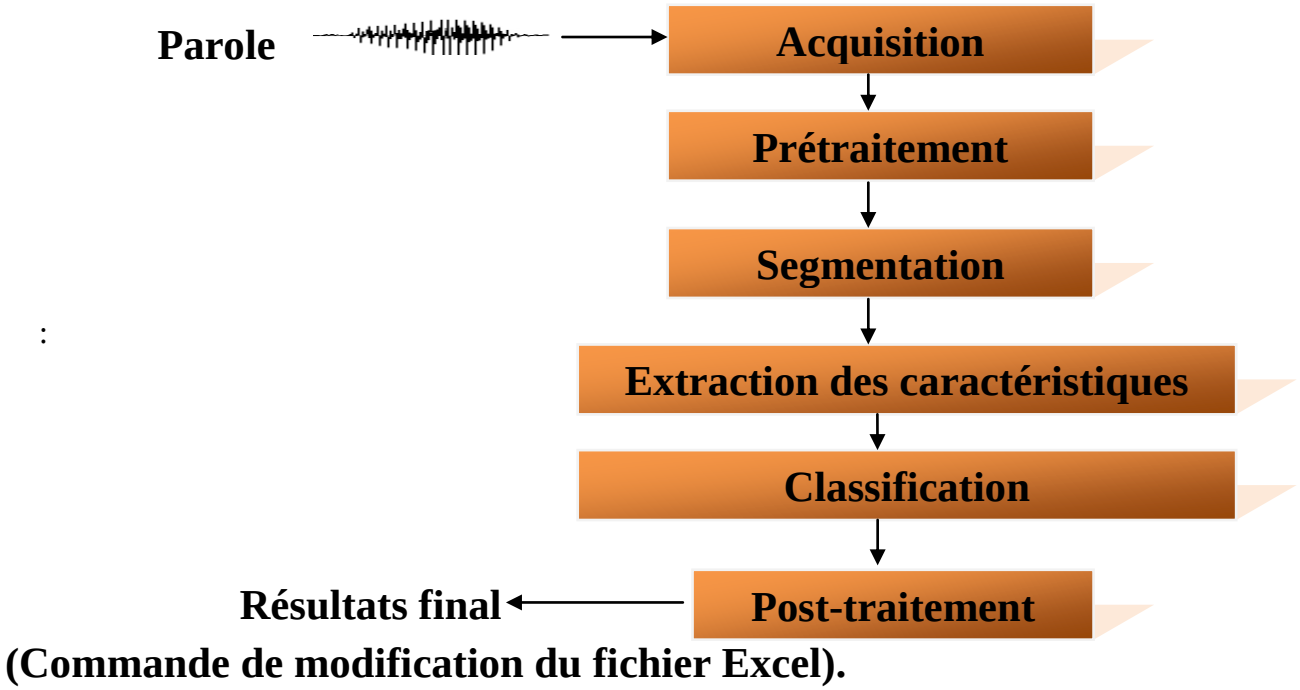


Figure 3.1 : Les différents composants du système.

2. Description des étapes

2.1. Acquisition

D'un sens, acquérir quelque chose c'est devenir propriétaire d'un bien. De ce sens l'acquisition du signal de parole (information) revient à l'appropriation des informations à un micro-ordinateur afin d'exécuter une tâche précise. L'acquisition est la première étape du processus de reconnaissance. Dans notre système nous allons utiliser le microphone comme outil d'acquisition à l'extérieur du PC ainsi que la carte son comme périphérique interne. Après la phase d'acquisition des voix des locuteurs, ces dernières seront automatiquement numérisées sous forme de tampon **[3]**.

2.1.1. Le capteur (microphone)

Le capteur représente le premier élément de l'acquisition. Il est considéré comme un transducteur, dispositif transformant une grandeur physique en une autre grandeur dépendante de la première. Bien qu'un microphone soit obstacle à la propagation des ondes sonores, pour l'acquisition du signal de parole, ce microphone est un capteur comportant un organe sensible aux variations de pression dues à l'onde sonore **[3]**.

Ces variations de pression sont utilisées pour exercer une force sur un système ne pouvant pratiquement pas se déplacer sans cette condition (existence de la force).Il existe plusieurs types de microphone (Microphone : à charbon, à condensateur, à magnétostriction, électrodynamique, électronique, thermique, ionique).On prend le microphone à condensateur comme exemple .Ce dernier se trouve dans un circuit comprenant une résistance et un générateur. L'intensité du courant dans le circuit dépend de ces variations .Ce genre de microphone est le plus performant parmi les microphones disponibles, en plus son avantage majeur est sa petite taille ainsi que sa simple construction **[3]**.

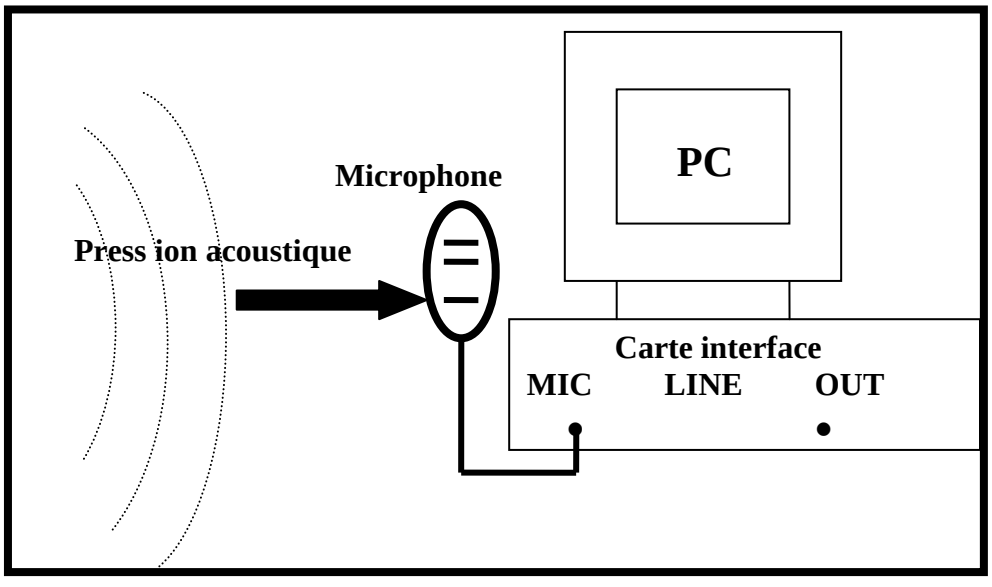


Figure 3.2: schéma synoptique de l'acquisition d'un signal de parole.

2.1.2. Carte interface (carte son)

Une fois le signal analogique, issu du microphone arrive à l'entrée MIC de la carte son, il doit passer par un circuit de conditionnement, qui permet l'amplification et le filtrage de ce signal, après quoi la conversion Analogique-Numérique est effectuée, dans le but de rendre l'information récupérée, traitable par le système numérique (micro-ordinateur).Cette conversion comprend l'échantillonnage, la quantification et le codage **[3]**.

Après la conversion Analogique-Numérique, la carte son passe à la mémorisation des données numérisées dans un espace mémoire ou tampon (buffer) sous forme de valeurs numérique. Ces donnés seront présentés par des vecteurs comportant une série de chiffre .On utilise ce genre de mémorisation plusieurs fois pour un même mot prononcé selon le choix de la taille du dictionnaire voulu, attribuée à l'apprentissage des données **[3].**

Il est à remarquer que la phase de près-traitement n'est pas incluse dans notre système parce que cette tache est gérée par le module d'acquisition du langage choisit et la carte son utilisée.

De ce fait nous allons passer directement à la phase suivante (phase de segmentation).

2.2. Segmentation

Dans ce composant, nous allons faire une analyse temporelle du signal. Une inspection minutieuse de la structure temporelle (forme d'onde), selon un certain nombre de critères, permet une segmentation primaire fiable et précise du signal en deux catégories : parole et silence.

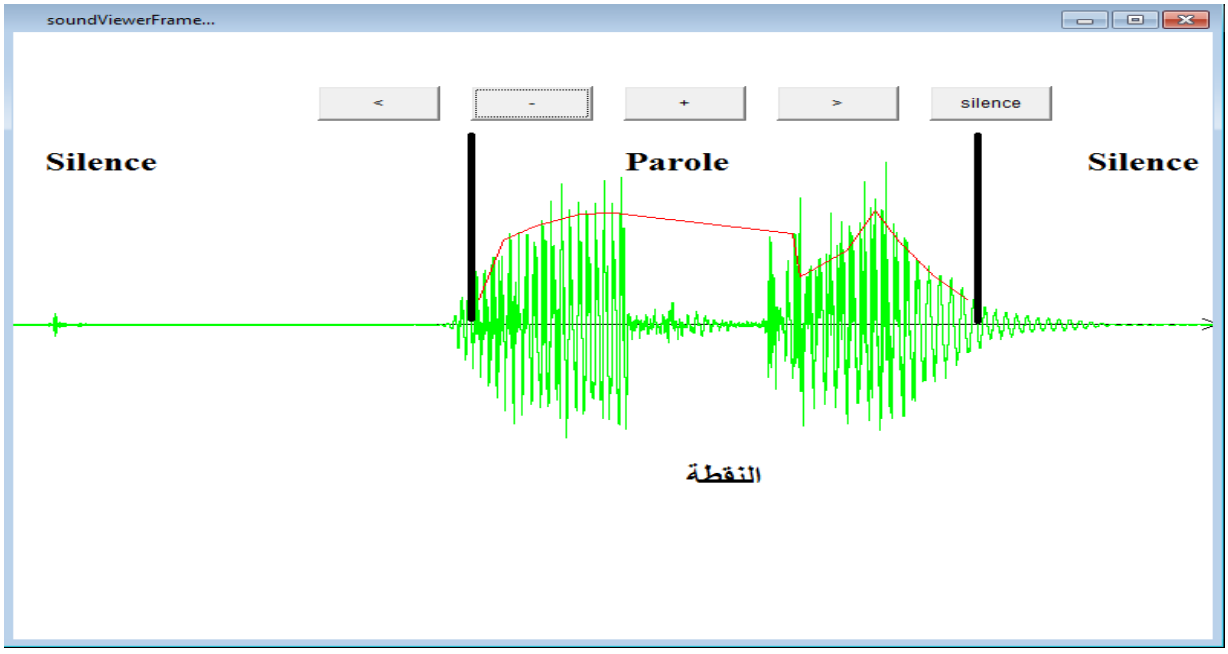


Figure 3.3 : segmentation d'un signale parole avec le mot **"النقطة"** .

Le signal est recodé premièrement selon le passage par zéro de sa dérivée. On obtient ainsi les extremums du signal à partir desquels les silences et les paroles peuvent être détectés. Un silence est une succession de valeurs inférieurs au niveau moyen du silence (par exemple 15) tandis qu'une parole est une succession de piques supérieures à la valeur moyenne du silence. Si une parole est détectée.

Pour segmenter un signale parole nous allons étudier le signale dans chaque petit intervalle et vérifier les valeurs maximales est supérieure à la valeur de silence ou non, si oui alors vérifier les autres valeurs dans le même intervalle, si oui alors partie parole, si non partie silence.

La fonction rouge dans la figure (3.3) est représenter l'intervalle de la signale parole alors le début de cette fonction est le début de mot et la fin est la fin de cette mot.

L'algorithme de segmentation est le suivant :

```
Début
    Signale : tableau de valeurs réel.
    Silence : entier. (donnée).
    Parole  : tableau de valeurs entier.
    Psilence : booléen.

    Pour  i=1  jusqu'à  N  faire
        Si (signal[i] >silence) alors
            Si  (Psilence) alors
                Psilence=false ;
                Parole [j]=i ; //l'indice de début partie parole dans le signal.
                j=j+1 ;
            Fin Si
        Si non
            Si  (non Psilence) alors
                Psilence=true
                Parole [j] =i; // l'indice de fin partie parole dans le signal.
                j=j+1 ;
            Fin Si
        Fin Si
    Fin pour
Fin
```

Généralement, la plupart des mots arabe contiennent un segment de parole qui se répète dans la partie de parole segmentée par la première étape (comme exemples voir les figures ci-dessous pour les mots arabe **"النقطة"** et **"ثلاثة"**).

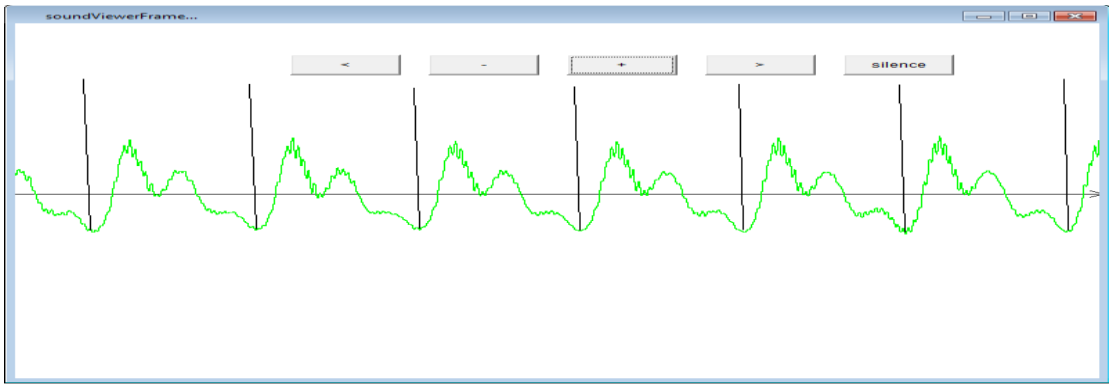


Figure 3.4 : Exemple de répétition du signale parole avec le mot **"النقطة"**.

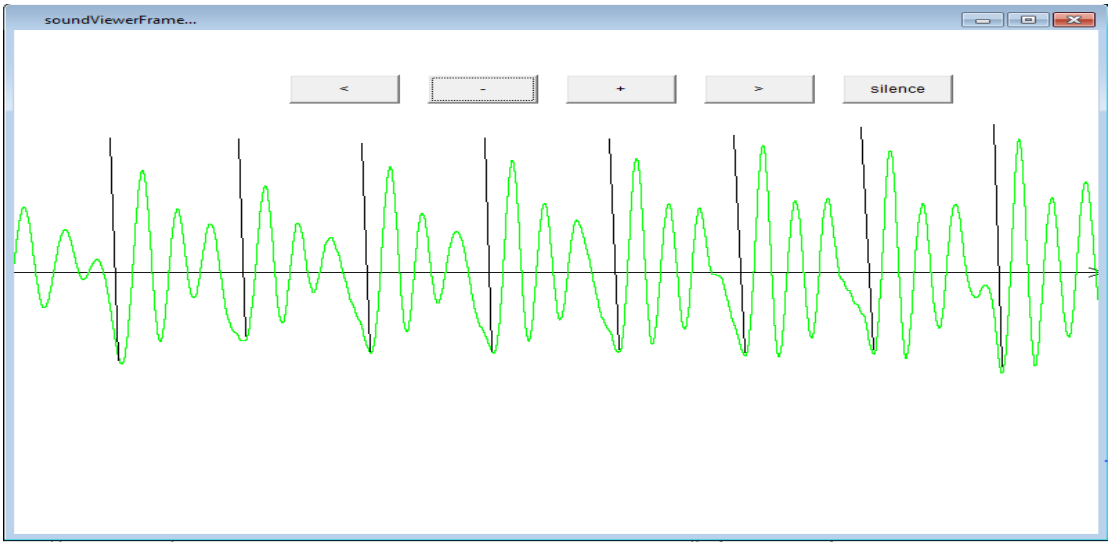


Figure 3.5 : Exemple de répétition du signale parole avec le mot **"ثلاثة"**.

2.3. Extraction des caractéristiques

L'extraction de caractéristiques joue un rôle très important dans les systèmes de reconnaissance la parole ou de la langue. Il existe une diversité de méthodes pour extraire les caractéristiques d'un signal vocal, mais celle dont la fiabilité a été prouvée est bien le codage prédictif linéaire ou LPC (linear predictif coding) car elle extrait l'information d'une petite partie de l'enveloppe spectrale de la parole **[1,19]**.

LPC prend en entrée une parole et fournit des coefficients correspondants à ses caractéristiques statistiques les plus importantes.

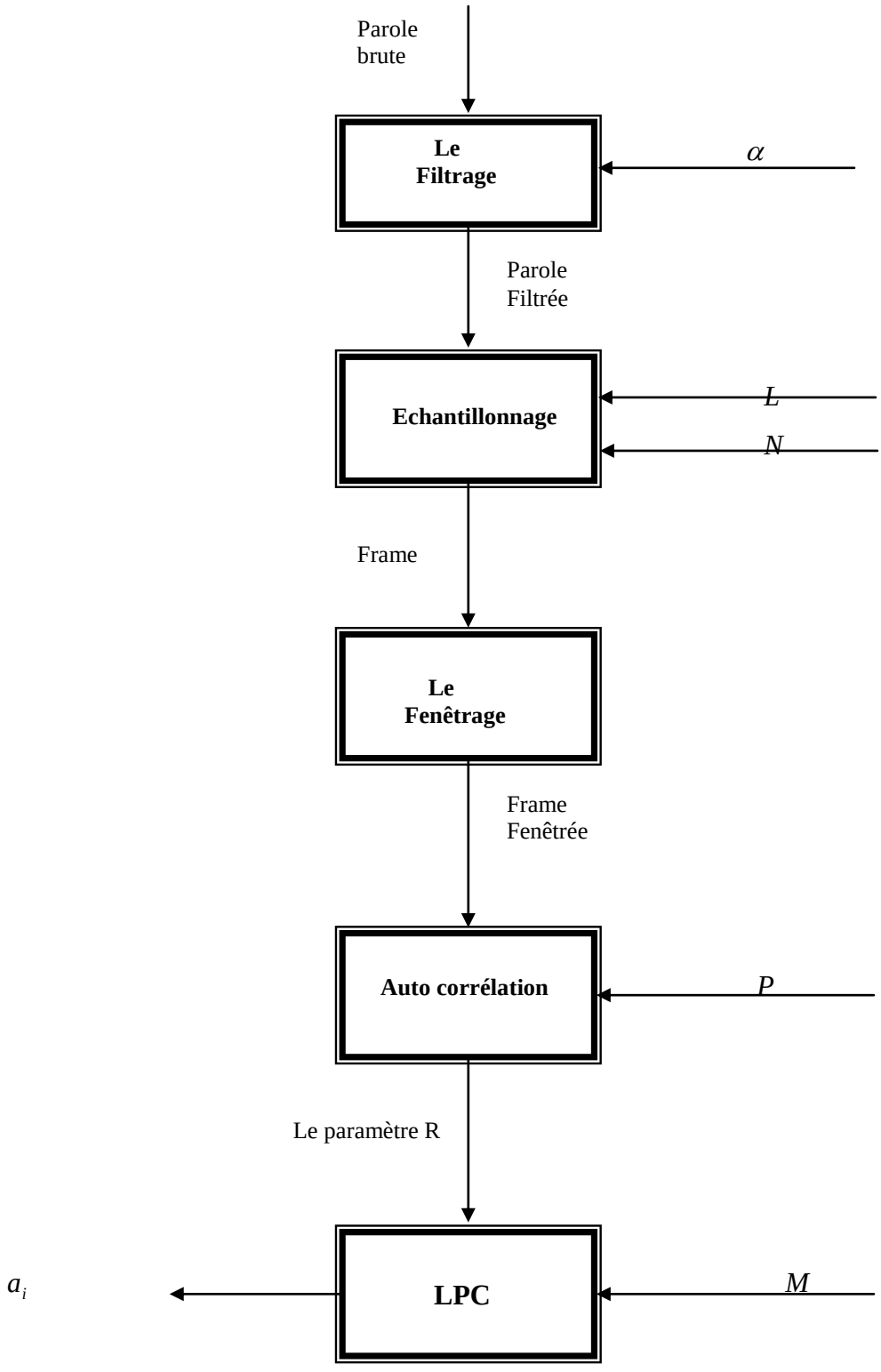


Figure 3.6: L'extraction des paramètres vocaux par LPC

Les coefficients de LPC représentent les caractéristiques les plus importantes**.** Les études faites par Atal montrent l'efficacité des coefficients de LPC dans la reconnaissance et l'identification.

2.3.1. Le principe de la prédiction linéaire

L'analyse par prédiction linéaire est une méthode de modélisation de type essentiellement temporel, bien que son principe repose sur l'hypothèse, selon laquelle un échantillon peut être prédictif comme une combinaison linéaire de N échantillons, qui le précédent d'où le nom de prédiction linéaire, abrégée LPC qui est bien adaptée à l'analyse du signal de parole **[1,19]**.

Cette méthode est fondée sur l'idée qu'un signal qui véhicule un message n'est jamais complètement aléatoire .Il y a une corrélation entre les échantillons successifs. Le codage par LPC utilise cette corrélation pour réduire les données manipulées tout en préservant l'information contenue dans le signal**.** Elle calcule des coefficients sur un échantillon de parole en tenant compte des échantillons précédents **[1,19]**.

$$\hat{y}(n) = -\sum_{i=1}^{P} a(i)y(n-i).$$

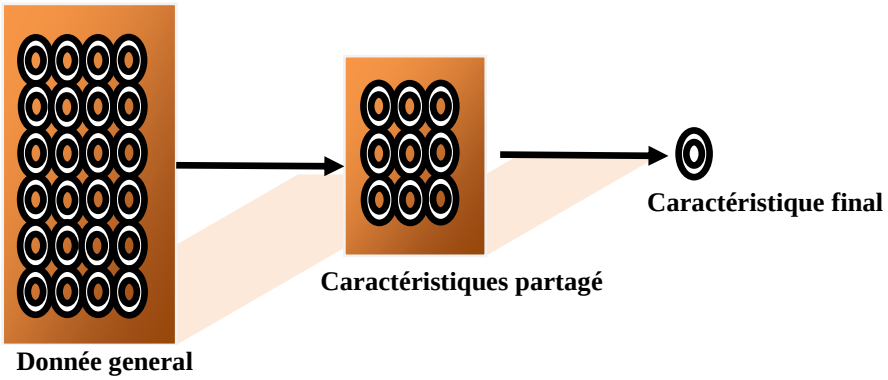


Figure 3.7 : Methode d'extraction de caractiristique.

2.3.2. Avantages de la méthode LPC

On cite les avantages les plus importants qui sont :

- Elle fournit un bon modèle du signal de parole .Cela est spécialement vrai pour la quasi-totalité des régions voisées du signal. Dans lesquelles un de ses modèles (tout pôles) fournit une représentation compacte et précise de l'enveloppe spectrale vocale.

- C'est une méthode d'analyse directe car une fois les critères de l'approximation du signal réel sont fixés, le résultat est obtenu par la résolution d'un système d'équations linéaires.

- C'est un modèle analytique traitable. Il est mathématiquement précis et simple à implémenter soit d'une manière logicielle soit matérielle.

- Le calcul demandé en LPC est considérablement inférieur à celui demandé dans l'implémentation d'autres méthodes comme la méthode de banc de filtre.

- Elle a prouvé son efficacité dans toutes les applications de reconnaissance.

- Les expériences ont montré que les performances des systèmes de reconnaissance basées sur la méthode LPC sont meilleures à celle des systèmes basés sur le banc de filtres**.**

2.3.3. Les phases de LPC

Il existe 5 phases d'extraction de caractéristiques du signal vocal par la méthode LPC :
Le filtrage, l'échantillonnage, le fenêtrage, l'auto corrélation, et le calcul des coefficients. Ces phases ont pour rôle de convertir le signal vocal en coefficients **[1,19].**

a) Le filtrage

Le signal vocal numérique s(n) est capturé par un convertisseur analogique numérique avec une fréquence d'échantillonnage f_s.Le signal est ensuite filtré par un filtre appelé FIR qui a la forme : $H(z)=1-\alpha z^{-1,}$ Où α est dans l'intervalle de 0.9 à 1.0 et reflète le degré du filtrage. La (figure 3.8) montre la réaction fréquentielle du filtrage avec $\alpha=0.95$. Quand $\omega=\pi$ la réaction du filtre est à 32 dB, c'est plus élevé que $\omega=0$.

Le concept de réaction du filtre est sans importance puisque ça n'a pas d'effet sur la perception de la parole. La sortie du filtre peut être liée à l'entrée par l'équation différentielle :

š (n)=s(n)-s(n-1) n=0,1,2,3,…….N-1.

Le filtrage doit généralement être appliqué sur les sons voisés. Quoi que ce soit c'est toujours possible d'appliquer le filtrage en utilisant un échantillon dont la valeur est dépendante de α.
$H(z)=1-\alpha(n)\, z^{-1}$
Où $\alpha(n)$ est une fonction concernant le nombre d'échantillons. Dans ce cas $\alpha(n)$ est dépendant, dans les coefficients des deux premières valeurs d'auto corrélation de l'échantillon courant **[1,19]**.

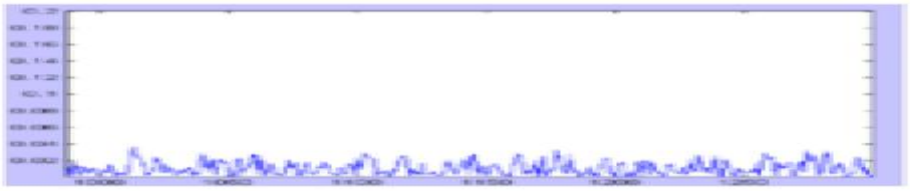


Figure 3.8: LPC de la lettre A.

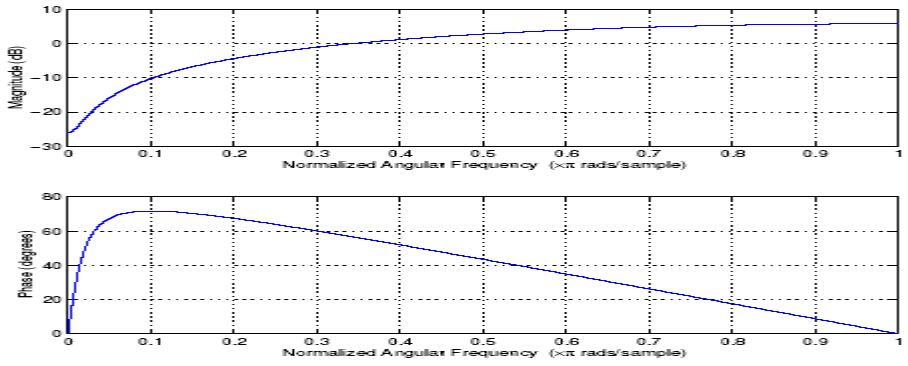


Figure 3.9 : La réaction fréquentielle du filtre.

L'algorithme de filtrage est le suivant :

Début

Signal : Tableau de valeur réel ;
N : réel (longueur de signal) ;
$\alpha = 0.95$: Constante réel ;
Pour i=1 **jusqu'à** N **faire**

Signal[i] =Signal[i]- $\alpha \times$ Signal [i-1];

FinPour

Fin

b) L'échantillonnage

Une fois le signal vocal filtré, il sera ensuite échantillonné en plusieurs échantillons équivalents de même longueur N.

Le début de chaque échantillon est à L samples du début de l'échantillon précédent. Le second échantillon commence à L, le 3ème doit commencer à 2L et ainsi de suite.

Si L<=N, les échantillons se chevauchent et l'estimation des coefficients calculés par LPC montrera un haut niveau de corrélation.

Dans un système où la fréquence d'échantillonnage est 8 KHZ., les valeurs de L et N seront respectivement 80 et 160 **[1,19]**.
Si on définit x_i comme étant le i^ème segment du signal échantillonné š et I échantillons sont requis donc l'échantillonnage sera décrit comme suit :

x_i (n)=š (L_i+ n). n=0, 1,2,……..N-1.

i=0, 1,2,…….I-1.

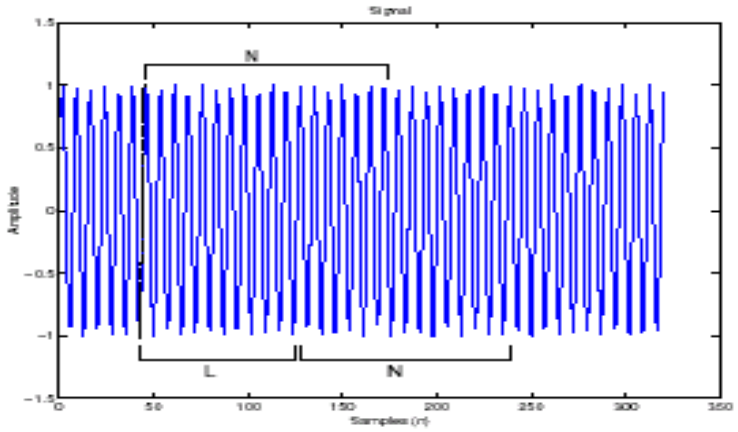


Figure 3.10 : La façon avec laquelle L et N sont utilisés dans

L'algorithme de l'échantillonnage est le suivant :

```
Début
    Signal : Tableau de valeur réel ;
    Frame : une Matrice de valeur réelle ;
    N : entier; (longueur de signal)
    L : entier ;
    I : entier ;
   Pour i=1   jusqu'à I faire
      Pour j=1   jusqu'à N faire
            Frame[i] [j] =Signal [L*i+j];
      FinPour
   FinPour
Fin
```

c) Le fenêtrage

Dans l'analyse du signal vocal, une fenêtre rectangulaire est implicitement employée. La raison d'être du concept de fenêtrage est que la fenêtre rectangulaire possède une discontinuité au début et à la fin de l'échantillon. La distorsion peut être réduite en utilisant une fonction de fenêtrage ω(n) .Il existe plusieurs fonctions de fenêtrage. Le résultat du fenêtrage des segments est défini comme suit:

$$x(n) = x_i(n)\omega(n) \qquad n=0,1,\ldots\ldots\ldots\ldots\ldots N-1.$$

L'algorithme de fenêtrage est le suivant :

```
Début
    W : Tableau de valeur réel ;
    Frame : une Matrice de valeur réelle ;
    N : entier ; // longueur de signal
    I : entier ;
   Pour i=1   jusqu'à I faire
      Pour j=1   jusqu'à N faire
            Frame[i] [j] =Frame[i] [j]*w [j];
      FinPour
   FinPour
Fin
```

d) Analyse et auto corrélation

L'analyse d'auto corrélation a pour rôle d'extraire une harmonie importante et les propriétés des formants à partir d'un signal vocal.

La fonction d'auto corrélation est définie comme suit :

$$R(i) = \frac{1}{N \sum_{n=0}^{N-i-1} x(n)x(n+i)} \quad . \qquad i=0,\ldots..p.$$

Où p est l'ordre de LPC., Ses valeurs sont inclues entre 8 et 16.
R (0) représente la mesure de l'énergie des segments de la parole.

Et peut être utilisé pour éliminer le silence. L'algorithme d'auto corrélation est le suivant :

```
Début

    Frame : Tableau de valeur réel ;
    R : Tableau de valeur réel ;
    N : entier; (longueur de Frame)
    P : entier ; (nombre de coefficients)
    Somme : réel ;

  Pour i=1   jusqu'à P faire

     Somme=0 ;
    Pour j=1   jusqu'à N-i-1faire

          Somme=somme + Frame[j]*Frame [i+j];
    FinPour

              R=Somme ;
  FinPour
Fin
```

e) Le calcul du LPC

Le prochain block fonctionnel converti le (p+1) coefficient d'auto corrélation en coefficient du LPC.

Si on suppose que l'appareil vocal est excité par un signal de bruit blanc ayant une unité de variance et sans sens, nous pouvons alors représenter ce système par un $p^{ème}$ ordre autorégressif.

$$H(z) = \frac{G_p^2}{1 + \sum_{k=1}^{p} a_{pk} Z^{-k}} .$$

Les inconnues dans cette équation :

(G_p^2 , a_{pk} Où k=1,2,……p) sont résolues en utilisant les deux équations suivantes :

$$R(0) = G_p^2 + \sum_{k=1}^{p} a_{pk} R(k).$$

Et

$$R(j) = -\sum_{k=1}^{p} a_{pk} R(j-k). \quad j=1,2,\ldots, p.$$

Ces équations sont communément référenciés comme les équations de Yule Walker.C'est possible de résoudre ces équations en utilisant une méthode récursive. La plus populaire des méthodes récursives est l'algorithme de Levinson Durlin. Cet algorithme est initialisé comme suit :

$$a_{i,i} = -\frac{R(1)}{R(0)}.$$

$$p_i = R(0)(1 - a_{1,1}^2).$$

Et récursivement implémenté pour m=2,…..,p. par :

$$a_{m,m} = -\frac{R(m) + \sum_{i=1}^{m-1} a_{m-1,i} R(m-i)}{p_{m-1}} .$$

$$a_{m,i} = a_{m-1,i} + a_{m,m}\ a_{m-1,m-i}.$$

$$p_m = p_{m-1}(1 - a_{m,m}^2).$$

La solution finale pour les coefficients LPC est donnée comme suit :

$$a_i = a_{p,i} \qquad 1 \le i \le p.$$

$$G_p^2 = p_p .$$

La dérivée de l'algorithme de Levinson est bien représentée sous forme de coefficients K_m (m=1,2,…...p) définis comme suit :

$K_i = a_{i\,i}$ $\quad 1 \le i \le p.$

Les coefficients de réflexion ou PARROR sont directement liés aux croisements de sections non uniformes du conduit vocal tout en formant ainsi un modèle de cet appareil vocal. Le système vocal peut être considéré comme une cascade de P cylindres de longueurs équivalentes **[1,19]**.

Quand l'air traverse l'appareil vocal, la différence au niveau des croisements cause une réflexion sur les frontières où les coefficients sont indiqués par k_m. Ces coefficients sont liés aux ceux de LPC dans la non linéarité. Ils ont été découverts pour être utilisés dans le domaine du codage de la parole **[1,19]**.

Parmi toutes les représentations du LPC, nous trouvons les coefficients cepstraux qui ont été inventés pour assurer une meilleure performance dans la reconnaissance de la parole et celle du locuteur.

L'algorithme de calcul LPC est le suivant :

```
Début
        PC : Tableau de P+1 valeur réel ;
        R : Tableau de P valeur réel ;
        CoefLpc : Tableau de P valeur réel ;
        P : entier ;
        a : Matrice de P+1*P+1 valeur réel ;
        Somme : réel ;

        a [ 1 ] [ 1 ] = - R [ 1 ]  / R [ 0 ] ;    // Calcul du premier coefficient
        PC [ 1 ] = R [ 0 ] * ( 1 – a [ 1 ]  [ 1 ] * a [ 1 ] [ 1 ] ) ;
        Pour  m = 2 jusqu'à  P+1 faire      // Calcul des autres coefficients


                Somme = 0 ;
                Pour i = 1 jusqu'à  m-1 faire
                        Somme = Somme + a [m-1] [i] * R [m-i];
                FinPour
                a [ m ] [ m ] = - ( R [ m ] + Somme  ) / PC [ m-1 ] ;
                Pour i = 1 jusqu'à m faire
                    a [ m ] [ i ] = a [ m-1 ] [ i ] + a [ m ] [ m ] * a [ m-1 ] [ m-i ] ;
                FinPour


                PC [ m ] = PC [ m-1 ] * ( 1 – a [ m ] [ m ] * a [ m ] [ m ] ;
        FinPour

        Pour  m = 1  jusqu'à  P  faire

            CoefLpc[m] =a[i] [i];

        FinPour
Fin
```

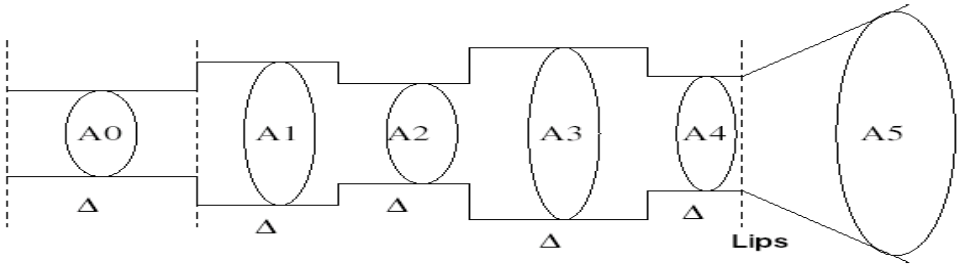


Figure 3.11 : Modèle du tube acoustique de production de la parole.

La Section.	Les Paramètres Utilisés.	Les valeurs Utilisés.
Le filtrage« preemphasis ».	**α.**	**0.95.**
L'échantillonnage« Frameblocker ».	**N.**	**480 (IISC-Microphone).**
	L.	**160 (IISC-Microphone).**
Le fenêtrage« Windowing ».	**ω (n) .**	**0.54-0.46 cos** $\frac{2\pi n}{M-1}$ **.**
L'autocorrélation et l'analyse LPC.	**P.**	**8, 10, 12,40.**
La conversion cepstrale.	**M.**	**P.**

Tableau 3.1: Les paramètres opérationnels utilisés dans l'extraction des caractéristiques avec LPC.

2.4. Classification (SVM et la parole)

Dans le chapitre précédent nous avons parlés des SVM et son utilisation générale, mais dans cette partie nous allons parler de l'utilisation des SVM pour la reconnaissance de la parole. Donc, après l'extraction des paramètres du signal vocal par la méthode LPC, ces paramètre sont utilisés comme une donnée d'entrée pour le composant de classification (SVM), qui va rechercher un hyperplan séparateur qui sépare les exemples dans la phase d'apprentissage et prend une décision de classification dans la phase d'identification.

Dans le module SVM, il y a deux phases : une pour l'apprentissage et l'autre pour la classification. La figure suivante représente la relation entre ces deux phases.

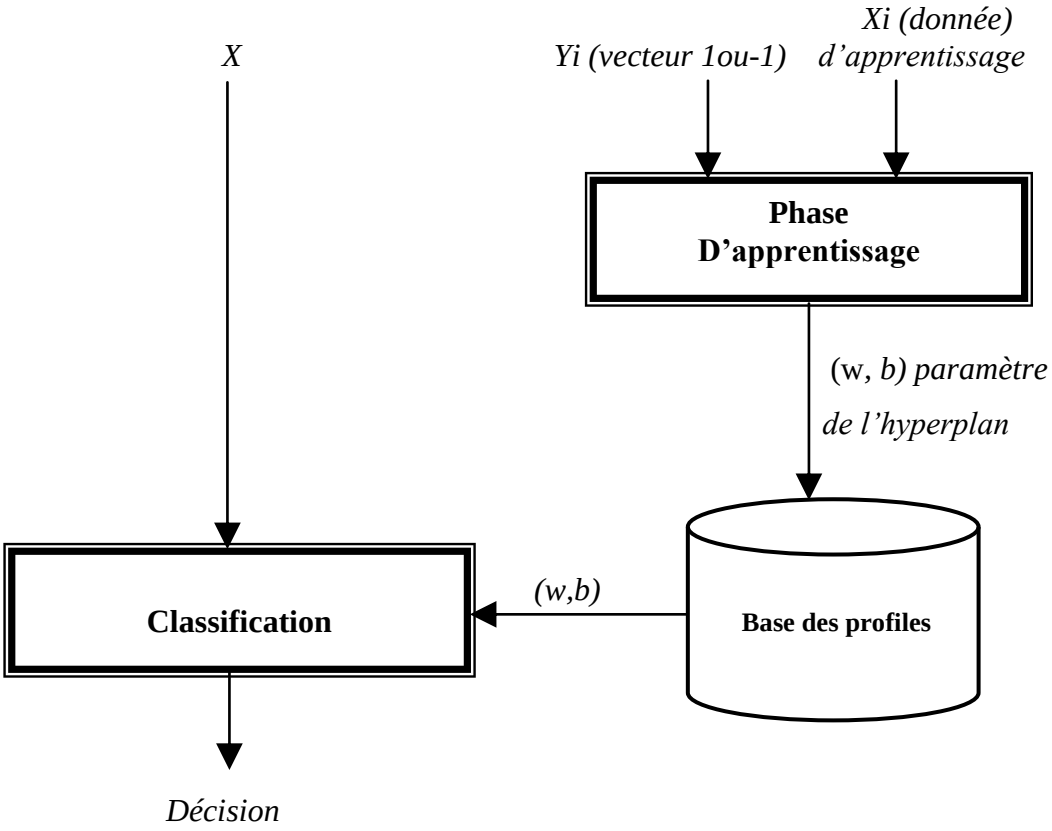


Figure 3.12 : les phases de SVM

Où *X*, *Xi* représentent le vecteur caractéristique des enregistrements du son ; et *Yi* les étiquettes de chaque classe.

Dans la phase précédente nous résolvons un problème dual. Nous connaissons alors *w et b* et nous pouvons définir la fonction de décision pour une nouvelle donnée *x* :

$$f(x) = signe(\langle w.\phi(x)\rangle - b\,.$$

2.5. Post-traitement

Dans la réalité, un nombre important des mots arabe sont composés de deux parties de parole et non pas une seule partie. Dans la phase de segmentation, nous avons parlé de la segmentation du signal en parties de parole ce qui fait le même mot peut être segmenté en deux, et par conséquent, une phase composition des résultats de classification de ces parties est nécessaire.

Par exemple le mot arabe **"النقطة"** est compose de deux parties .

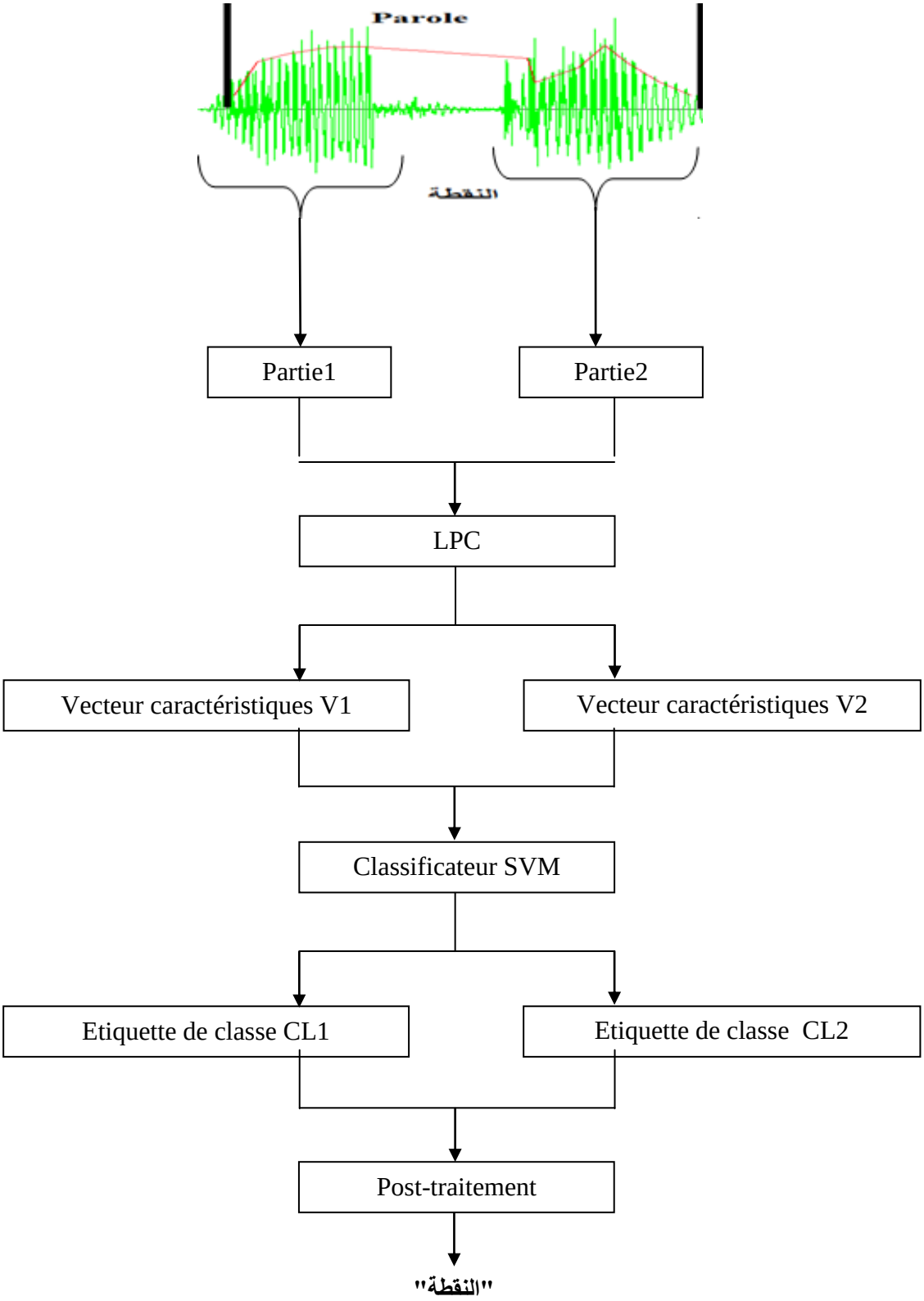


Figure 3.13 : La phase du post-traitement.

Le résultat obtenu par la phase de post-traitement représente une commande qui va engendrer une modification dans le fichier Excel cible.

Conclusion

Nous avons présenté dans ce chapitre les différentes étapes qui peuvent conduire à une conception convenable d'un système de reconnaissance de la parole. Notre système est mono locuteur, on a choisi la méthode de codage LPC pour la paramétrisation du signal acoustique et la méthode SVM pour la reconnaissance des mots isolés. Par la suite, une phase de post-traitement a été utilisée pour améliorer et contrôler les résultats de classification.

Chapitre 4

Réalisation et test de résultats

- **Introduction**
- **Choix du langage de programmation**
- **Interface et fenêtre**
- **Test et résultats**
- **Conclusion**

Introduction

Dans le chapitre précédent nous avons présenté une conception du système en donnant une vue globale du système, en suite nous avons détaillé chaque module composant le système séparément. Dans ce chapitre nous allons voir la réalisation du système : le choix du langage, l'implémentation des différents modules, quelques tests et enfin quelques résultats concernant le taux de reconnaissance.

1. Choix du langage de programmation

Nous avons choisis comme environnement de programmation le langage JAVA qui offre une grande simplicité de manipulation du son, soit en enregistrement (acquisition) ou en génération des fichiers son (sortie). Ce langage possède avantages très intéressants tel que :

- La portabilité des logiciels ;
- La réutilisation de certaines classes déjà développées ;
- La possibilité d'ajouter à l'environnement de base des composants fournis par l'environnement soit même ;
- La quasi-totalité de contrôle de windows (boutons, boites de saisies, listes déroulantes, menus …etc.) qui sont représentés par classes;

2. Interface et fenêtres

En lançant l'application nous allons voir l'interface présentée dans la figure 4.1. L'application peut être utilisée en deux modes :

- Développement ;
- Reconnaissance.

2.1. Mode développement

Le mode développement est chois pour utiliser l'application en mode apprentissage ou mode test.

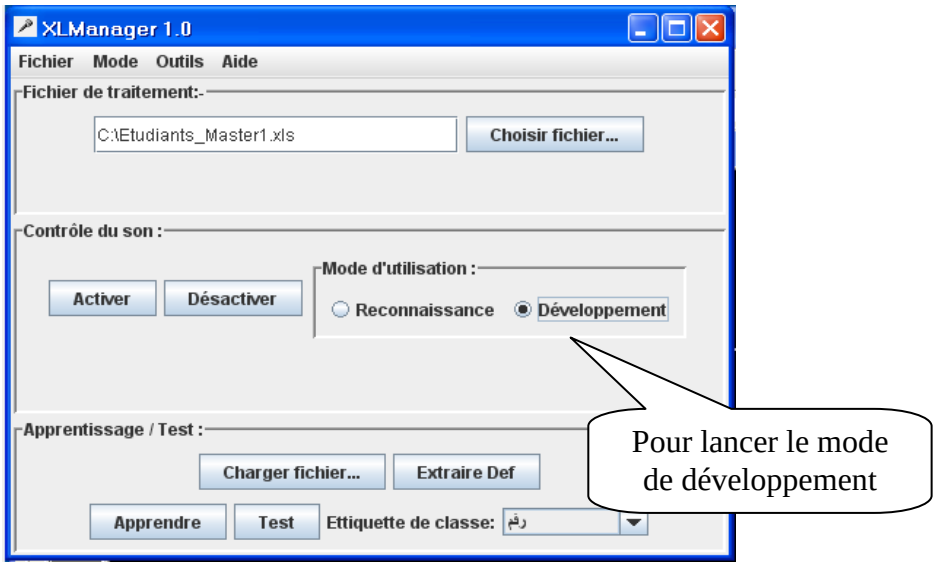


Figure 4.1 : Mode de développement.

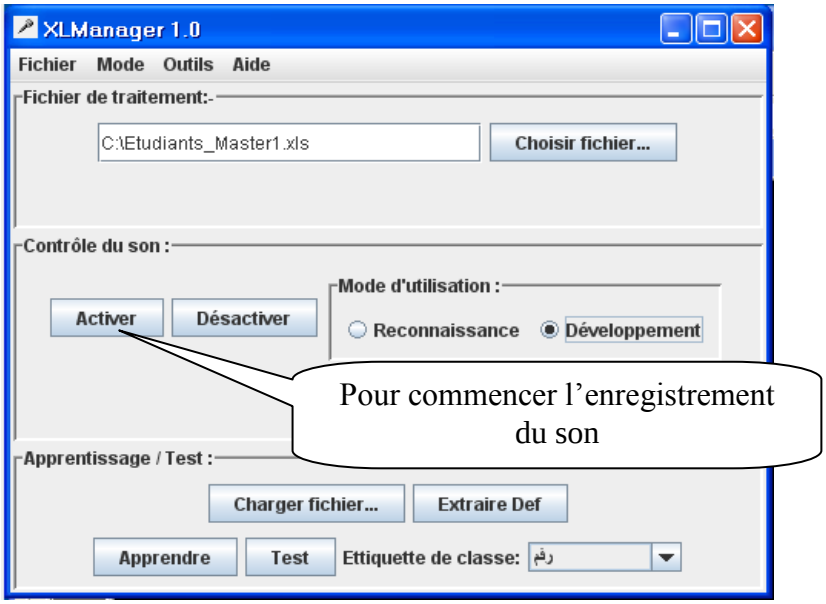


Figure 4.2 : Démarrage d'enregistrement.

Dans ce mode l'application peut être utilisée selon deux sous-modes:

- Apprentissage: comme phase initiale pour aider le système à apprendre les différentes classes.
- Test: pour tester et calculer le taux de reconnaissance, et éventuellement utiliser XLManager.

2.1.1. Mode apprentissage

Ce mode peut être vu comme phase initiale ou d'initialisation de la base de connaissance du système, pour le faire on procède comme suit:

1) choisir une classe cible (l'un des formes primitives), à l'aide du combo box comme le montre la figure suivante:

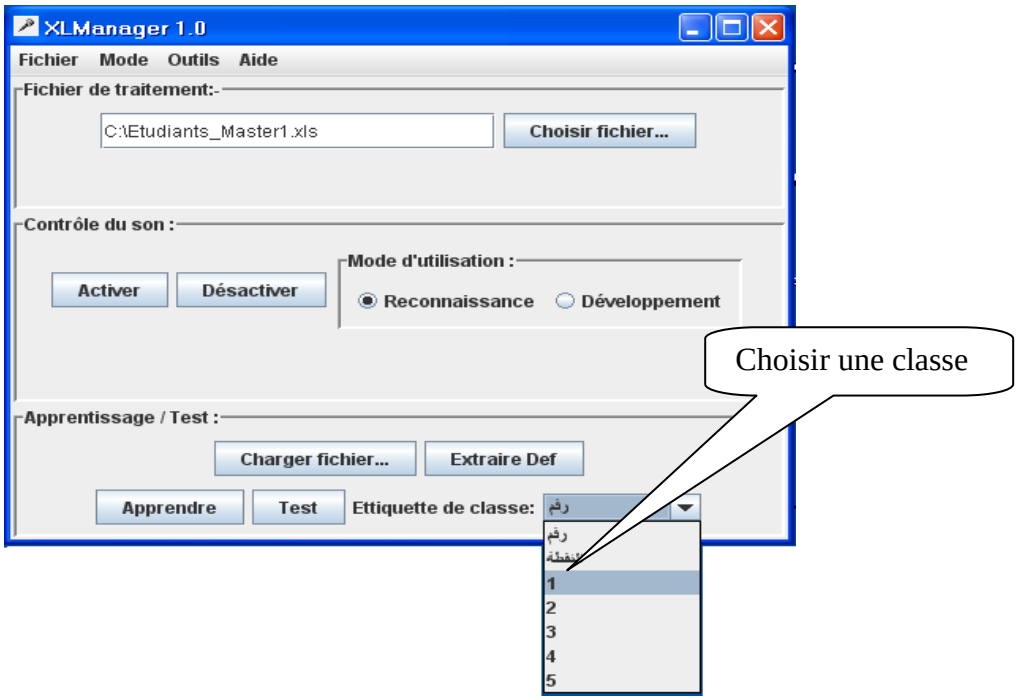


Figure 4.3 : Illustration du choix d'une classe.

2) commencer l'enregistrement de l'exemple de la classe cible, en appuyant sur le bouton *Activer*:

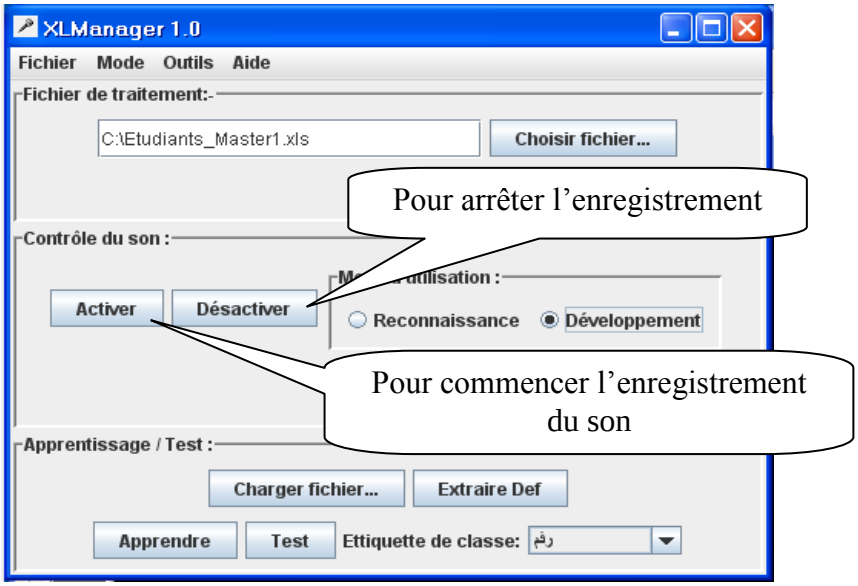


Figure 4.4 : Enregistrer l'exemple de la classe.

En terminant le dicté de l'exemple en question, il faut appuyer sur le bouton *Désactiver*.

3) Cliquer sur le bouton *Extraire Def* pour extraire les caractéristiques de l'exemple enregistré, et les écrire dans le block note du Classifieur.

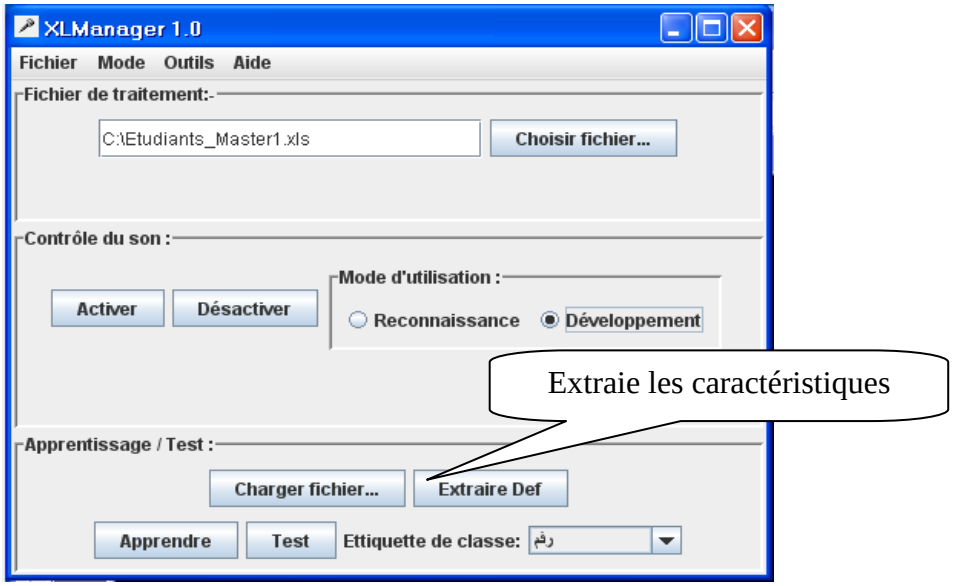


Figure 4.5 : Extraction des caractéristiques de l'exemple enregistré.

Ces étapes décrites ci-dessus sont répétées autant de fois qu'on veut enregistrer d'exemples pour chaque classe. A la fin, nous pouvons appuyer sur le bouton *Apprendre* pour lancer le processus d'apprentissage.

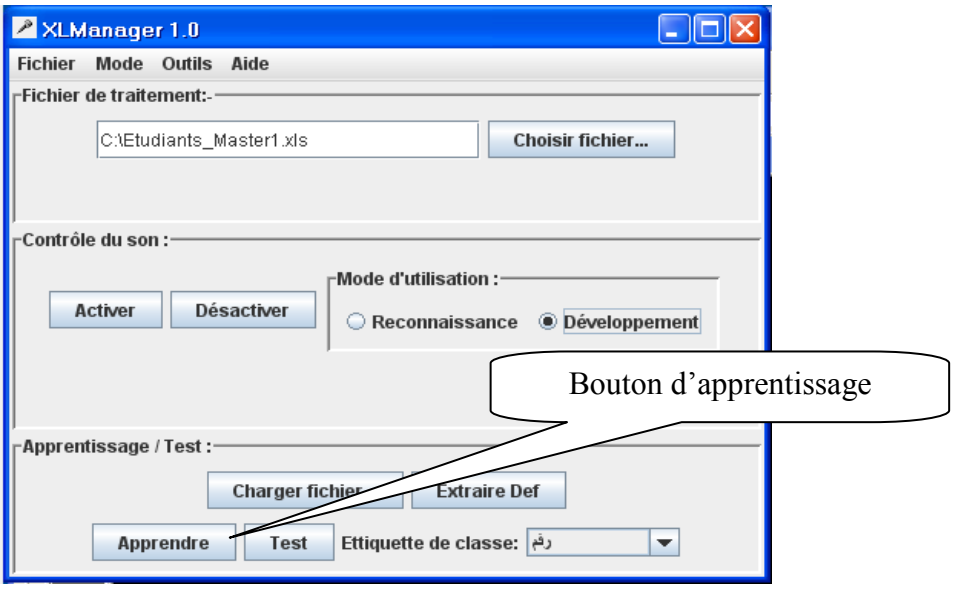


Figure 4.6 : Illustration du mode apprentissage.

2.1.2. Mode test

Ce mode ne peut être exploitable qu'après avoir terminé la phase d'apprentissage et il suit presque les mêmes étapes avec des différences légères. Les étapes 1 et 2 sont les mêmes, pour l'étape 3, on clique sur le bouton de test *Test*.

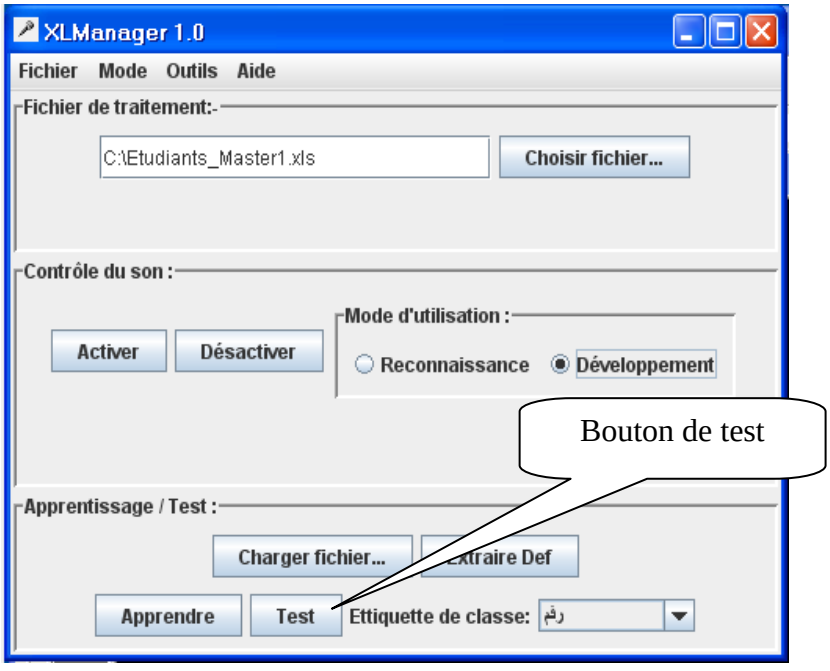


Figure 4.7 : Illustration du mode de test.

Dans la phase de test, l'identité de la commande vocale est d´exterminée à partir d'une comparaison entre les caractéristiques actuelle et les caractéristiques de référence.

2.2. Mode Reconnaissance

Ce mode peut être choisit en cliquant le bouton radio *Reconnaissance* (voir la figure ci-dessous). Après avoir choisir ce mode, il aussi faire le choix d'un fichier Excel cible en appuyant sur le bouton *Choisir fichier*.

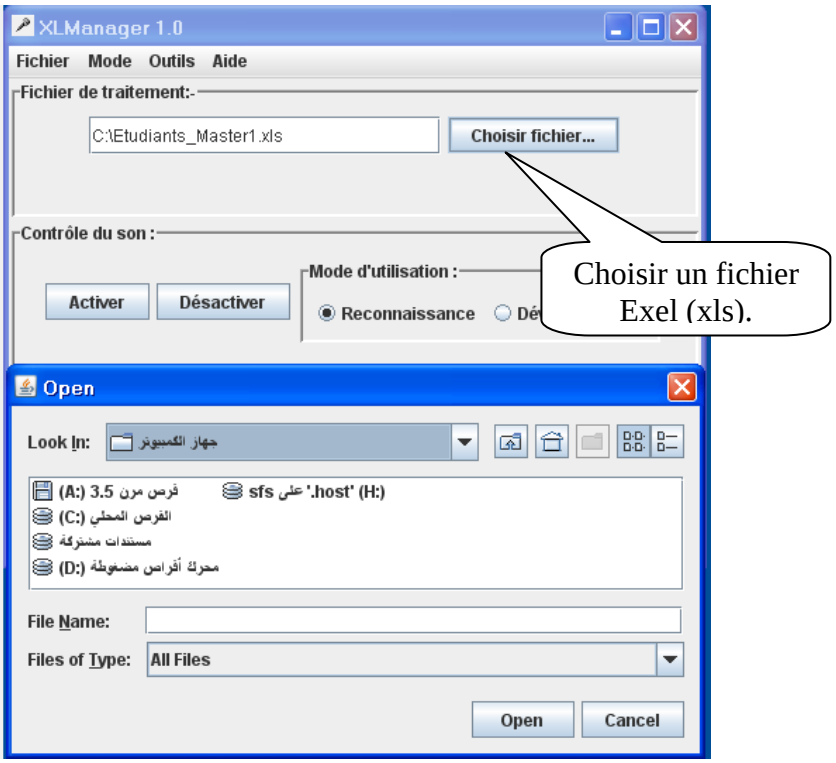


Figure 4.8 : Choisir un fichier Excel (xls).

C'est le mode d'utilisation normale de l'application, où on lance l'enregistrement continu par le bouton *Activer* et le système commence à faire la reconnaissance en temps réel la parole dictée la traduire en commandes qui les appliquent par la suite en modifiant le fichier Excel choisit.

Dans ce travail, nous avons choisis d'utiliser la bibliothèque java JExcelApi qui permet de lire, écrire et modifier des fichiers Excel sous le format 97-2003 (xls).

3. Test et résultats

Dans cette partie nous allons tester la méthode d'apprentissage SVM, et précisément comment déterminer les paramètres qui influent sur l'efficacité de la classification par SVM. Ces paramètres sont le paramètre de pénalisation d'erreur C et pour le noyau choisit (noyau gaussien) on doit déterminer un certain nombre de paramètres pour ajuster sa forme quand à la distribution des données d'apprentissage.

Le choix des paramètres adaptés est une étape cruciale, un ensemble trop encadré peut ne pas parvenir à séparer les données initiales, et au contraire un ensemble trop libre peut aboutir à l'incapacité de généralisation. La méthode de validation croisée est utilisée pour trouver les valeurs les plus adaptées.

Les paramètres C et gamma (sigma) ne sont pas les seuls facteurs qui influent sur le résultat de classification par SVM mais en plus le type de jeu de données a une influence directe, pour bien montrer ça, on a choisi deux classes qui sont facilement séparable « 1 ,2 » et deux autres qui ne le sont pas « 5,7 ».

La configuration de quelques paramètres peut être modifiée par l'utilisateur mais avec prudence, ces paramètres sont les suivant :

- Le paramètre P de LPC =40 ;
- La fréquence d'échantillonnage =44100 ;
- Les paramètres C =100, et gamma =1 de SVM.

On a pris pour chaque classe des exemples dans la phase d'apprentissage, les résultats de test sont résumés dans le tableau suivant :

Classe	Nombre d'exemples	Taux de reconnaissance
1	60	75.66 %
2	60	98.33 %
5	50	97.33 %
Total	170	90.44%

Table 4.1 : Illustration du taux de reconnaissance pour les classes « 1, 2 et 5».

Pour mieux comprendre ces résultats, nous les avons présentés sur l'histogramme et le cercle de secteurs suivants :

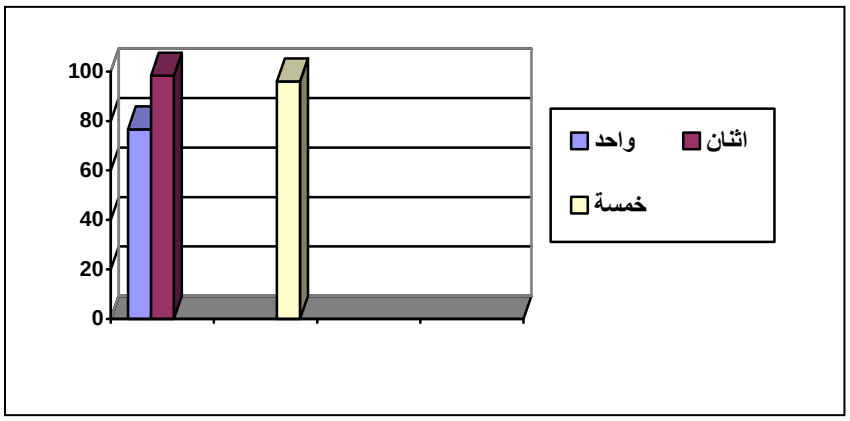


Figure 4.9: Le taux de reconnaissance des classe« 1, 2 et 5».

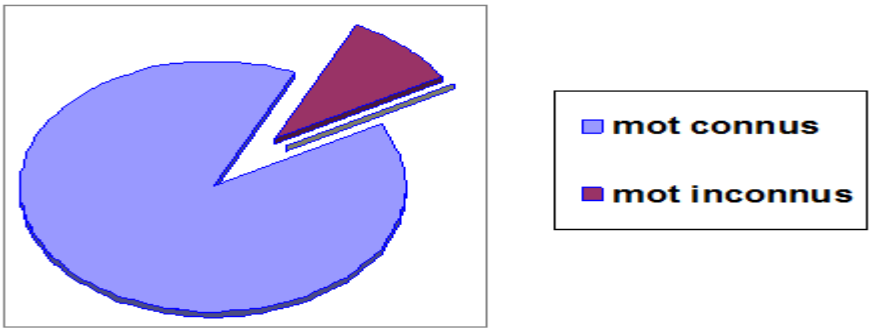


Figure 4.10 : Le taux de reconnaissance globale.

Où mot connus veut dire classes facilement séparables et mot inconnus classes non facilement séparables.

Conclusion

Les résultats de la classification par SVM obtenus peuvent être améliorés en optimisant les paramètres suivants :

- Les paramètres C et gamma.
- Le nombre d'exemples.
- Qualité de matérielle.
- Le bruit et La qualité de Sound traité
- L'environnement d'enregistrement.
- L'état du locuteur.
- …etc.

Comme solution au problème des classes qui ne sont pas facilement séparables, nous pouvons augmenter le nombre d'exemples de l'apprentissage et l'extraction de caractéristiques de la parole.

Conclusion et perspectives

Dans ce mémoire nous présentons les technologies basiques intervenants dans la réalisation d'un système de dicter basée sur un système de reconnaissance de la parole en temps réelle de mots arabe isolées ; l'analyse acoustique du signal par le traitement du signal (LPC) et une méthode de classification SVM (Support Vector Machine), méthode de classification binaire inspiré de la théorie d'apprentissage statistique.

Plusieurs ambiguïtés ont été rencontrées durant notre étude, parmi les quelles nous citons :

- Les conditions d'enregistrement ne répondent pas aux contraintes d'application (bruit, le matérielle, position et sensibilité du microphone…)
- Les états variés des locuteurs (le tempérament du locuteur, état émotif, état de fatigue…). Ces conditions ont une influence sur les résultats obtenus.
- L'outil capteur utilisé n'est pas vraiment fiable.
- La diversité des notions liées au concept de la parole(la reconnaissance de mots prononcés, La dictée vocale , La différentiation entre locuteur masculin, féminin et enfant , la dépendance ou non dépendance du texte…. etc..).
- Le choix de la méthode d'extraction des caractéristiques vocales qui convient au modèle de classification choisi.
- La difficulté d'obtention de caractéristiques vocales différenciables.
- La méthode SVM nécessite des paramètres obtenus par expérience.

Les résultats obtenus sont acceptables, il faut appliquer cette étude dans des circonstances meilleures.

Comme perspective, il est préférable d'utiliser la méthode d'extraction de paramètres MFCC (Mel Frequency Cepstrum Coefficients) au lieu de LPC parcequ'elle est plus robuste dans les environnement bruités. L'utilisation du modèle phonétique au lieu du modèle de mots isolés est beaucoup plus mieux pour minimiser le temps d'apprentissage et le nombre d'exemples.

Bibliographie

[1] **René Boite et Murat Kunt**
« *Traitement de la parole* ».
Presses polytechniques romandes, Lausanne.
1987.

[2] **Maurice Bellanger**
« *Traitement numérique du signal théorie et pratique* ».
Dunod, Paris.
1998-2002.

[3] **Othmani.C et Mazouzi.M**
« *Conception et réalisation d'un système de reconnaissance de locuteur par réseau de neurones artificiels* **».**
Mémoire de fin d'étude en vue de l'obtention du diplôme d'ingénieur en Informatique, Biskra.
Session 2005.

[4] **Rodolphe BATTAULT**
« *La reconnaissance vocale, techniques utilisées, applications actuelles et Futures* »**.**
Examen probatoire pour l'obtention du diplôme d'ingénieur du C.N.A.M, Paris.
1998.

.

[5] http://perso.orange.fr/xcotton/electron/coursetdocs.htm

[6] **José Anibal ARIAS AGUILAR**
« *Méthodes à vecteurs de support et Indexation Sonore* ».
Laboratoire IRIT (**I**nstitut de **R**echerche en **I**nformatique de **T**oulouse).
Année 2003-2004.

[7] http://membres.lycos.fr/guillaumerey/reconnaissance_principes.htm

[8] **Marc Sebban et Gilles Venturini**
« *Apprentissage automatique* ».
Hermes Science Publications, Paris.
1999.

[9] **Jeremy Mary**
« *Méthodes d'apprentissage avancées* ».
***C**entre **N**ational de la **R**echerche **S**cientifique*.
janvier2006.

[10] **Pierre Mahé et Laure Ait-Ali**
« *Projet d'apprentissage statistique SVM pour l'apprentissage non Supervisé* ».
DEA MVA.
Février 2003.

[11] **Antoine Cornuéjols**
« *Une nouvelle méthode d'apprentissage : les SVM. Séparateur à vaste marge* ».
Juin 2002**.**

[12] **Pierre Mahé**
« *Noyaux pour graphes et Support Vector Machines pour le criblage virtuel de molécules* ».
Septembre 2003.

[13] **Pascal Vincent**
« *Modèles à noyau à structure Locale* ».
Thèse présentée à la faculté des études supérieures en de l'obtention du grade de Philosophiæ.
Octobre 2003.

[14] **Olivier Bousquet**
« *Introduction aux Support Vector Machines* ».
Centre de Mathématiques Appliquées, Ecole Polytechniques, Palaiseau.
Novembre 2001.

[15] **Jérôme CALLUT**
« *Implémentation efficace des Support Vector Machines pour la Classification* ».
Mémoire présenté en vue de l'obtention du grade de maître en informatique.
2002-2003.

[16] **Mohamadally Hasan** et **Fomani Boris**
« *SVM : Machines à Vecteurs de Support ou Séparateurs à Vastes Marges* ».
janvier 2006.

[17] **Anderzej Drigajlo**
Traitement de la parole. *«Speech Processing and Biometrics Group (GTPB)»*.
juin 2006.

[18] **Thomas Styger, Bernard Gabioud, Eric Keller.**
« *Méthodes informatiques pour l'analyse de paramètres primaires en parole pathologique* ».
1993.

[19] **Guy Almouzni**
« *Traitement de la parole* ».
Cours et Tps.
2006-2007.

[20] **BENAMMAR Ryadh**
«Traitement Automatique De La Parole Arabe Par Les HMMs : Calculatrice Vocale».
Septembre 2012.

[21] **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 sciences spécialité Informatique
2011-2012.

Sommaire

Chapitre 1: Reconnaissance de la parole

Chapitre 2: Support Vector Machines

Chapitre 3: Conception et implémentation du système

Chapitre 4: Réalisation et test de résultats

Liste des figures

Chapitre 1: Reconnaissance de la parole

Chapitre 2: Support Vector Machine

Chapitre 3: Conception et implémentation du système

Chapitre 4:Réalisation et test de résultats

Liste des tableaux

Chapitre 1: Reconnaissance de la parole

Chapitre 3: Conception et implémentation du système

Chapitre 4:Réalisation et test de résultats

Résumé

Si l'homme a la faculté de comprendre un message vocal provenant d'un locuteur quelconque, dans des environnements souvent perturbés, quelques soient son mode d'élocution, la syntaxe et le vocabulaire utilisés, la machine est-elle capable d'en faire autant ? Une solution peut-elle répondre en globalité à ces difficultés ? Le problème de la reconnaissance vocale est un sujet d'actualité et pour l'instant, seules les solutions partielles sont aptes à répondre aux différentes tâches que la machine doit effectuer.

Ce document est destiné à la conception et à la réalisation d'un système de saisie à l'aide des commandes vocales basée sur l'apprentissage par SVM (support vector machines) l'une des méthodes d'apprentissage inspirée de la théorie de statistique de l'apprentissage de Vladimir Vapnik. C'est une méthode de classification binaire par apprentissage supervisé qui fut introduite par Vapnik en 1995. Cette méthode se base sur la recherche d'un hyperplan séparateur entre les classes dans la phase d'apprentissage et l'utilisation d'une fonction de décision dans la phase de décision.

Une commande vocale issus d'un locuteur passe par une succession d'opérations (Acquisition, Segmentation et extraction de vecteur acoustique, Classification, exécution ou calcul et synthèse de résultat) afin qu'elle soit interprétée et exécutée. Le signal acoustique est premièrement numérisé, ensuite soumis à la méthode LPC pour extraire un vecteur caractéristique. Ce dernier est comparé ensuite par la méthode SVM à d'autres vecteurs pour trouver un ressemblant dans une base de sons. Une fois la classe trouvée la commande est décodée et exécutée.

Mots clé

RAP : Traitement de la parole, Reconnaissance de la parole.
LPC : Codage prédictif linéaire.
SVM : Support Vector Machine.

Introduction générale

Aujourd'hui, l'impact des systèmes de RAP est encore minime dans la vie courante, et la commande des ordinateurs ne s'effectue toujours pas par la voix, malgré les promesses de fabricants de logiciel ou de matériel informatique (Microsoft, Apple). L'annonce de la commercialisation du système de dictée vocale d'IBM pour les ordinateurs PC en 1994 a suscité de l'intérêt, mais aussi des réserves quant aux performances actuelles du système. Pourtant les progrès réalisés depuis 25 ans en RAP sont très importants, grâce à un grand nombre de recherches traitant du problème sous tous ses aspects. Les limitations de la capacité des systèmes de reconnaissance, imposées à l'origine par la complexité de la tâche, sont progressivement repoussées, et des systèmes efficaces pour des applications spécialisées sont maintenant disponibles et commercialisés.

Notre étude s'intègre dans le cadre du développement d'un système de dictée vocale indépendant du locuteur (logiciel de saisie des notes des étudiants par dicter initialiser les nombres de l'inscription). La modélisation acoustique par les méthodes les plus performantes de l'état de l'art reste insuffisante; cette faiblesse est un facteur limitant des systèmes de RAP. Nous cherchons à améliorer la qualité de la modélisation acoustique, en intégrant certains traitements adaptés à un processus de décodage de la parole continue.

Le chapitre 1 présente les méthodes classiques employées en reconnaissance de la parole. Les difficultés rencontrées pour la mise au point des systèmes de RAP proviennent de la variabilité du signal de parole et de la continuité du processus de production.

Parmi les méthodes développées, l'approche statistique SVM (Support Vector Machine) semble la plus efficace. En introduisant un nombre d'exemples présentant des échantillons pour chaque classe plus les étiquettes de chaque classe nous pouvons définir un hyperplan séparant chaque classe de l'autre, ce qui est présenté en chapitre 2.

Le chapitre 3 présente la conception du système où en définissant les différents modules du système leur architecture générale, ensuite nous illustrant notre implémentation ainsi que la validation de notre système abordé dans le chapitre 4.

Finalement, nous terminons notre mémoire par une conclusion et les perspectives de notre projet.

Chapitre 1

Reconnaissance de la parole

- **Introduction**
- **Généralité**
- **Traitement de la parole**
- **La reconnaissance de parole**
- **Conclusion**

Introduction

Dans ce chapitre, nous allons parler sur les principales techniques associées au prétraitement du signal de la parole, on proposant un état de l'art de la reconnaissance vocale et suivant le processus de génération de la parole jusqu'à sa reconnaissance. On y trouvera les domaines d'application, le mécanisme de production de la parole et les paramètres qui la caractérisent, les principes des techniques dominantes d'analyse du signal.

L'information portée par le signal de parole peut être analysée de bien des façons. On en distingue généralement plusieurs niveaux de description non exclusifs : acoustique, phonétique, phonologique, morphologique, syntaxique,

1. Généralités

1.1. Panorama sur la reconnaissance de la parole

1.1.1. Historique

Une évolution rapide :

Les premiers travaux qui se relient directement à la reconnaissance automatique de la parole furent ceux de **J.Dreyfus-Graf**, en Suisse puis en France.

1949 : Visualisation sur un oscilloscope du signal de parole filtré dans six bandes de fréquence différentes.

1952 : reconnaissance des 10 chiffres par un dispositif électronique câblé.

1956 : Système de distinction des voyelles pour différents locuteurs et première " machine à écrire phonétiquement ", dix syllabes par locuteur.

1960 : Utilisation des méthodes numériques.

1960 - 1970 : L'ordinateur recherche automatiquement des sons spécifiques et les met en mémoire pour référence ultérieure.

1968 : reconnaissance de mots isolés par des systèmes implantés sur gros ordinateurs (jusqu'à 500 mots)

1970 : Utilisation des niveaux syntaxiques et sémantiques.

1972 : premier appareil commercialisé de reconnaissance de mots.

1990 : premières véritables applications de dialogue oral homme-machine.

1994 : IBM lance son premier système de reconnaissance vocale sur PC.

1.1.2. Problèmes rencontrés durant ces années

1.1.2.1. Continuité

Contrairement au langage écrit, où les mots sont séparés par des "blancs" dans les textes manuscrits ou par des espaces dans les textes dactylographiés, les séparateurs, symbolisés par les silences entre les mots, sont parfois très difficiles à repérer **[4].**

1.1.2.2. Variabilité

Elle provient de la position d'un phonème par rapport aux autres (coarticulation), des locuteurs aux timbres différents : homme, femme, enfant et à leur mode d'élocution : voix chantée, criée, enrouée, sous stress,....Elle est due aussi à la qualité du moyen d'acquisition et du bruit environnemental **[4].**

1.1.2.3. Reconnaissance des informations en fonction de la tâche à accomplir

La reconnaissance vocale peut s'effectuer sur les sons eux-mêmes, sur la structure syntaxique d'une phrase (dictée), sur la signification d'une phrase (robots) ou sur l'identité du locuteur et son état émotionnel (joyeux, en colère,...) **[4]**.

1.1.2.4. Depuis 1970

Les difficultés rencontrées durant ces débuts ont amené les scientifiques à classifier, puis à déterminer des axes de recherches suivant le tableau 1.1 **[4]**.

		Type d'élocution	
		Mots isolés	**Parole continue**
Approche	**Globale**	Reconnaissance de mots (petits vocabulaires) Systèmes existants	Localisation de mots Dans les phrases
	Analytique	Reconnaissance de mots (grands vocabulaires)	Localisation de mots Reconnaissance et compréhension de phrases

Tableau 1.1 : Les problèmes de la reconnaissance de la parole.

1.1.2.5. L'approche globale

Ce domaine de recherche concerne la reconnaissance, après une phase d'apprentissage, de quelques mots isolés pour un même locuteur. Elle se concrétisa en 1972 par l'industrialisation du VIP100, puis du VNC par la société Threshold Technologie (30 mots reconnus avec un taux proche de 100%) **[4]**.
La fin de cette décennie fut marquée en France par Martine Kempf et son "Katalavox".

1.1.2.6. L'approche analytique

C'est une voie de recherche fondamentale qui concerne la reconnaissance et la compréhension de la parole continue, multi locuteur, à grand vocabulaire et langage peu contraint.

Cette méthode, basée sur l'identification d'éléments phonétiques, engendra ces années là un recours massif aux traitements du type intelligence artificielle pour pallier aux erreurs de décodage des phonèmes.

Trois systèmes issus du projet "ARPA/SUR "virent le jour aux USA.

La recherche française aussi active produisit les systèmes "Myrtille 1 et 2" au C.R.I.N., Keal au C.N.E.T. de Lannion, Esope au L.I.M.S.I. à Orsay et Arial II au C.E.R.F.I.A. de Toulouse.

On remarque déjà, à la fin de ces années, l'importance prise par la modélisation "Markovienne du langage **[4]**.

1.1.2.7. Le mécanisme de la parole

L'appareil phonatoire humain (Figure 1.1) peut être assimilé, et est même souvent modélisé comme un système composé d'une source et d'un filtre. La source est un élément qui vibre soit dans un mode harmonique, soit dans un mode "aléatoire" quand il y a une constriction au niveau des cordes vocales et donc un écoulement turbulent de l'air. Le filtre résulte du conduit vocal qui est formé d'une cavité résonante complexe **[1,4]**.

1.1.3. Les résonateurs

Les cordes vocales sont les éléments vibreurs ; et comme une anche d'un instrument de musique, elles possèdent la particularité de produire, en plus de leur fréquence fondamentale, un spectre riche en harmoniques.

Mais un élément vibreur, placés devant une cavité résonante, produira alors un son dont les fréquences seront filtrées par la bande passante du résonateur.

Les ordres de grandeur des fréquences fondamentales sont de 120Hz pour les hommes, 250Hz pour les femmes et de 450Hz pour les enfants **[1,4]**.

1.1.4. L'appareil phonatoire

Le résonateur de l'appareil phonatoire est composé de quatre cavités principales en “série (Figure.1.1): le Pharynx ou arrière gorge, les deux cavités buccales délimitées par la langue et que l'on simplifiera à une seule et l'ajutage labiale situé entre les dents et les lèvres. La cavité nasale, en “parallèle” sur l'ensemble série précédent, vient compléter ce résonateur.

La source de ce résonateur est en fait décomposable en deux émissions distinctes et d'origines différentes. Les cordes vocales, en fournissant un spectre riche en harmoniques, produisent les sons voisés. Le bruit d'écoulement de l'air en provenance des poumons, dont le spectre est similaire à un bruit blanc, crée les sons non voisés.

Les sons et donc la parole naissent de l'excitation d'un résonateur et sont formés par les ouvertures et les volumes de ce dernier qui varient très rapidement.

L'observation spectrale du conduit vocal laisse apparaître des pics de résonance, appelés formants. Les affaiblissements constatés dans le spectre, nommés anti-formants, sont introduits par les sons nasalisés **[1]**.

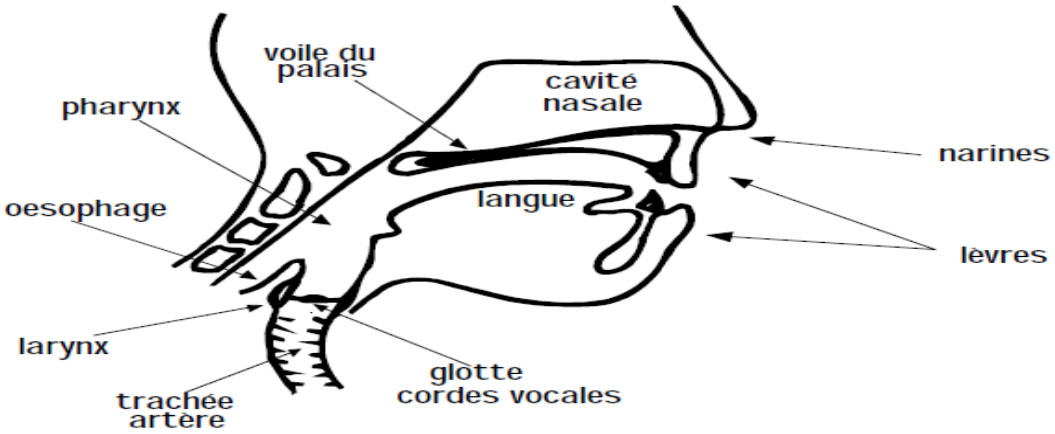


Figure 1.1 L’appareil phonatoire.

1.2. L'information vocale

Le signal de la parole véhicule plusieurs types d'informations, tels que le fondamental, la prosodie, le timbre et les phonèmes. Par conséquent, ceci impose, aux systèmes de reconnaissance vocale, de n'extraire que l'information nécessaire à son application, les phonèmes pour les machines de dictée par exemple.

La parole est surtout contenue dans les deux premiers formants, mais l'information proprement dite provient des transitions for mantiques.

En général, on considère que la plage de fréquence d'un signal de parole se situe dans la bande de 100Hz-5KHz (300Hz-3.4KHz pour la téléphonie) **[4]**.

1.3. L'appareil auditif

1.3.1. Echelles des hauteurs

1.3.1.1. L'échelle des Mels

Après 500Hz, l'oreille perçoit moins d'une octave pour un doublement de la fréquence. Des expériences psycho acoustiques ont alors permis d'établir la loi qui relie la fréquence et la hauteur perçue : l'échelle des Mels où le « Mel » est une unité représentative de la hauteur perçue d'un son (Figure 1.2) **[4]**.

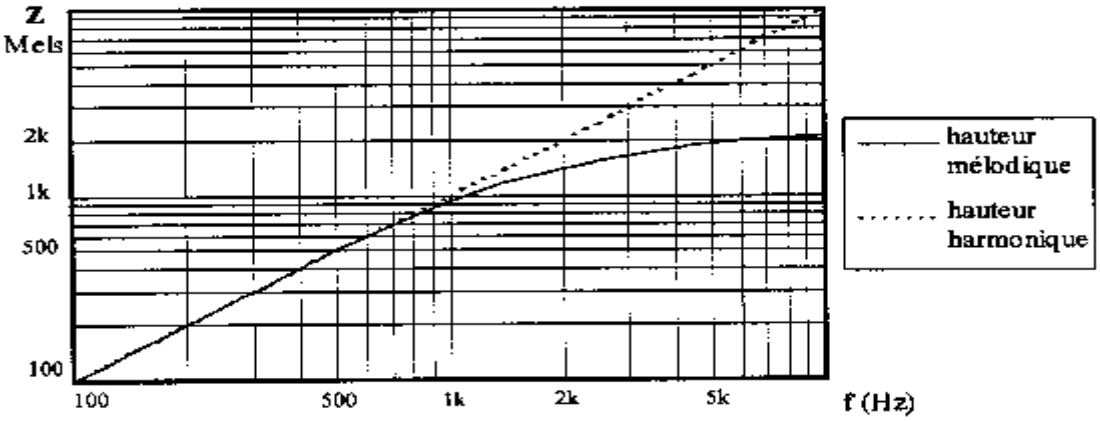


Figure 1.2 : L'octave d'un son de 2000Hz (1800 Mels) sonnera l'octave supérieure à 4600Hz (3600 Mels) au lieu de 4000Hz.

1.3.1.2. L'échelle de Bark

Le système auditif se comporte comme un banc de filtres dont les bandes, appelées “bandes critiques”, se chevauchent et dont les fréquences centrales s'échelonnent continûment. Cette bande critique correspond à l'écartement en fréquence nécessaire pour que deux harmoniques soient discriminées dans un son complexe périodique.

Remarque : Les échelles Mel ou de Bark sont approchées par un banc de 15 à 24 filtres triangulaires espacés linéairement jusqu'à 1KHz, puis espacés logarithmique ment jusqu'aux fréquences maximum **[4]**.

1.4. Les méthodes de reconnaissance vocale

1.4.1. Technologie analogique : Le spectrographe

Le spectrographe de la parole est un appareil inventé voilà plus d'un demi-siècle et commercialisé plus tard sous le nom de Sonagraphe. Historiquement, ce premier outil d'analyse pour les phonéticiens (Figure1.3) était composé d'un banc de filtres analysant les différentes fréquences successivement.

Une autre technique de cet appareil est basée sur le filtrage hétérodyne : on fait défiler le signal vocal, modulé en amplitude par une sinusoïde variable en fréquence, sous un filtre fixe. On recueille alors l'énergie pour chaque incrément de fréquence. Le signal évoluant dans le temps, on obtient alors une représentation graphique à deux dimensions (fréquence temps), nommée "sonagramme" et dont l'intensité est représentée par une échelle de gris **[4].**

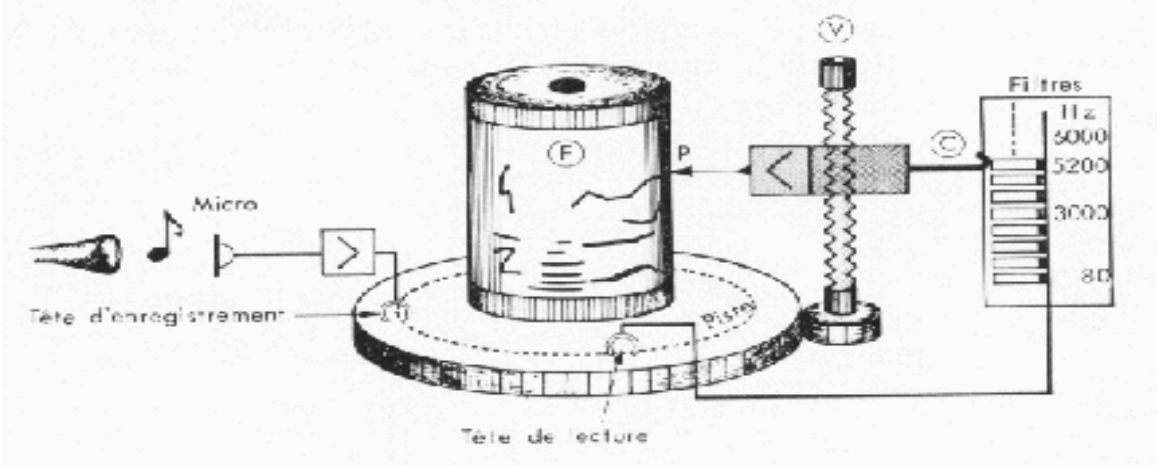


Figure 1.3 : Le spectrographe de la parole

1.4.2. Technologie numérique, introduction

Les systèmes de reconnaissance vocale numériques sont caractérisés par :

- le prétraitement qui comprend l'acquisition du signal de la parole et l'extraction des paramètres,
- l'apprentissage du vocabulaire et la comparaison aux références,
- le traitement des résultats en fonction de l'application finale.

Ces trois fonctions sont réalisées suivant deux approches : l'approche globale et l'approche analytique **[4]**.

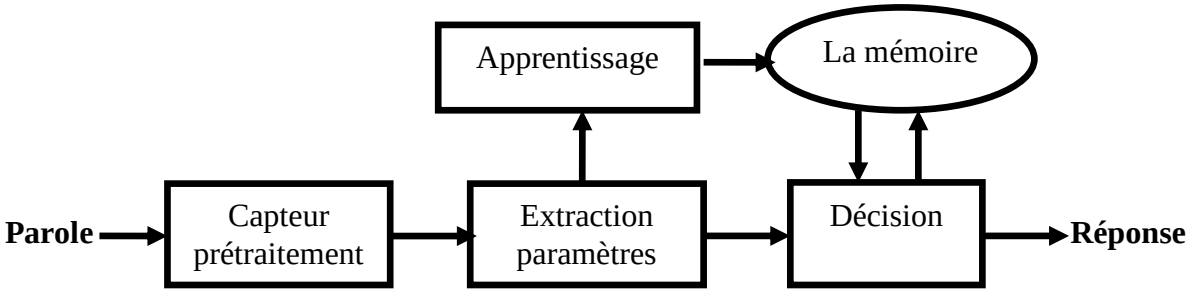


Figure 1.4: système de la reconnaissance vocale numérique.

1.4.3. La reconnaissance globale

Dans cette approche globale, dite aussi acoustique, on considère le message, mot ou groupe de mots, comme une forme insécable en lui attribuant une classe d'appartenance : le mot ou la

phrase sont donc les unités de base du décodage et ne sont définis qu'à partir de paramètres purement acoustiques numérisés **[4]**.

La reconnaissance globale comprend deux phases distinctes :

La phase d'apprentissage pendant laquelle un ou plusieurs locuteurs prononcent une ou plusieurs fois chacun des mots de l'application prévue. Ces prononciations sont toutes prétraitées puis conservées telles quelles ou bien moyennées dans un dictionnaire de références en tant que “ images acoustiques ”.

Puis la phase de reconnaissance où le signal à reconnaître subit le même prétraitement que la phase précédente. Il est ensuite comparé aux références contenues dans le dictionnaire. Le calcul d'une « distance » et sa comparaison à un seuil permet ou non de retenir la ou les références les plus proches.

Mais les différences de prononciations et les variations de débit d'élocution, parfois importantes et non linéaires imposent l'utilisation d'algorithmes de comparaison tels que la comparaison dynamique ou les chaînes de Markov.

C'est une méthode bien adaptée aux applications mono locuteur, à faible vocabulaire et plutôt à mots isolés.

1.4.4. La reconnaissance analytique

Par cette approche, appelée aussi analyse **phonétique,** on considère la segmentation du message en **constituants élémentaires** tels que les phonèmes, les di phonèmes ou les tris phonèmes (Figure 1.5). En effet, ces éléments présentent l'avantage d'être en nombre réduit : 37 phonèmes permettent de décrire le français parlé et une analyse statistique réalisée au LIMSI a montré, qu'à partir d'un répertoire de 627 di phonèmes, il était possible de reconstituer n'importe quelle phrase en français.

Quant aux tris phonèmes, ou triplet phonétique, il est constitué d'un phonème et de ses transitions antérieures et postérieures. Ils sont bien sûr en plus grand nombre, mais ils ont l'avantage de prendre en compte la coarticulation des phonèmes.
Le caractère continu du signal vocal complique beaucoup la reconnaissance de la parole : aucun indice acoustique ne permet de localiser les frontières de mots. Ce problème est abordé, après la phase de prétraitement, d'une part par un décodage acoustique phonétique (DAP) permettant la transcription de la phrase sous forme d'une suite d'éléments phonétique du langage, et d'autre part par un traitement linguistique faisant appel à diverses sources d'informations (lexicales, syntaxiques, sémantiques) permettant la reconnaissance des mots.
Ce sont donc des systèmes à architectures logicielles complexes à plusieurs sources de connaissances qui pallient aux problèmes de reconnaissances des phrases **[4]**.

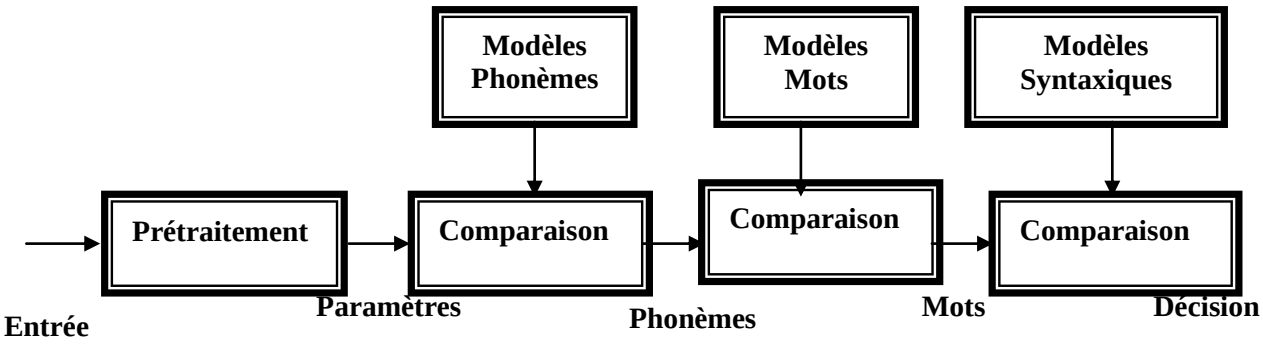


Figure 1. 5 : Schéma synoptique d'un système de reconnaissance de parole selon une approche comparaison.

2. Traitement de la parole

2.1. Introduction

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage (c'est-à-dire du traitement de données symboliques), cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications. L'importance particulière du traitement de la parole dans ce cadre plus général s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine **[1].**

2.2. Le niveau acoustique

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La phonétique acoustique étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone (lui-même associé à un préamplificateur).

De nos jours, le signal électrique résultant est le plus souvent numérisé. Il peut alors être soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les traits acoustiques : sa fréquence fondamentale, son énergie, et son spectre. Chaque trait acoustique est lui-même intimement lié à une grandeur perceptuelle : pitch, intensité, et timbre.

L'opération de numérisation, schématisée à (la figure 1.6), requiert successivement : un filtrage de garde, un échantillonnage, et une quantification.

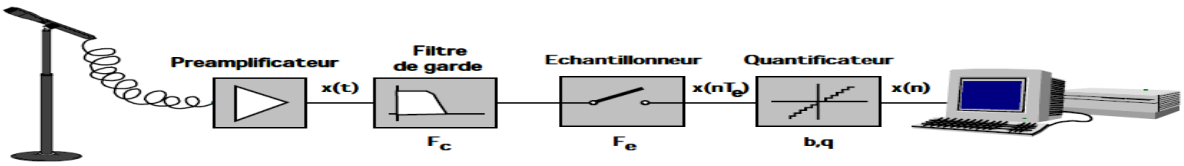


Figure 1.6: Enregistrement numérique d'un signal acoustique.

La fréquence de coupure du filtre de garde, la fréquence d'échantillonnage, le nombre de bits et le pas de quantification sont respectivement notés fc, fe, b, et q **[3]**.

2.2.1. Audiogramme

L'échantillonnage transforme le signal à temps continu x (t) en signal à temps discret X (nTe) défini aux instants d'échantillonnage, multiples, entiers de la période d'échantillonnage Te, celle-ci est elle-même l'inverse de la fréquence d'échantillonnage fe.

Pour ce qui concerne le signal vocal, le choix de fe résulte d'un compromis. Son spectre peut s'étendre jusque 12 kHz **[3]**.La figure 1.7 représente l'évolution temporelle, ou audiogramme du signal vocal pour les mots "بسم الله الرحمان الرحيم".

Figure 1.7 : Audiogramme de signaux de parole.

2.2.2. Transformée de Fourier à court terme

La transformée de Fourier à court terme est obtenue en extrayant de l'audiogramme une trentaine de ms de signal vocal, en pondérant ces échantillons par une fenêtre de pondération (souvent une fenêtre de Hamming) et en effectuant un transformée de Fourier sur ces échantillons.(La figure 1.9) illustre la transformée de Fourier d'une tranche voisée et celle d'une tranche non voisée. Les parties voisées du signal apparaissant sous la forme de successions de pics spectraux marqués, dont les fréquences centrales sont multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière. La forme générale de ces spectres, appelée **enveloppe spectrale**, présente elle-même des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal et sont appelés **formants** et **anti-formants [2]**.

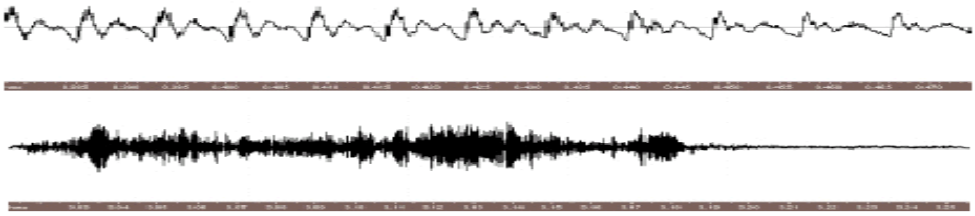


Figure 1.8: Exemples de son voisé (haut) et non voisé (bas) **[1]**.

L'évolution temporelle de leur fréquence centrale et de leur largeur de bande détermine le timbre du son. Il apparaît en pratique que l'enveloppe spectrale des sons voisés est de type passe bas, avec environ un formant par kHz de bande passante, et dont seuls les trois ou quatre premiers contribuent de façon importante au timbre. Par contre, les sons non-voisés présentent souvent une accentuation vers les hautes fréquences **[2]**

2.2.3. Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un **spectrogramme**. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. On parle de spectrogramme **à large bande** ou **à bande étroite** selon la durée de la fenêtre de pondération. Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms), ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales **[3]**.

2.2.4. Fréquence fondamentale

Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la fréquence fondamentale ou pitch. La figure 1.10 donne l'évolution temporelle de la fréquence fondamentale de la phrase « les techniques de traitement de la parole ». On constate qu'à l'intérieur des zones voisées la fréquence fondamentale évolue lentement dans le temps.
Elle s'étend approximativement de 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes, et de 200 à 600 Hz chez les enfants **[1]**.

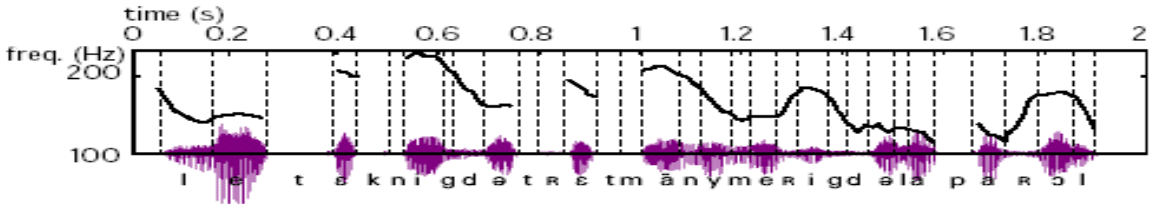


Figure 1.9 : Evolution de la fréquence de vibration des cordes vocales.

2.3. Généralités sur le signal vocal

La parole transmet la pensée, considérée comme information, à travers le canal acoustique par l'intermédiaire de sons articulés différenciés. Cette information transmise est de trois dimensions, vu qu'elle représente trois composantes d'informations transmises, dont la plus importante est l'information linguistique qui correspond souvent à la signification d'une articulation. La deuxième composante étant l'information sociolinguistique (région de l'interlocuteur et sa classe sociale). La dernière composante correspond à l'information personnelle de l'interlocuteur (son identité, sa qualité de voix et ses habitudes d'articulation).Ces trois composantes d'information sont combinées en un seul signal de parole.

Le signal vocal transmet simultanément deux types de messages : un message sémantique convoyé par la parole, expression verbale de la pensée, et un message esthétique perceptible au travers des qualités esthétiques de la voix (timbre, intonation, débit,.. etc.) **[5].**

2.3.1. Caractéristique d'un signal vocale

En plus de ses caractéristiques, en tant qu'onde longitudinale (l'oscillation a la même direction que la propagation), qui sont la fréquence, la longueur d'onde, et la vitesse de propagation qui dépend du milieu matériel de propagation. Le son a par conséquent d'autres caractéristiques, qui sont :

2.3.1.1. La hauteur

C'est la qualité qui fait distinguer un son gras d'un son aigu .La hauteur d'un son est liée à la fréquence des vibrations de la source sonore .Les sons aigus sont dus aux mouvements vibratoires de fréquence élevée, à la différence des sons graves qui sont dus aux mouvements de basse fréquence. **[3].**

2.3.1.2. L'intensité

C'est la qualité qui fait distinguer un son fort d'un faible. L'intensité est liée à la pression de l'air en amont du larynx, qui fait varier l'amplitude des vibrations sonores **[3]**.

2.3.1.3. Le timbre

Le timbre est l'ensemble des caractéristiques qui permettent de différencier une voix.Il provient en particulier de la résonance dans la poitrine, la gorge la cavité buccale et le nez sont les amplitudes relatives des harmoniques du fondamental qui déterminent le timbre du son.

Les éléments physiques du timbre comprennent :

- Les relations entre les parties du spectre, harmoniques ou non.
- les bruits existant dans le son (qui n'ont pas de fréquence particulière, mais dont l'énergie est limitée à une ou plusieurs bandes de fréquence).
- L'évolution dynamique globale du son.
- L'évolution dynamique de chacun des éléments les uns par rapport aux autres [20].

2.3.1.4. Fréquence et amplitude

La répétition d'une forme d'onde périodique est appelée un cycle, et la fréquence fondamentale de la forme d'onde est le nombre de cycles qui se produit par seconde.
Lorsque la longueur du cycle appelée longueur d'onde ou période augmente, la fréquence en cycles par seconde diminue et vice versa.
Nous substituons Hz pour 'cycle par seconde' en conformité avec la terminologie standard de l'acoustique (Hz est une abréviation de Hertz) d'après le nom de l'acousticien allemand Heinrich HERTZ) **[3]**.

2.3.1.5. Le théorème de l'échantillonnage

Définit la relation entre le taux d'échantillonnage et la largeur de bande du signal transmis.
Il fut énoncé par Harold NYQUIST (1928) comme suit :
« Pour toute déformation donnée du signal reçu, le domaine de fréquence transmis doit être augmenté en proportion directe avec la vitesse du signal. La conclusion est que la largeur de fréquence est directement proportionnelle à la vitesse. » Le point essentiel du théorème de l'échantillonnage peut être établi précisément comme ceci :
«Afin d'être capable de reconstruire un signal, la fréquence d'échantillonnage doit être le double de la fréquence du signal échantillonné ».
En raison de sa contribution à la théorie de l'échantillonnage, la plus haute fréquence qui puisse être produite dans un système audionumérique. C'est-à-dire la moitié du taux d'échantillonnage est appelée 'la fréquence de NYQUIST'. Dans les applications musicales, la fréquence de NYQUIST est en général dans le domaine supérieur à celui de l'écoute humaine, au dessus de 20 KHZ. Ainsi la fréquence d'échantillonnage peut être spécifié comme étant au moins le double : au dessus de 40 KHZ **[5]**.

2.3.1.6. Fréquence d'échantillonnage idéale

La question de savoir quelle fréquence d'échantillonnage est idéale pour l'enregistrement et la reproduction musicale de haute qualité est un débat encore en cours.
L'une des raisons est que la théorie mathématique et la pratique des ingénieurs rentrent souvent en conflit :

Les horloges des convertisseurs ne sont pas stables, leurs voltages ne sont pas linéaires, les filtres introduisent de la distorsion de phase et ainsi de suite. Une autre des raisons est que beaucoup de personnes entendent des informations 'on emploie alors le terme 'ambiance' 'dans la région située autour de la 'limite' humaine d'écoute de 20KHZ**.**

Dans les applications d'échantillonnage et de déplacement des hauteurs, le manque de hauteur libre nécessite un filtrage passe bas des échantillons avant que ceux-ci ne soient déplacés vers le haut. Il est clair que des enregistrements à un taux d'échantillonnage élevé sont préférables d'un point de vue artistique, bien qu'ils posent des problèmes pratiques de stockage et la nécessité d'avoir des systèmes de reproduction de haute qualité afin que cet effort en vaille la peine **[3].**

2.3.1.7. Quantification

L'échantillonnage à intervalles de temps discrets dont nous avons parlé dans les parties précédentes, constitue l'une des différences majeures entre les signaux analogiques et les signaux numériques.
Une autre différence est la quantification ou résolution d'amplitude discrète. Les valeurs du signal échantillonné ne peuvent pas prendre n'importe quelle valeur. Ceci en raison du fait que les nombres numériques ne peuvent être représentés qu'à l'intérieur d'un certain domaine et avec une certaine exactitude, qui varie selon le matériel utilisé. Les implications de ceci sont un facteur important de la qualité audionumérique. **[2].**

2.3.1.8. Définition du bruit

On appelle bruit tout phénomène perturbateur (interférence, bruit de fond, etc.) gênant la perception ou l'interprétation d'un signal, ceci par analogie avec les puissances acoustiques de même nom (Figure 1.11) **[3].**

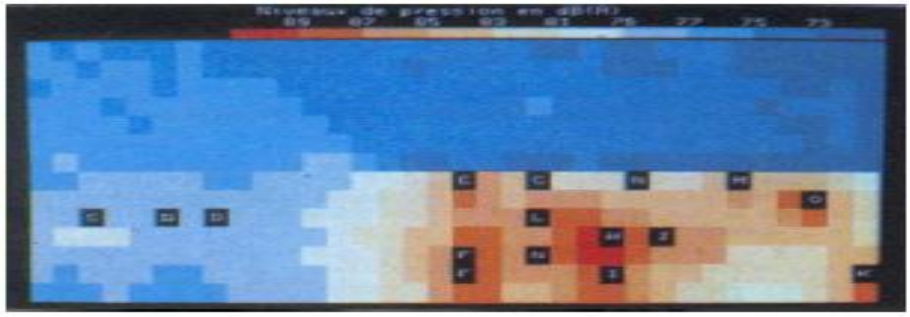


Figure 1.10 : Exemple de carte prévisionnelle de niveaux sonores du plus bruyant (rouge) au plus silencieux (bleu foncé).

2.3.2. Conversion Analogique Numérique

Le son atteint les oreilles de l'auditeur après avoir été transmis par l'air depuis sa source. Les auditeurs entendent des sons car la pression de l'air change légèrement dans leurs oreilles. Si la pression varie selon un modèle répétitif, nous disons que le son a une forme d'onde périodique s'il n'y a pas de modèle discernable, on parle de bruit. Entre ces deux extrêmes se trouve le vaste domaine des sons quasi périodique et quasi bruit eux **[5].**

Un appareil 'le convertisseur Analogique Numérique 'se charge de convertir les tensions en chaînes de nombres binaires à chaque période de l'horloge d'échantillonnage. Les nombres binaires sont stockés sur un support d'enregistrement numérique sorte de mémoire (Figure 1.12)**.**

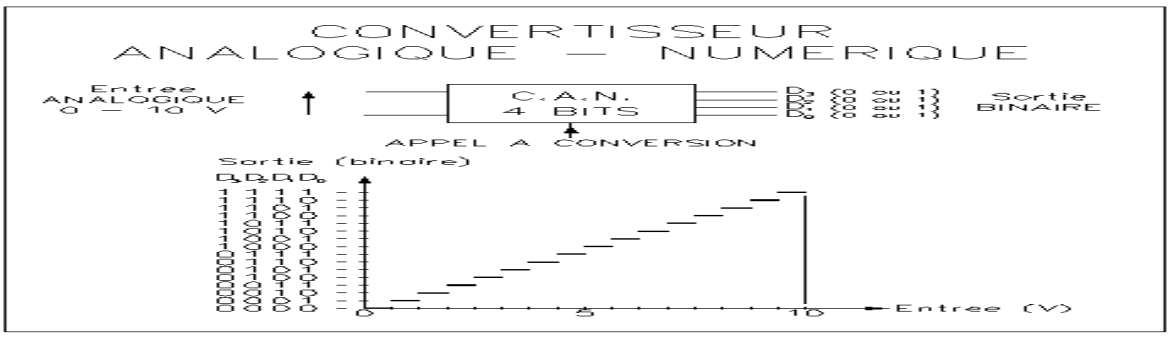


Figure 1.11: Conversion analogique numérique.

2.3.3. Méthodes d'analyse du signal vocal

L'arsenal des méthodes d'analyse et de traitement du signal est considérable. Nous présentons les méthodes générales couramment utilisées pour l'analyse du signal de parole, puis les méthodes utilisées pour l'extraction des paramètres.

Les méthodes d'analyse du signal de la parole peuvent être divisées en deux grandes classes, les méthodes paramétriques et les méthodes non paramétriques **[6].**

2.3.3.1. Les méthodes temporelles

Le signal de parole jouissant de quelques propriétés, exploitables à partir de la représentation temporelle. Les méthodes d'analyses temporelles se basent essentiellement sur la mesure du max, du min, du nombre des passages par zéro,de la fonction d'auto corrélation,du calcul de l'énergie et autres.

Du fait que le signal vocal est considéré comme étant un signal quasi-stationnaire, le traitement doit se faire sur des tranches de 5 à 30ms.

2.3.3.2. Les méthodes d'analyse spectral

Les propriétés spectrales du signal vocal présentent un intérêt majeur pour la perception auditive. L'analyse spectrale est une technique qui met en évidence les caractéristiques fréquentielles des signaux. Elle est souvent utilisée dans les techniques d'analyse synthèse du signal vocal, notamment dans l'analyse par banc de filtre **[6].**

2.3.3.3. Les méthodes non paramétriques

Ces méthodes sont basées principalement sur le calcul de la transformée de Fourrier, soit sur le signal directe, soit sur sa fonction d'auto corrélation. Le calcul de la TF permet l'obtention de la densité spectrale de puissance, qui nous mène ainsi l'extraction des paramètres nécessaires à l'analyse et la synthèse du signal vocal **[6].**

2.3.3.4. Les méthodes paramétriques

Les techniques basées sur l'analyse spectrale présentent quelques limitations, liées à l'hypothèse que le signal est nul au-delà de la fenêtre d'analyse. Pour remédier à ce problème, des méthodes paramétriques sont apparues. Parmi ces méthodes ont trouve les méthodes dites autorégressive **[6].**

2.3.3.5. Le codage Prédictif Linéaire (LPC)

C'est une méthode de type essentiellement temporel qui permet de calculer des coefficients appelés coefficients de la prédiction linéaire.

2.4. Méthodes d'extraction des paramètres

Ces méthodes consistent à extraire les paramètres essentiels qui caractérisent généralement le signal de parole à savoir l'énergie, la fréquence fondamentale et les formants.

2.4.1. Extraction de la fréquence fondamentale (pitch)

L'extraction de la fréquence fondamentale (ou pitch) est comme son nom l'indique fondamentale. Les variations de la fondamentale pour un locuteur donné constituent ce qu'on

appelle la prosodie. Celle-ci influe considérablement sur l'oreille humaine pour permettre la différentiation entre locuteurs et ainsi la reconnaissance du locuteur **[6].**

2.4.2. Extraction des formants

Le début d'utilisation des méthodes d'extraction des formants remonte à 1934. Ces formants sont les résonances du conduit vocal considéré comme un filtre et correspondant aux pôles de la fonction de transfert de ce dernier. Ce sont des paramètres privilégiés dans l'étude et l'analyse de la parole, ils apparaissent plus clairement pour les sons voisés **[6].**

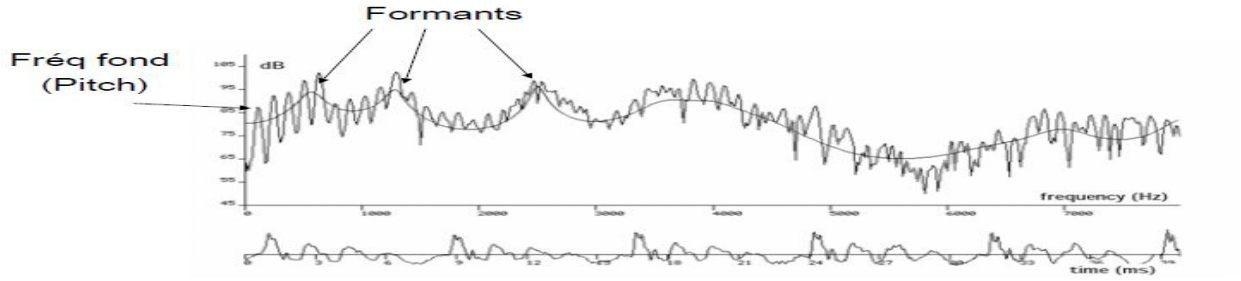


Figure 1.12 : Exemple d'extraction des paramètres.

2.5. Conclusion

Les méthodes illustrées si dessus sont destinées à l'analyse et l'extraction des paramètres à savoir l'énergie, la fréquence fondamentale et les formants. L'extraction de ces paramètres revêt une importance remarquable, notamment en synthèse et en reconnaissance de la parole. Les méthodes précédentes et beaucoup d'autres s'identifient à leurs avantages et leurs inconvénients propres. La qualité d'une méthode peut se chiffrer par les critères suivants :

- La précision de la méthode et sa fiabilité dans la détection du fondamentale et les formants pour différents locuteurs et pour larges gammes de sons.
- Le temps de calcul et les possibilités de câblage en temps réel.
- Le coût de l'implantation (câblage) relativement à son usage et à ses utilisateurs.
- La méthode que nous allons utiliser pour extraction de caractéristiques d'un signal vocal est Le codage prédictif linéaire LPC, car elle a prouvé son efficacité au niveau de l'identification et la classification de locuteurs **[5].**

3. Reconnaissance de la parole

3.1. Introduction

Notre Dieu tout puissant, nous a délégué un organisme biologique très complexe et très développé, notre espèce humaine est l'unique à être privilégiée de « la pensé », le message représentant cette dernière, en générale, peut prendre trois aspects, l'aspect écrit, l'aspect signé et celui verbal, en prenant le dernier aspect, la forme la plus simple qui le concrétise est **la parole.**

L'expression répandue 'ce ne sont que des paroles' banalise ce terme, cependant nous apercevons son véritable poids du point de vue phénomène à étudier, seulement chez les chercheurs de ce domaine, car la parole pour eux est un phénomène très complexe, non seulement en tenant compte de la difficulté du mécanisme interne qui la génère et celui qui la transmet, mais aussi de l'entrave de l'organisme qui la reconnaît.

Sans sa reconnaissance, la parole n'a aucun sens, car elle est produite pour être reconnue dans le but de transmettre une pensé précise afin de satisfaire un certain besoin.

Par l'esprit scientifique, de savoir, de développement et de création, les chercheurs optent pour dépasser la compréhension de la communication entre Hommes, en se dirigeant vers un nouvel horizon qui englobe la reconnaissance automatique de la parole.

La reconnaissance automatique de la parole est la manière évoluée pour établir un dialogue artificiel « Homme-Machine », dans le but d'adapter une machine à un vocabulaire limité, qui traduit un besoin issu d'un locuteur.

Le domaine de cette application peut atteindre plusieurs domaines tels que : le pilotage d'avion, la composition du numéro téléphonique du correspondant, faire acquérir des informations à un PC, différentes aides à des handicapés …etc.

Ce domaine fait l'objectif des chercheurs depuis de longues années, par conséquent un bon nombre de méthodes sont incorporées, telles que les méthodes globales, analytiques, probabilistes et encore les méthodes connexionnistes qui sont adoptées depuis les années quarante **[6].**

3.2. Définition

La reconnaissance automatique de la parole est l'un des deux domaines du traitement automatique de la parole, l'autre étant la synthèse vocale.

La reconnaissance automatique de la parole permet à la machine de comprendre et de traiter des informations fournies oralement par un utilisateur humain. Elle consiste à employer des techniques d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots mais aussi, plus récemment, de phonèmes (unité sonore minimale). En revanche, le système de synthèse de la parole permet de reproduire d'une manière sonore un texte qui lui est soumis, comme un humain le ferait.

Ces deux domaines et notamment la reconnaissance vocale, font appel aux connaissances de plusieurs sciences : l'anatomie (les fonctions de l'appareil phonatoire et de l'oreille), les signaux émis par la parole, la phonétique, le traitement du signal, la linguistique, l'informatique, l'intelligence artificielle et les statistiques. Il faut bien distinguer ces deux mondes : un système de synthèse vocale peut très bien fonctionner sans qu'un module de reconnaissance n'y soit rattaché. Evidemment le contraire est également tout à fait possible. Par contre, dans certains

domaines bien précis, l'un ne va pas sans l'autre. Il est bien entendu que l'étude se portant sur la reconnaissance automatique de la parole, l'autre aspect du traitement de la parole.

Le traitement automatique de la parole ouvre des perspectives de nouveaux comptes tenus de la différence considérable existant entre la commande manuelle et vocale.

L'utilisation du langage naturel dans le dialogue personne/machine met la technologie à la portée de tous et entraîne sa vulgarisation, en réduisant les contraintes de l'usage des claviers, souris et codes de commandes à maîtriser. En simplifiant le protocole de dialogue personne/machine, le traitement automatique de la parole vise donc aussi un gain de productivité puisque c'est la machine qui s'adapte à l'homme pour communiquer, et non l'inverse. De plus, il rend possible l'utilisation simultanée des yeux ou des mains à une autre tâche.
Il permet d'humaniser les systèmes informatiques de gestion de l'information, en axant leur conception sur les utilisateurs **[3].**

Figure 1.13 : Processus de reconnaissance de la parole.

3.3. Principe de fonctionnement

3.3.1. Problématique

Pour bien appréhender le problème de la reconnaissance automatique de la parole, il est bon d'en comprendre les différents niveaux de complexité et les différents facteurs qui en font un problème difficile. Le système doit-il être optimisé pour un unique locuteur ou est-il destiné à devoir se confronter à plusieurs utilisateurs ?

On peut aisément comprendre que les systèmes dépendants d'un seul locuteur sont plus faciles à développer et sont caractérisés par de meilleurs taux de reconnaissance que les systèmes indépendants du locuteur étant donné que la variabilité du signal de parole est plus limitée.

Cette dépendance au locuteur est cependant acquise au prix d'un entraînement spécifique à chaque utilisateur. Ceci n'est néanmoins pas toujours possible. Par exemple, dans le cas d'applications téléphoniques, on comprend bien que les systèmes puissent être utilisés par n'importe qui et donc être indépendants du locuteur. Bien que la méthodologie de base reste la même, cette indépendance au locuteur est obtenue par l'acquisition de nombreux locuteurs (couvrant si possible les différents dialectes) qui sont utilisés simultanément pour l'entraînement de modèles susceptibles d'en extraire toutes les caractéristiques majeures. Une solution intermédiaire parfois utilisée consiste à développer des systèmes capables de s'adapter rapidement (de façon supervisée ou non) au nouveau locuteur. Le système est-il robuste ?

Autrement dit, le système est-il capable de fonctionner proprement dans des conditions difficiles? En effet, de nombreuses variables pouvant affecter significativement les performances des systèmes de reconnaissance ont été identifiées :

Bruits d'environnement (dans une rue, un bistrot, etc.…).

- Déformation de la voix par l'environnement (réverbérations, échos, etc.…).
- Qualité du matériel utilisé (micro, carte son, etc.…).
- Bande passante fréquentielle limitée (fréquence limitée d'une ligne téléphonique).
- Elocution inhabituelle ou altérée (stress, émotions, fatigue, etc.…) **[6].**

3.3.2. Fonctionnement

Le problème de la reconnaissance automatique de la parole consiste à extraire l'information contenue dans un signal de parole, typiquement par échantillonnage du signal électrique obtenu à la sortie d'un microphone, afin qu'il puisse être comparé à des modèles sous forme numérique. Parmi plusieurs techniques de reconnaissance, il y en a deux qui sont majoritairement utilisées afin de parvenir à résoudre ce problème : la comparaison à des exemples et la comparaison d'unités de parole **[6].**

3.3.2.1. Reconnaissance par comparaison à des exemples

Les premiers succès en reconnaissance vocale ont été obtenus dans les années 70 à l'aide d'un paradigme de reconnaissance de mots. L'idée, très simple dans son principe, consiste à faire prononcer un ou plusieurs exemples de chacun des mots susceptibles d'être reconnus, et à les enregistrer sous forme de vecteurs acoustiques (représentation numérique du signal sonore).

Puisque cette suite de vecteurs acoustiques caractérise complètement l'évolution de l'enveloppe spectrale du signal enregistré, on peut dire qu'elle correspond à l'enregistrement d'un spectrogramme.
L'étape de reconnaissance proprement dite consiste alors à analyser le signal inconnu sous la forme d'une suite de vecteurs acoustiques similaires, et à comparer la suite inconnue à chacune des suites des exemples préalablement enregistrés. Le mot «reconnu» sera alors celui dont la suite de vecteurs acoustiques «spectrogramme» colle le mieux à celle du mot inconnu. Il s'agit en quelque sorte de voir dans quelle mesure les spectrogrammes se superposent **[6].**

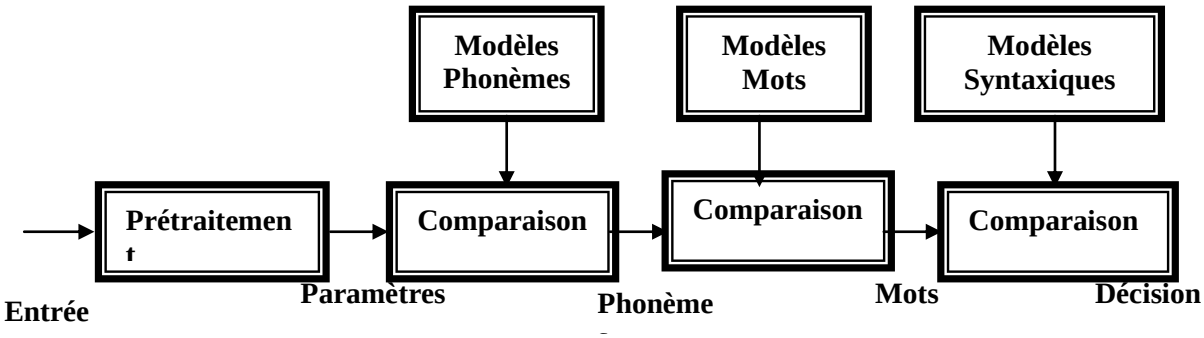


Figure 1.14 : Schéma synoptique d'un système de reconnaissance de parole selon une approche comparaison.

3.3.2.2. Reconnaissance par modélisation d'unités de parole

La plupart des systèmes de reconnaissance de la parole sont de nos jours basés sur ce mode là. Dés que l'on cherche à concevoir un système réellement multi locuteur, à plus grand vocabulaire et s'adaptant facilement à une application, il devient nécessaire de mener la reconnaissance sur base d'unités de parole de plus petite taille, que l'on appelle phonèmes.

En effet, la parole est constituée d'une suite de sons élémentaires : «a», «é», «ss». Ils sont produits par la vibration des cordes vocales. Ces sons mis bout à bout composent des mots. On ne se contente plus alors d'exemples de ces unités, mais on cherche plutôt à en déduire un modèle (un modèle par unité), qui sera applicable pour n'importe quelle voix. Il apparaît ainsi dans de nombreuses publications que l'on peut décomposer la reconnaissance de la parole en quatre modules **[7].**

Un module d'acquisition et de modélisation du signal qui transforme le signal de parole en une séquence de vecteurs acoustiques. Pour être utilisable par un ordinateur, un signal doit tout d'abord être numérisé. Cette opération tend à transformer un phénomène temporel analogique, le signal sonore dans notre cas, en une suite d'éléments discrets, les échantillons. Ceux-ci sont obtenus avec une carte spécialisée courante de nos jours dans les ordinateurs depuis l'avènement du multimédia. La numérisation sonore repose sur deux paramètres : la quantification et la fréquence d'échantillonnage.

La quantification définit le nombre de bits sur lesquels on veut réaliser la numérisation. Elle permet de mesurer l'amplitude de l'onde sonore à chaque pas de l'échantillonnage. Le choix de la fréquence d'échantillonnage est aussi déterminant pour la définition de la bande passante représentée dans le signal numérisé.

Un module acoustique qui peut produire une ou plusieurs hypothèses phonétiques pour chaque segment de parole (par exemple de 10 ms, pour chaque vecteur acoustique), associées en général à une probabilité. Ce générateur d'hypothèses est généralement basé sur des modèles statistiques de phonèmes, qui sont entraînés sur une grande quantité de données de parole (par exemple, enregistrement de nombreuses phrases) contenant plusieurs fois les différentes unités de parole dans plusieurs contextes différents. Ces modèles statistiques sont le plus souvent constitués de lois statistiques paramétriques dont on ajuste les paramètres pour «coller» au mieux aux données ou de réseaux de neurones artificiels.

Un module lexical dans le cadre de la reconnaissance de la parole continue, même si le système acoustique est basé sur des phonèmes, il faut obtenir, pour chaque entrée du dictionnaire phonétique, un modèle qui lui est propre. Un tel module lexical embarque en général des modèles des mots de la langue (les modèles de base étant de simples dictionnaires phonétiques, les plus complexes sont de véritables automates probabilistes, capables d'associer une probabilité à chaque prononciation possible d'un mot). A l'issue de ce module, il peut donc y avoir plusieurs hypothèses de mots qui ne pourront être départagées que par les contraintes syntaxiques.

Un module syntaxique qui interagit avec un système d'alignement temporel pour forcer la reconnaissance à intégrer des contraintes syntaxiques, voire sémantiques. Les connaissances syntaxiques sont généralement formalisées dans un modèle de la langue, qui associe une probabilité à toute suite de mots présents dans le lexique. Ainsi le système est capable de choisir entre plusieurs mots selon le contexte de la phrase ou du texte en cours, et de son modèle lexical.On peut ajouter à cela un module de filtrage pouvant corriger le signal après l'acquisition afin de retirer les distorsions ou les bruits provenant du matériel ou de l'environnement du locuteur. Ce module est aussi appelé «traitement du canal de transmission». Du fait de sa complexité et du peu d'amélioration qu'il apporte, ce module n'est pas toujours intégré aux systèmes. Cependant la recherche de meilleurs traitements du canal de transmission sera sûrement nécessaire à l'amélioration des systèmes de reconnaissance vocale.

3.4. Reconnaissance de petits vocabulaires

Ça concerne la reconnaissance de mots isolés, multi locuteurs dans des conditions difficiles, par exemple : reconnaissance de chiffres à travers le réseau téléphonique **[6].**

3.5. Reconnaissance de petits vocabulaires de mots isolés

La reconnaissance de mots isolés, le plus souvent mono locuteur, pour des vocabulaires de quelques dizaines jusqu'à quelques centaines de mots est un problème assez bien résolu. Les premiers systèmes commerciaux de cette catégorie sont apparus il y a un peu plus de vingt ans **[3].**

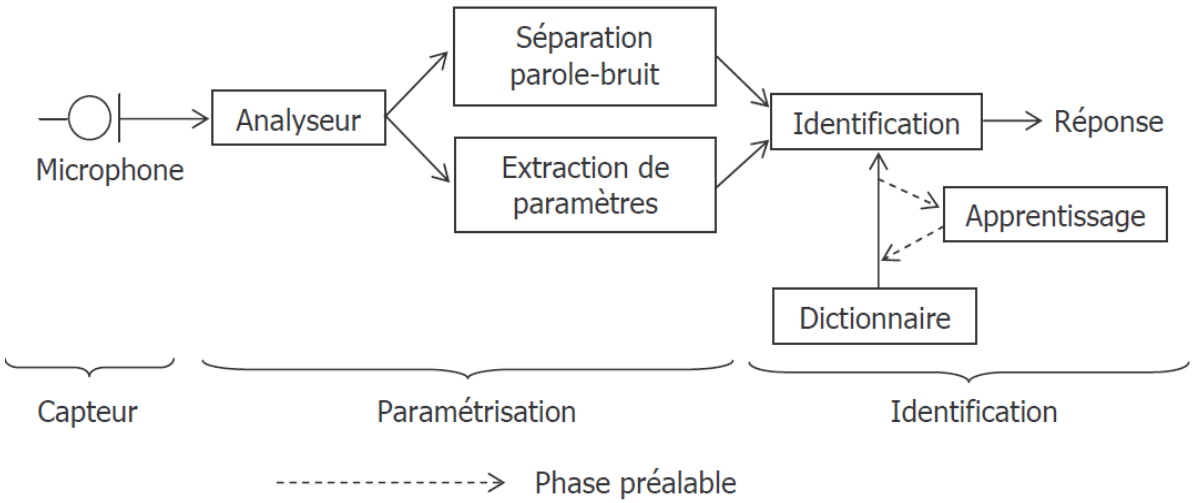


Figure 1.15 : Système de reconnaissance de mots isolés **[7]**.

3.6. Reconnaissance de grands vocabulaires

Par exemple par IBM, Kurzeil et Dragon systèmes mono locuteur en particulier pour des tâches de dictée de textes dans des domaines d'application fixés .Des systèmes de ce type sont présentés, fondés sur une modélisation stochastique de la parole, méthode actuellement la plus

performante. Dans cette catégorie apparaissent aussi des systèmes utilisant une reconnaissance phonétique des mots, c'est notamment le cas d'un produit de speech system **[3].**

Microsoft, en passant par Apple et IBM, de nombreux industriels travaillent sur des projets de reconnaissance vocale, généralement en complément d'une activité de recherche sur la synthèse de la parole, le tout s'insérant dans des projets plus généraux d'interface Homme Machine.

Il faudra attendre encore plus longtemps avant que la machine remplace purement et simplement la secrétaire dactylo pour la saisie de textes sur ordinateur. Les systèmes de reconnaissance vocale actuels sont encore bien trop grossiers pour comprendre toutes les finesses qui peuvent se glisser dans la **syntaxe** et dans les intonations de la langue parlée en continu et non plus sous la forme de mots clés ou de petites phrases sommaires **[3].**

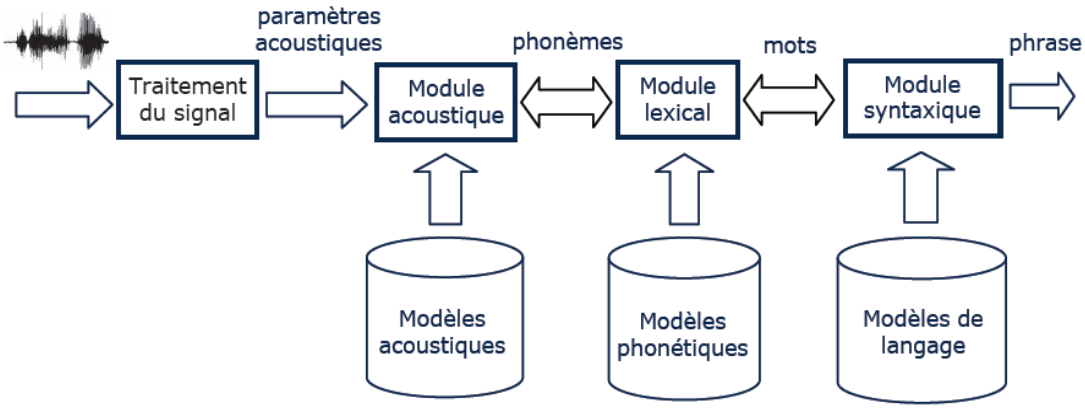


Figure 1.16 : Système de reconnaissance de grands vocabulaires.

3.7. Caractéristiques du système de reconnaissance de la parole

On peut caractériser un tel système par **:**

3.7.1. Le mode de fonctionnement

La reconnaissance de la parole pour la voix d'un seul locuteur est déjà en elle-même un problème élémentaire, en raison de la variabilité intra locuteur, inhérent au processus humain de production. Cette variation est liée aux différences du débit d'élocution, à l'émotion, au stress, aux rhumes, … etc **[3].**

3.7.2. Le mode d'élocution

Si le locuteur marque une pause après chaque mot de l'énoncé, la complexité du problème est réduite, puisque les frontières du mot sont alors disponibles (contrairement au cas de la parole continue).
Dans ce cas on parle des mots isolés. Les phrases et les mots enchaînés présentent d'autres genres de modes d'élocution **[3].**

3.7.3. La taille du vocabulaire

Il est plus facile de travailler sur un vocabulaire très étendu (quelques milliers ou dizaines de milliers de mots) que sur vocabulaire très restreint (quelques dizaines de mots).

La taille du vocabulaire est cependant un paramètre insuffisant, un ensemble de mots très différents les uns des autres étant plus faciles à traiter que des mots proches phonétiquement.

3.7.4. Le langage

La prise en compte de la syntaxe du langage produit par l'utilisateur sera plus facile pour un langage ' rigide', très contraint, que si toute la souplesse de la langue naturelle palée peut être rencontrée.

3.7.5. Le mode de décodage de l'information

Ce mode de décodage de l'information est lié aux types d'approches utilisées pour le traitement du signal :

3.7.5.1. Approche analytique

Cette approche utilise tout d'abord une segmentation à priori du signal en unités de tailles phonétiques, puis chacun des segments est identifié en comparant les mesures acoustiques à des formes de reconnaissance **[5].**

3.7.5.2. Approche globale

Approche globale ou pragmatique consiste à faire totalement abstraction des phonèmes linguistiques pour ne retenir que l'aspect acoustique de la parole.

Elle applique des hypothèses simplificatrices (au risque de dénaturer la communication vocale) au problème de façon à le rendre plus abordable **[5].**

3.7.6. L'environnement

Indépendamment de ce qui précède, l'environnement acoustique et les conditions de prise du son, constituent un facteur important : la présence du bruit, même stationnaire, dégrade en général fortement les performances des systèmes de reconnaissance. De plus si ce bruit est intense, il induit une augmentation de la variabilité chez le locuteur **[3].**

3.8. Reconnaissance de la parole continue

Tout d'abord, qu'est ce que la parole continue ? C'est un discours, des phrases où les mots s'enchaînent sans moyen de séparer, contrairement aux mots isolés. Le but de cette partie n'est pas de rentrer dans les détails de la programmation d'un logiciel de reconnaissance de la parole

continue, cela serait trop long et fastidieux. On va donc présenter les " ficelles " de la reconnaissance de la parole continue de manière très générale.

Les objectifs de cette partie étant donc éclaircis, on peut entamer la réflexion autour de la reconnaissance de la parole continue. Pourquoi, après tout, s'évertuer à attribuer à une machine de telles capacités ? Est-ce par pure fantaisie que les auteurs de science-fiction inventent des dialogues entre un héros et sa machine ? Non, ceci relève d'un besoin qui pourrait se résumer à une chose : la recherche d'un confort et d'amélioration de l'interaction de l'homme avec la machine. Les avantages d'un tel progrès sont simples à imaginer **[6].**

3.9. Quelques applications

De façon générale, le choix d'une application doit faire l'objet d'une étude attentive, fondée sur un ensemble de critères objectifs. En particulier, il est important d'examiner si la voix apporte véritablement un accroissement des performances ou un meilleur confort d'utilisation. Par ailleurs, il ne faut pas trop attendre de la commande vocale mais la considérer, en tout état de cause, comme un moyen complémentaire parmi d'autres moyens d'interaction Homme-Machine plus traditionnels**.** Bien entendu, à chaque type d'application correspondent des critères de performance différents. Ainsi, pour des applications en reconnaissance de la parole, on jugera la qualité d'une application sur les quatre critères principaux suivants :

- Le débit du flux de parole correctement reconnu. Si le locuteur prononce les mots séparément avec de petites pauses (environ 200 ms) entre chaque mot, on parlera de reconnaissance par mots isolés, sinon ce sera de la reconnaissance de parole continue.
- La taille du vocabulaire correctement reconnu. Ce vocabulaire variera de quelques mots (la cabine téléphonique à entrée vocale) à plusieurs milliers de mots (la machine à écrire à entrée vocale).
- Les contraintes imposées par le système sur l'environnement de fonctionnement : acceptation de bruits de fond et parasites divers. Des critères de qualité positifs dans certaines applications peuvent être négatifs dans d'autres : l'indifférence au locuteur est recherchée pour une cabine téléphonique à numérotation vocale alors qu'au contraire c'est la capacité de discrimination entre locuteurs qui déterminera la qualité d'une serrure à commande vocale.
- Les contraintes imposées par le système sur l'utilisateur : est-il unique ou multiple, doit-il s'astreindre à une phase d'apprentissage préalable ? **[7].**

3.9.1. Services vocaux

Les serveurs passifs (sans reconnaissance vocale) existent depuis de nombreuses années tels que l'horloge parlante, la météo, les résultats des courses, du loto, etc.… Mais lorsque la quantité d'information est importante, il devient nécessaire pour l'utilisateur de pouvoir sélectionner ce qu'il veut entendre. Dans des cas simples la sélection de touches «multifréquences» (DTMF) peut suffire. Mais pour des applications plus complexes des systèmes sont en voie de développement pour que l'utilisateur puisse naviguer sur un serveur vocal en prononçant les

mots de contrôle de l'application. Ces services pourront s'étendre à tout un ensemble de domaines : la réservation de place d'avion, de train, de théâtre, de chambre d'hôtel, les déclarations de sinistre à l'assureur, les consultations et transactions bancaires, les opérations boursières, la facturation automatique des appels à distance, etc.…**[3].**

3.9.2. Contrôle de qualité, saisie des données

Dans de nombreux environnements de travail la possibilité de décharger le travailleur, grâce à une interface vocale, apporte un gain incontestable de liberté et de rapidité de mouvement. Pendant qu'il observe un processus complexe, il peut par exemple décrire des informations visuelles. Il a aussi la possibilité de commander à distance un automate évoluant en milieu hostile (apesanteur, sous-marin, industrie pétrolière) **[5].**

3.9.3. Avionique

A bord des avions, les tâches étant de plus en plus complexes et le tableau de bord de plus en plus réduit, la parole permet au pilote d'avoir à sa disposition un moyen supplémentaire d'interaction avec la machine, sans cependant gêner l'accomplissement des tâches courantes qui requièrent de sa part toute son attention visuelle. Les autorités canadiennes ont été les précurseurs des techniques vocales dans l'avionique. Ainsi, l'Institut de Recherche Aérospatiale (IRA) a effectué des travaux de recherche sur la technologie vocale depuis la fin des années 70.
Actuellement, un système de reconnaissance de la parole est à bord du dernier des avions de l'armée française, le Rafale, capable de traiter un vocabulaire de près de 300 mots **[5].**

3.9.4. Formation

Les enfants, mais aussi les adultes, sont attirés par des jeux doués de parole (poupées qui parlent, jeux de société, jeux vidéo, jeux éducatifs). L'enseignement assisté par ordinateur et notamment les laboratoires de langue commence à intégrer de plus en plus de possibilités vocales, et évolue vers une interactivité plus grande : les systèmes d'aide à l'apprentissage des langues étrangères, permettant d'acquérir une prononciation correcte, une maîtrise du vocabulaire et de la syntaxe, ne peuvent que bénéficier des technologies vocales qui leur confèrent en outre un aspect ludique.
Du côté des applications proprement dites, la société Auralog a été précurseur avec ses logiciels d'apprentissage de langues (TeLL me More, Atout Clic Anglais). Ainsi, grâce à la technologie avancée de la reconnaissance vocale, l'utilisateur engage un véritable dialogue avec son PC. Suivant son niveau, l'apprenant paramètre la reconnaissance vocale pour la rendre plus tolérante ou plus exigeante quant à la qualité de sa prononciation. L'utilisateur s'entraîne à prononcer une phrase ou un mot et obtient un score lui permettant d'évaluer la qualité de son accent, de sa prononciation et de son intonation **[7].**

3.9.5. Aide aux handicapés

Différents programmes européens ont permis de mieux cerner les différents types d'handicap dont souffre la population, ainsi que le nombre de personnes concernées. On dénombre actuellement en Europe 12 millions de malvoyants dont 1 million de non-voyants, 81 millions de

malentendants, dont 1 million de non entendants, environ 30 millions de personnes ayant un handicap moteur des membres supérieurs et 50 millions ayant un handicap des membres inférieurs. L'intérêt des technologies vocales apparaît évident dans la mesure où celles-ci permettent aux personnes handicapées de retrouver une certaine autonomie et de bénéficier d'une meilleure insertion dans leur environnement tant professionnel que familial, la parole se substituant au sens défaillant.
Ainsi beaucoup de systèmes existent pour cela, tel que le contrôle de fauteuil roulant, le contrôle de fonctions secondaires dans la voiture, le contrôle d'appareil électrique à la maison, le contrôle de l'ordinateur, …etc.**[7].**

3.9.6. Dictée vocale

L'orientation actuelle des logiciels tend de plus en plus à offrir un contrôle total de l'environnement permettant de se passer du clavier et de la souris pour utiliser l'ordinateur. Les nouveaux systèmes d'exploitation couplés aux logiciels à venir devraient enfin permettre d'offrir un ordinateur fonctionnant réellement «sans les mains» **[7].**

3.9.7. Relation avec la télécommunication

Dans le secteur de la téléphonie, les grandes sociétés de télécommunication ont engagé une course à l'innovation. Ainsi, il suffit de dire le nom du correspondant désiré dans le récepteur, à condition de l'avoir préalablement encodé, pour obtenir la communication souhaitée. Ceci peut-être très utile pour téléphoner depuis une voiture.

L'information au public est aussi un domaine concerné par la numérisation de la parole. Dans les gares ou les aéroports, par exemple, on pourra bientôt voir des bornes interactives qui remplaceront les agents préposés aux renseignements. Pour connaître l'horaire d'un train, il suffira de demander de vive voix à la machine où on veut aller et quand, et elle répondra dans la langue de notre choix, avant de nous souhaiter un agréable voyage.

Plus précisément, aujourd'hui, deux gammes de services dominent le marché des services de Télécommunication à commande vocale : ce sont les services à opérateurs partiellement automatisés et les services de répertoires vocaux, évoluant progressivement vers des services plus complets d'assistants téléphoniques **[3].**

3.9.8. Et aussi …

On peut aussi citer les modules de reconnaissance vocale embarqués, comme dans les téléphones mobiles ou les assistants numériques ainsi que les futures possibilités en cours de développement chez les fabricants automobile avec le contrôle de différents éléments de la voiture grand public : autoradio, climatisation, navigation de bord. On peut même de nos jours surfer sur Internet grâce à des commandes vocales, c'est ce que propose la société Interactive Speech. L'utilisation de la reconnaissance automatique de la parole devient courante et devrait très bientôt apparaître dans la plupart des domaines d'activités et la plupart des applications futures **[3].**

Conclusion

Au terme de ce bilan rapide sur la reconnaissance vocale, on a pu constater que ce domaine est particulièrement vaste et qu'il n'existe pas de produit miracle capable de répondre à toutes les applications. Le bruit, par exemple, non traité par ce document, reste un frein à la généralisation des systèmes de reconnaissance.

La reconnaissance vocale reste un compromis entre la taille du vocabulaire, ses possibilités multi locuteur, son encombrement physique, sa rapidité, temps d'apprentissage, et...

La puissance des outils de calcul actuels et les capacités d'intégration des systèmes ont provoqué un regain d'intérêt depuis ces dernières années chez les industriels. En effet, ces derniers voient dans la reconnaissance vocale, « le plus commercial »", permettant de faire la différence avec la concurrence.

Chapitre 2

Support Vector Machine

- **Introduction**
- **Méthodes de classification**
- **Apprentissage statistique et SVM**
- **SVM principe de fonctionnement général**
- **Fondements mathématiques**
- **SVMs et analyse des bases de données**
- **Les domaines d'application**
- **Conclusion**

Introduction

Parmi les méthodes à noyaux, inspirées de la théorie statistique de l'apprentissage de Vladimir Vapnik, les SVM constituent la forme la plus connue. SVM est une méthode de classification binaire par apprentissage supervisé, elle fut introduite par Vapnik en 1995. Cette méthode est donc une alternative récente pour la classification. Cette méthode repose sur l'existence d'un classificateur linéaire dans un espace approprié. Puisque c'est un problème de classification à deux classes, cette méthode fait appel à un jeu de données d'apprentissage pour apprendre les paramètres du modèle. Elle est basée sur l'utilisation de fonctions dites noyau (kernel) qui permettent une séparation optimale des données. Dans la présentation des principes de fonctionnements, nous schématiserons les données par des « points » dans un plan.

La notion d'apprentissage étant importante, nous allons commencer par effectuer un rappel. L'apprentissage par induction permet d'arriver à des conclusions par l'examen d'exemples particuliers. Il se divise en apprentissage supervisé et non supervisé. Le cas qui concerne les SVM est l'apprentissage supervisé. Les exemples particuliers sont représentés par un ensemble de couples d'entrée/sortie. Le but est d'apprendre une fonction qui correspond aux exemples vus et qui prédit les sorties pour les entrées qui n'ont pas encore été vues. Les entrées peuvent être des descriptions d'objets et les sorties la classe des objets donnés en entrée.

1. Méthodes de classification

Parmi les méthodes de classification, deux sont particulièrement classiques dans ce domaine : la méthode des plus proches voisins (*k-ppv*) et les arbres de décision et une méthode moderne appeler SVM.

1.1. K-ppv

La méthode des K plus proches voisins (k-ppv) est une méthode particulièrement élémentaire. Comme son nom l'indique elle consiste à rechercher dans la base d'apprentissage les k individus qui sont les plus proches d'une nouvelle donnée, et la règle de décision consiste à faire un vote majoritaire sur les classes de ces *k-ppv*. On peut noter qu'il est préférable de prendre k impair pour ne pas avoir de problèmes d'égalité lors de la prise de décision. La méthode repose donc sur un critère de similarité qu'il faut définir a priori pour comparer les données. Le seul paramètre à régler est alors le nombre de voisins à considérer. Cette méthode peut paraître élémentaire mais dans de nombreux cas réels elle s'avère efficace, et même plus performante que des modèles plus complexes. Elle peut par conséquent constituer une bonne référence pour quantifier les performances de classification d'autres méthodes **[12]**.

1.2. Arbres de décision

Les arbres de décision représentent une méthode très efficace d'apprentissage supervisé. Il s'agit de partitionner un ensemble de données en des groupes les plus homogènes possible du point de vue de la variable à prédire. On prend en entrée un ensemble de données classées, et on fournit

en sortie un arbre qui ressemble beaucoup à un diagramme d'orientation où chaque nœud final (feuille) représente une décision (une classe) et chaque nœud non final (interne) représente un test. Chaque feuille représente la décision d'appartenance à une classe des données vérifiant tous les tests du chemin menant de la racine à cette feuille [21].
Les arbres de décisions sont très répandus, à cause de la simplicité de lecture de leurs résultats et leur traitement naturels des cas multi classe. Néanmoins, ils posent beaucoup de problèmes tel que :

- La difficulté de manipulation des attributs numériques.
- L'espace nécessaire pour leur déduction.

1.3. Machines à vecteurs de support (SVM)

L'algorithme des machines à vecteurs de support a été développé dans les années 90 par Vapnik. Il a initialement été développé comme un algorithme de classification binaire supervisée. Il s'avère particulièrement efficace de par le fait qu'il peut traiter des problèmes mettant en jeu de grands nombres de descripteurs, qu'il assure une solution unique (pas de problèmes de minimum local comme pour les réseaux de neurones) et il a fourni de bons résultats sur des problèmes réels L'algorithme sous sa forme initiale revient à chercher une frontière de décision linéaire entre deux classes, mais ce modèle peut considérablement être enrichi en se projetant dans un autre espace permettant d'augmenter la séparabilité des données. On peut alors appliquer le même algorithme dans ce nouvel espace, ce qui se traduit par une frontière de décision non linéaire dans l'espace initial.

Nous reviendrons à cette méthode dans ce chapitre puisque cette approche fait précisément l'objet du sujet **[12]**.

2. Apprentissage statistique et SVM

2.1. Objectif de l'apprentissage statistique

Effectuer une classification consiste à déterminer une règle de décision capable, à partir d'observations externes, d'assigner un objet à une classe parmi plusieurs. Le cas le plus simple consiste à discriminer deux classes. D'une manière plus formelle, la classification bi-classe revient à estimer une fonction $f : x \rightarrow \{+1,-1\}$ à partir d'un ensemble d'apprentissage constitué de couples (xi, yi), qu'on suppose i.i.d. suivant une distribution de probabilité P(x, y) inconnue, tels que

$$(xi, yi) \in X \times Y \text{ où } i = 1,\ldots, N \text{ x et } Y = \{+1,-1\},$$

de sorte à ce que f classifie correctement des exemples inconnus(xt, yt). Par exemple, on peut assigner xt à la classe (+1) si $f(xt) \geq 0$, et à la classe (-1) sinon. Les exemples inconnus sont supposés suivre la même distribution de probabilité P(x, y) que ceux de l'ensemble d'apprentissage. La meilleure fonction f est celle obtenue en minimisant le risque :

$$R[f] = \int L\,[f(x), y]\, d\,P(x, y). \qquad (2.1)$$

Où L désigne une fonction de coût, comme par exemple :

$$L\,[f(x),y] = (f(x)-y)^2$$

Malheureusement, le risque (2.1) ne peut être directement minimisé dans la mesure où la distribution de probabilité sous-jacente P(x, y) est inconnue. Aussi, on va chercher une fonction de décision proche de celle optimale à partir de dont on dispose, c'est-à-dire l'ensemble d'apprentissage et la classe de fonctions F est à laquelle la solution f appartient. Pour ce faire, on approxime le minimum du risque théorique par le minimum du risque empirique qui s'écrit :

$$R_{emp}[f]=\frac{1}{Nx}\sum_{i=1}^{Nx} L\,[f(x_i),y_i]\ . \qquad (2.2)$$

Il est possible de donner des conditions au classifieur pour qu'asymptotiquement (si $Nx\rightarrow \infty$), le risque empirique (2.2) converge vers le risque (2.1). Cependant, si on dispose de peu d'exemples pour faire l'apprentissage (i.e Nx petit), on s'expose au risque de sur-apprentissage (Figure 2.1). Pour éviter le sur-apprentissage, on peut restreindre la complexité de la classe F à laquelle appartient *f*. Intuitivement, une fonction de décision simple (la classe la plus simple se constituant des fonctions linéaires) capable de discriminer correctement les données est préférable à une fonction complexe. Pour cela, on introduit un terme de régularisation pour limiter la complexité des fonctions de F.

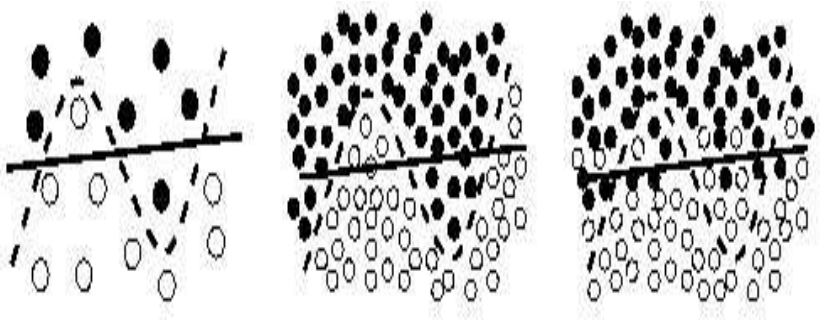


Figure 2.1 : Illustration du problème de sur apprentissage.

Etant donné un petit ensemble d'apprentissage (schéma de gauche), deux frontières de discrimination (représentées par les lignes continue et discontinue) sont possibles. La ligne discontinue est plus complexe mais minimise davantage le risque empirique. Seul un ensemble d'exemples plus grand permet de déterminer la meilleure des deux frontières de décision. S'il s'agit de la ligne discontinue, alors la ligne continue n'est pas suffisamment discriminante (schéma du milieu) ; s'il s'agit de la ligne continue, alors la ligne discontinue ne convient pas et caractérise un sur apprentissage (schéma de droite) **[12]**.

2.2. Théorie de Vapnik-Chervonenkis

Une manière de contrôler la complexité d'une classe de fonctions est donnée par la théorie de Vapnik-Chervonenkis (VC) et le principe de minimisation du risque structurel. Ici, le concept de

complexité de la fonction de décision f s'exprime par la dimension de VC (notée h) de la classe de fonctions F à laquelle appartient f . Grossièrement, la dimension de VC mesure combien d'échantillons de l'ensemble d'apprentissage peuvent être séparés par toutes les classifications possibles issues des fonctions de la classe **[8]**.

Considérons une famille imbriquée de classes de fonctions

$$F_1 \subset F_2 \subset \ldots \subset F_k ;$$

avec une dimension de VC non-décroissante, et $f_1 \ldots f_k$ les fonctions minimisant le risque empirique dans chacune de ces classes.

La minimisation du risque structurel consiste à choisir la classe F_i (et la fonction $f_{i)}$ de sorte à ce qu'une borne supérieure de l'erreur de généralisation puisse être minimisée (grâce, par exemple, au théorème suivant) **[8]**.

Théorème 1 : Soient h la dimension de VC de la classe de fonctions F, Remp[f] le risque empirique défini par (2.2) avec la fonction perte 0/1 (i.e. L [f(xi),yi] =H(-yf(x)) Où H désigne la fonction de Heaviside). Pour tout $\delta>0$ et $f \in F$, l'inégalité bornant le risque

$$R[f] = R_{emp}[f] + \sqrt{\frac{h(\ln \frac{2Nx}{h} + 1) - \ln(\frac{\delta}{4})}{Nx}} \qquad (2,3)$$

est vraie avec une probabilité de moins $(1- \delta)$pour $Nx>h$ *[12]*.

Cette borne n'est qu'un exemple et des formulations du même type ont été démontrées pour d'autres fonctions perte et d'autre mesures de complexité. Le but recherché ici est de minimiser l'erreur de généralisation R[f] en obtenant un faible risque empirique Remp [f] tout en gardant la plus petite classe de fonctions possible.

L'inégalité (2.3) fait apparaître deux cas extrêmes:

- une très petite classe de fonctions (par exemple F_1) fait décroître rapidement le terme de complexité (celui en racine carrée), mais le risque empirique demeure grand,
- une très grande classe de fonctions (par exemple F_k) implique un risque empirique petit, mais le terme de complexité explose.

La meilleure classe de fonctions est généralement intermédiaire entre la plus petite et la plus grande, puisque l'on cherche une fonction qui explique au mieux les données tout en préservant un faible risque empirique (Figure 2.2) **[12].**

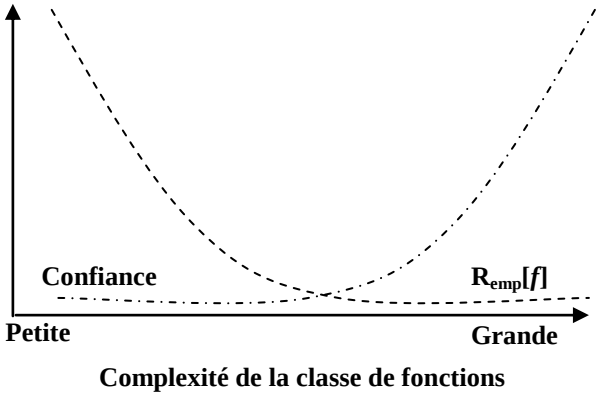


Figure 2.2 : Illustration de l'inégalité (2.3).

La courbe croissante, appelée confiance, correspond à la borne supérieure du terme de complexité. Les comportements du terme de complexité et de l'erreur empirique sont clairement opposés. On recherche donc le meilleur compromis entre complexité et erreur empirique **[12]**.

2.3. Marge et dimension de VC

Supposons pour l'instant que les échantillons de l'ensemble d'apprentissage sont séparables par un hyperplan (Figure 2.3), i.e on choisit des fonctions de décision de la forme :

$f(x)=\langle w, x\rangle +b.$ (2.4)

La marge est la distance minimale entre les échantillons de l'ensemble d'apprentissage et la frontière de décision.

Il a été montré que pour la classe des hyperplans, la dimension de VC peut être bornée en fonction de la marge. La marge peut à son tour être mesurée grâce au vecteur poids *w*: puisque nous supposons que les échantillons sont séparables, on peut redéfinir *w* et *b* de sorte à ce que les échantillons *x* les plus proches de l'hyperplan satisfassent $|\langle w, x\rangle +b|=1$.

Considérons maintenant deux échantillons x_1 et x_2 de classes différentes telles qu'on ait $\langle w, x_1\rangle +b = +1$et $\langle w, x_2\rangle +b = -1$. La marge γ correspond alors à la distance entre x_1 et x_2 mesurée perpendiculairement à l'hyperplan :

$$\gamma=\langle w / ||w||, x_1 - x_2\rangle =2/||w|| ;$$

Les résultats liant la dimension de VC de la classe des hyperplans de séparation à la marge et à la longueur du vecteur poids *w* sont respectivement donnés par les inégalités suivantes :

Où R est le rayon de la plus petite boule englobant les données. Ainsi, en bornant la marge de la classe de fonction, on peut contrôler sa dimension de VC **[8,12]**.

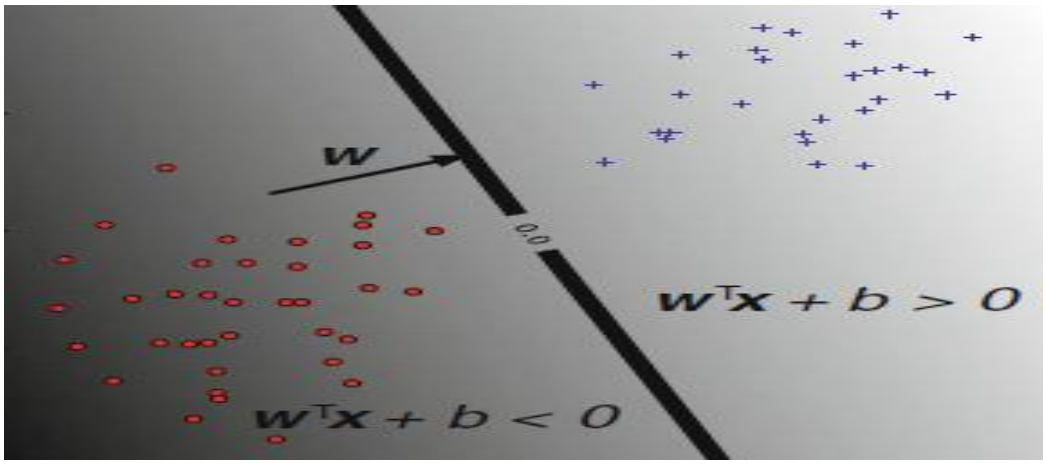


Figure2.3 : Classifieur linéaire et marge.

Un classifieur linéaire est défini par un vecteur normal à l'hyperplan w et un biais b:la frontière de décision est {x| $\langle w, x \rangle$ +b=0}(ligne continue).Chacun des deux sous-espaces séparés par l'hyperplan correspond à une classe, i.e. f(x)=signe($\langle w, x \rangle$ +b). La marge du classifieur linéaire est la distance minimale entre les échantillons de l'ensemble d'apprentissage et la frontière de décision. Sur le schéma, il s'agit de la distance entre la ligne continue et les lignes discontinues **[12]**.

3. SVM principe de fonctionnement général

3.1. Notions de base: Hyperplan, marge et support vecteur

Pour deux classes d'exemples donnés, le but de SVM est de trouver un classificateur qui va séparer les données et maximiser la distance entre ces deux classes. Avec SVM, ce classificateur est un classificateur linéaire appelé hyperplan.

Dans le schéma qui suit, on détermine un hyperplan qui sépare les deux ensembles de points **[16]**.

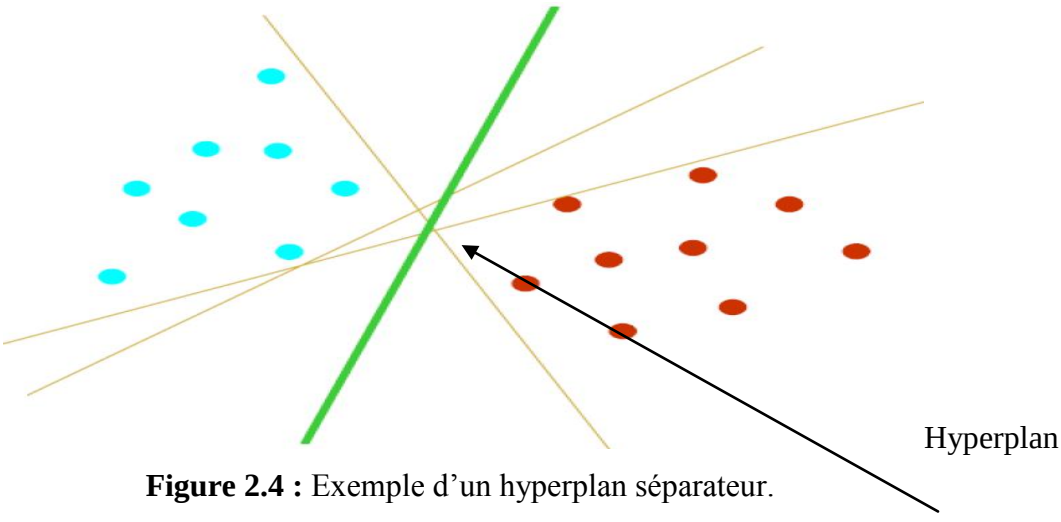


Figure 2.4 : Exemple d'un hyperplan séparateur.

Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support.

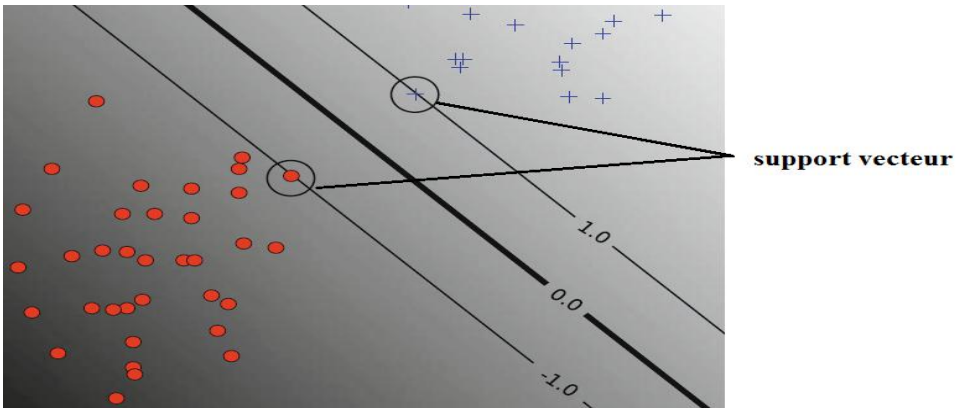


Figure 2.5 : Exemple de vecteurs de support.

Il est évident qu'il existe une multitude d'hyperplan valide mais la propriété remarquable des SVM est que cet hyperplan doit être optimal. Nous allons donc en plus chercher parmi les hyperplans valides, celui qui passe « au milieu » des points des deux classes d'exemples. Intuitivement, cela revient à chercher l'hyperplan le « plus sûr ». En effet, supposons qu'un exemple n'ait pas été décrit parfaitement, une petite variation ne modifiera pas sa classification si sa distance à l'hyperplan est grande. Formellement, cela revient à chercher un hyperplan dont la distance minimale aux exemples d'apprentissage est maximale **[11]**.

On appelle cette distance « marge » entre l'hyperplan et les exemples. L'hyperplan séparateur optimal est celui qui maximise la marge. Comme on cherche à maximiser cette marge, on parlera de séparateurs à vaste marge.

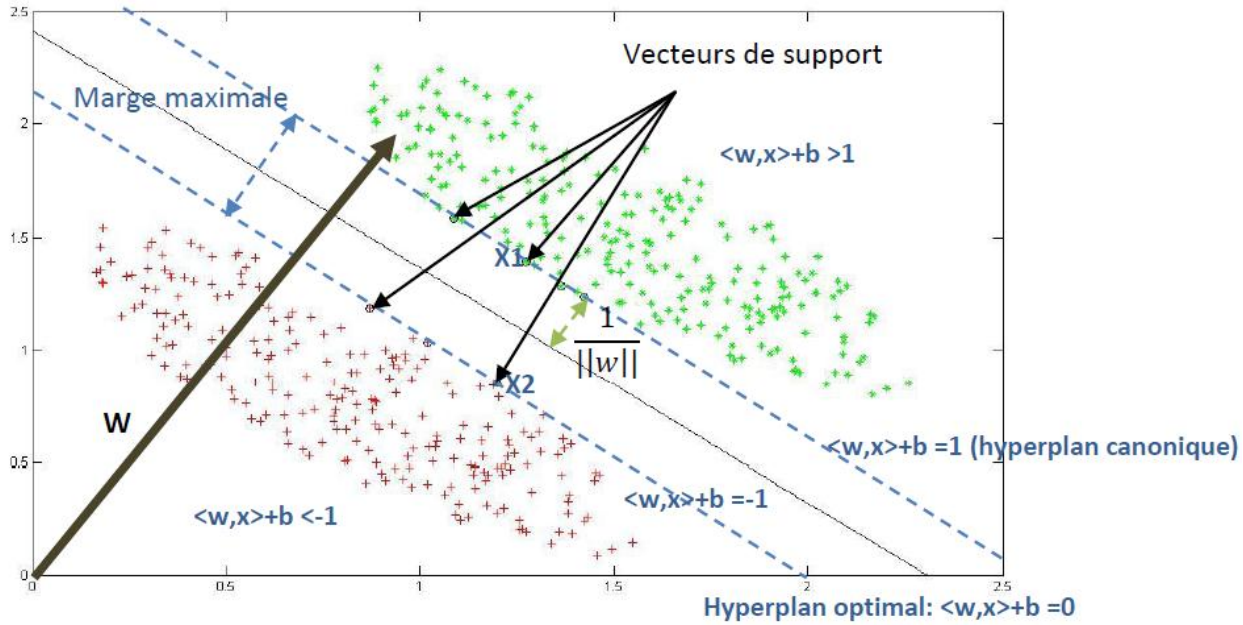


Figure 2.6 : Exemple de marge maximal (hyperplan optimal)

3.2. Pourquoi maximiser la marge ?

Intuitivement, le fait d'avoir une marge plus large procure plus de sécurité lorsque l'on classe un nouvel exemple. De plus, si l'on trouve le classificateur qui se comporte le mieux vis-à-vis des données d'apprentissage, il est clair qu'il sera aussi celui qui permettra au mieux de classer les nouveaux exemples. Dans le schéma qui suit, la partie droite nous montre qu'avec un hyperplan optimal, un nouvel exemple reste bien classé alors qu'il tombe dans la marge. On constate sur la partie gauche qu'avec une plus petite marge, l'exemple se voit mal classé **[16]**.

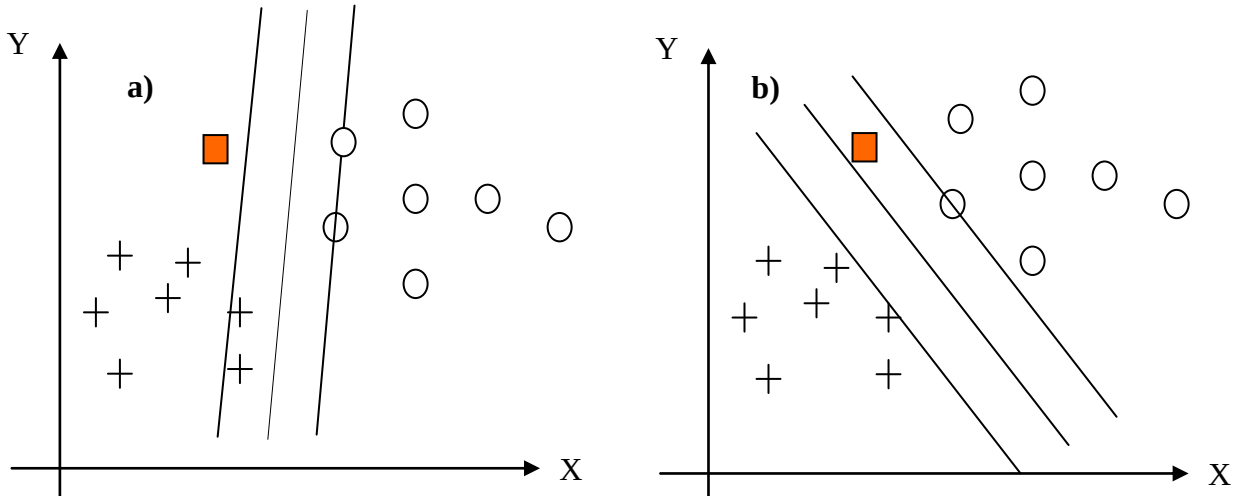


Figure 2.7 :a) Hyperplan avec faible marge **b)** Meilleur hyperplan séparateur **[16]**.

En général, la classification d'un nouvel exemple inconnu est donnée par sa position par rapport à l'hyperplan optimal. Dans le schéma suivant, le nouvel élément sera classé dans la catégorie des « + ».

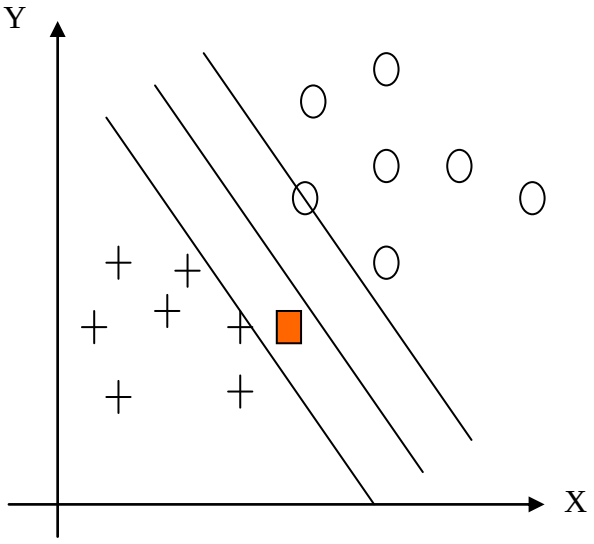


Figure 2.8 : Exemple de classification d'un nouvel élément **[16]**.

3.3. Linéarité et non-linéarité

Parmi les modèles des SVM, on constate les cas linéairement séparable et les cas non linéairement séparable. Les premiers sont les plus simples de SVM car ils permettent de trouver facilement le classificateur linéaire. Dans la plupart des problèmes réels il n'y a pas de séparation linéaire possible entre les données, le classificateur de marge maximale ne peut pas être utilisé car il fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables **[16]**.

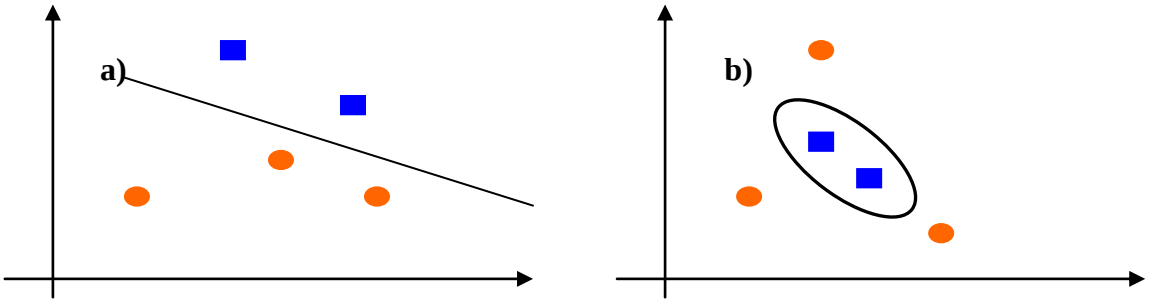


Figure 2.9 : a) Cas linéairement séparable **b)** Cas non linéairement séparable **[16]**.

3.4. Cas non linéaire

Pour surmonter les inconvénients des cas non linéairement séparable, l'idée des SVM est de changer l'espace des données. La transformation non linéaire des données peut permettre une séparation linéaire des exemples dans un nouvel espace. On va donc avoir un changement de dimension. Cette nouvelle dimension est appelé « espace de re-description ».En effet, intuitivement, plus la dimension de l'espace de re-description est grande, plus la probabilité de pouvoir trouver un hyperplan séparateur entre les exemples est élevée. Ceci est illustré par le schéma suivant :

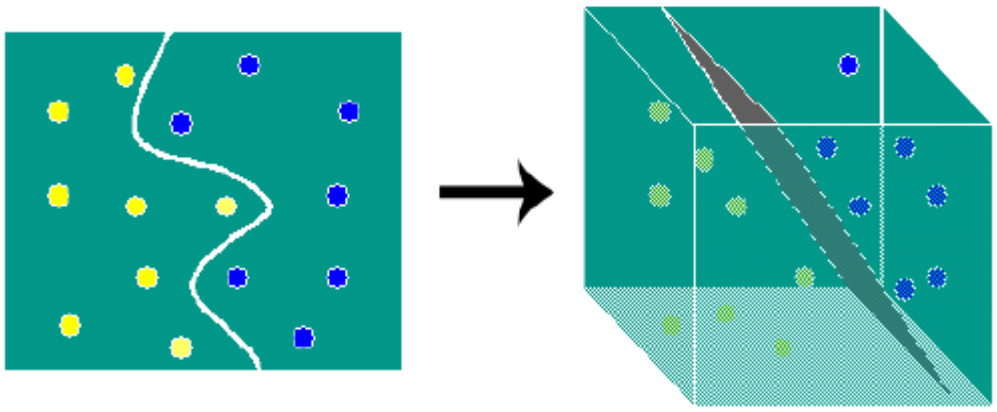


Figure 2.10 : Exemple de changement de l'espace de données **[16]**.

On a donc une transformation d'un problème de séparation non linéaire dans l'espace de représentation en un problème de séparation linéaire dans un espace de re-description de plus grande dimension. Cette transformation non linéaire est réalisée via une fonction noyau. En pratique, quelques familles de fonctions noyau paramétrables sont connues et il revient à l'utilisateur de SVM d'effectuer des tests pour déterminer celle qui convient le mieux pour son application. On peut citer les exemples de noyaux suivants : polynomiale, gaussien, sigmoïde et laplacien **[16]**.

4. Fondements mathématiques

Nous allons détailler dans les paragraphes ci-dessous les principes mathématiques sur lesquels repose SVM.

4.1. Problème d'apprentissage

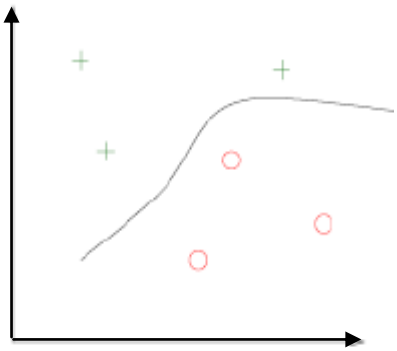
On s'intéresse à un phénomène f (éventuellement non déterministe) qui, à partir d'un certain jeu d'entrées x, produit une sortie y = f(x).

Le but est de retrouver cette fonction f à partir de la seule observation d'un certain nombre de couples entrée-sortie {(xi; yi) : i = 1,…, n} afin de « prédire » d'autres évènements.

On considère un couple (X, Y) de variables aléatoires à valeurs dans X x Y.
Seul le cas : Y = {-1, 1} (classification) nous intéresse ici (on peut facilement étendre au cas : card(Y) = m > 2 et au cas $Y = \Re$. La distribution jointe de (X, Y) est inconnue.

Sachant qu'on observe un échantillon S = {(X1, Y1),..., (Xn, Yn)} de n copies indépendantes de (X, Y), on veut: construire une fonction h : $X \rightarrow Y$ telle que P (h(X) != Y) soit minimale **[14]**.

Illustration :
Trouver une frontière de décision qui sépare l'espace en deux régions (pas forcément connexes).

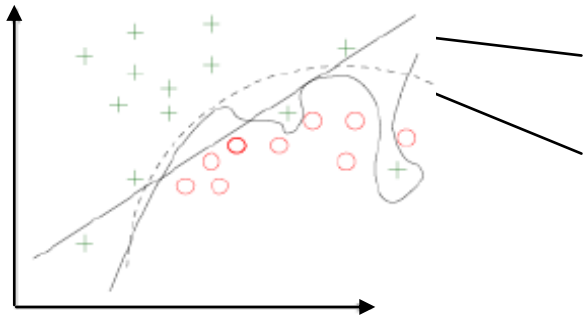


Connaissant h, on peut en déduire la classification des nouveaux points c'est à dire trouver une frontière de décision.

Le problème est de trouver une frontière assez éloignée des points de différentes classes. C'est ce qui constituera l'un des problèmes majeurs de classification grâce aux SVMs [14].

Figure 2.11 : Illustration du problème détermination de frontière assez éloignée des points de différentes classes **[14]**.

Sur et sous- apprentissage :



Si les données sont générées par un modèle quadratique :
Le modèle linéaire est en situation de sousapprentissage
Le modèle de haut degré est en situation de surapprentissage (apprentissage par coeur)

Il faut donc trouver un compromis entre adéquation aux données et complexité pour pouvoir généraliser.

Figure 2.12 : Illustration des sous et sur apprentissage **[14]**.

4.2. Classification à valeurs réelles

Plutôt que de construire directement h : $\boldsymbol{X} \rightarrow$ {-1, 1}, on construit :
f : $X \rightarrow R$ (ensemble des réels).La classe est donnée par le signe de f ;
h = signe (f).

L'erreur se calcule avec P (h(X) != Y) = P (Yf(X) ≤0). Ceci donne une certaine idée de la confiance dans la classification. Idéalement, |Yf(X)| est proportionnel à P (Y|X).Yf(X) représente la marge de f en (X, Y). Le but à atteindre est la construction de f et donc h. Nous allons voir comment y parvenir **[16]**.

4.2.1. Transformation des entrées

Il est peut être nécessaire de transformer les entrées dans le but de les traiter plus facilement. X est un espace quelconque d'objets. On transforme les entrées en vecteurs dans un espace F (feature space) par une fonction: $\Phi : X \rightarrow F$; $\boldsymbol{F}$ n'est pas nécessairement de dimension finie mais dispose d'un produit scalaire (espace de Hilbert). L'espace de Hilbert est une généralisation de l'espace euclidien qui peut avoir un nombre infini de dimensions. La non linéarité est traitée dans cette transformation, on peut donc choisir une séparation linéaire (on verra plus loin comment on arrive à ramener un problème non linéaire en un problème linéaire classique) **[16]**.

Dès lors, il s'agit de choisir l'hyperplan optimal qui classifie correctement les données (Lorsque c'est possible) et qui se trouve le plus loin possible de tous les points à classer.

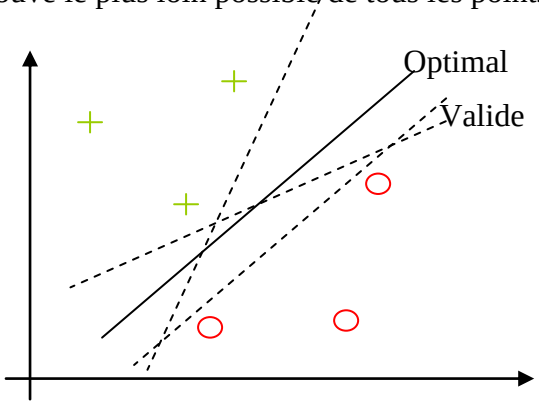


Figure 2.13 : Exemple de recherche d'un hyperplan optimal **[14]**.

Mais l'hyperplan séparateur choisi devra avoir une marge maximale.

4.2.2. Maximisation de la marge

La marge est la distance du point le plus proche à l'hyperplan.

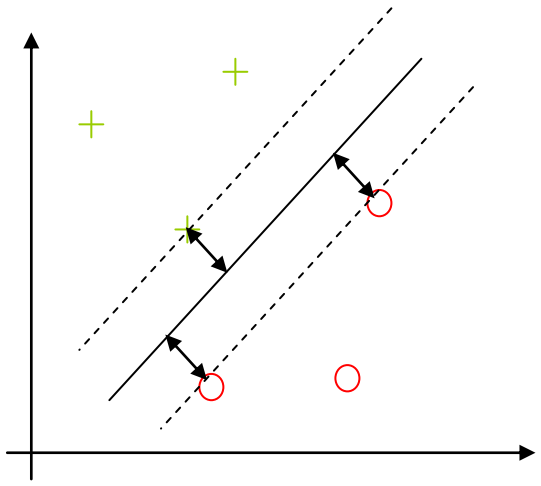


Figure 2.14 : Illustration de la relation entre marge, points de vecteurs de support et hyperplan optimal **[9]**.

4.3. Temps de calcul et convergence

4.3.1. Complexité

Nous allons évaluer la complexité (temps de calcul) de l'algorithme SVM. Elle ne dépend que du nombre des entrées à classer (d) et du nombre de données d'apprentissage (n).

On montre que cette complexité est polynomiale en n.

$$dn^2 \leq \text{Complexité} \leq dn^3$$

Taille de la matrice hessienne = n²
En effet, on doit au moins parcourir tous les éléments de la matrice ainsi que toutes les entrées. Pour un très grand nombre de données d'apprentissage, le temps de calcul explose. C'est pourquoi les SVMs sont pratiques pour des « petits » problèmes de classification **[16]**.

4.3.2. Pourquoi SVM marche?

Les noyaux précédents qui sont les plus utilisés, remplissent les conditions de mercer (facile à vérifier une fois qu'on a le noyau).

Normalement, la classe (le nombre) des hyperplans de R^d est de dH = d + 1. Mais la classe des hyperplans de marge 1/||w|| tels que ||w||² ≤ c est bornée par : dH ≤ Min (R² c, d) + 1 Où R est le rayon de la plus petite sphère englobant l'échantillon d'apprentissage S Donc dH peut être beaucoup plus petit que la dimension d de l'espace d'entrée X ; il est donc toujours possible d'en trouver un c'est la raison pour laquelle **[11]**.

5. SVMs et analyse des bases de données

5.1. Introduction

Les machines à vecteur support présentées au chapitre précédent sont des outils très puissants pour plusieurs tâches d'analyse des bases de données, et qui peuvent faire face à des problèmes difficiles à résoudre par les méthodes classiques d'analyse statistiques et de classification.

Les SVMs viennent même d'être intégrées dans des systèmes de gestion de bases de données tels qu'Oracle, qui présente à partir de sa version 10 g une intégration complète des SVMs.

En effet, l'analyse des bases de données dans les différentes étapes du processus du data mining peut profiter de la robustesse des SVMs pour améliorer ses performances.

5.2. Entrepôt de données

Les bases de données analysées sont généralement rassemblées dans des entrepôts de données Data warehouse. Un entrepôt de données est un environnement structuré conçu pour stocker et analyser toutes les parties significatives d'un ensemble de données [21].

Les données sont physiquement et logiquement transformées de plusieurs applications sources dans une structure commerciale maintenue est la mise à jour pour une longue période. Un entrepôt de données est généralement organisé autour d'un sujet majeur dans une entreprise tel que le client, le vendeur, le produit ou l'activité, ce qui affecte directement la conception et l'implémentation des données dans l'entrepôt de données. Les données de l'application source qui ne sont pas utilisées dans l'analyse pour atteindre l'objectif sont exclues de l'entrepôt de données. La figure 2.16 présente une architecture typique d'un entrepôt de données.

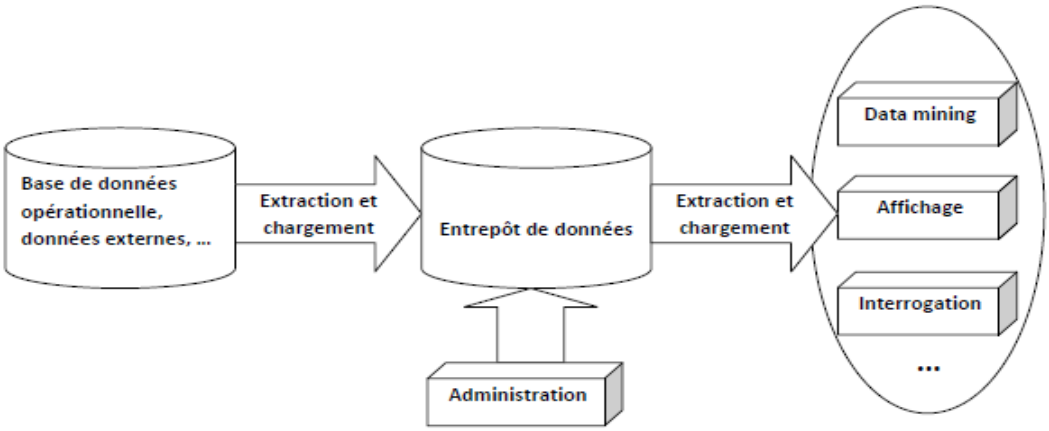


Figure 2.15 : Architecture d'un entrepôt de données.

L'analyse des bases de données se fait généralement pour la découverte des informations qui se cachent dans les grandes quantités de données. Cependant la manipulation des bases de données peut être confrontée dans plusieurs phases du chemin vers la découverte des informations.

En effet, les bases de données peuvent être elles-mêmes des sources naturelles de données telles que dans le cas des systèmes d'information dans une banque ou un supermarché.

Les bases de données peuvent être construites après la phase d'extraction des caractéristiques d'un autre type de données. Après leur prétraitement, les données extraites, sont enregistrés dans un entrepôt de données sous forme de bases de données.

Après la phase d'extraction des connaissances, les informations extraites peuvent être enregistrée sous forme de bases de données.

Les étapes dans lesquelles les bases de données nécessitent d'être analysées sont les étapes de d'acquisition et d'extraction des connaissances [21].

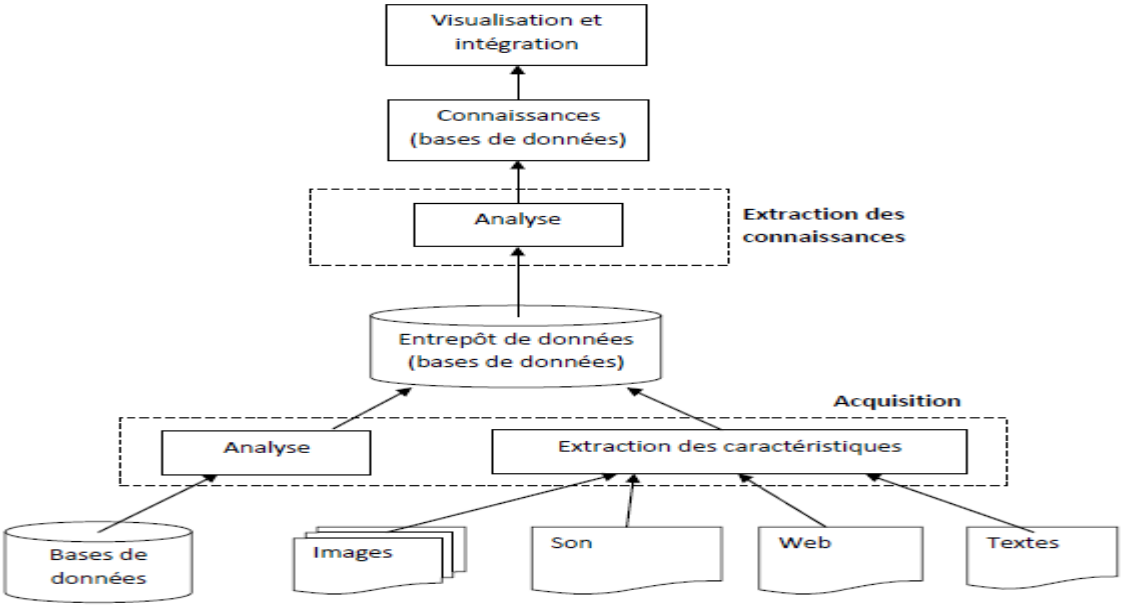


Figure 2.16 : Analyse des BDDs dans le processus de data mining.

6. Les domaines d'applications

SVM est une méthode de classification qui montre de bonnes performances dans la résolution de problèmes variés. Cette méthode a montré son efficacité dans de nombreux domaines d'applications tels que le traitement d'image, la catégorisation de textes ou le diagnostiques médicales et ce même sur des ensembles de données de très grandes dimensions La réalisation d'un programme d'apprentissage par SVM se ramène à résoudre un problème d'optimisation impliquant un système de résolution dans un espace de dimension conséquente. L'utilisation de ces programmes revient surtout à sélectionner une bonne famille de fonctions noyau et à régler les paramètres de ces fonctions. Ces choix sont le plus souvent faits par une technique de validation croisée, dans laquelle on estime la performance du système en la mesurant sur des exemples n'ayant pas été utilisés en cours d'apprentissage. L'idée est de chercher les paramètres permettant d'obtenir la performance maximale. Si la mise en oeuvre d'un algorithme de SVM est en général peu coûteuse en temps, il faut cependant compter que la recherche des meilleurs paramètres peut requérir des phases de test assez longues **[16]**.

7. Avantages et inconvénients

Avantage:

- Absence d'optimum local.
- contrôle explicite du compromis entre la complexité du classifier et l'erreur.
- Possibilité d'utilisation de structure de données comme les chaines de caractères et arbres comme des entrées.
- traitement des données a grandes dimensions.

Inconvénients :

- Demande des données négatives & positives en même temps.
- Besoin d'une bonne fonction Kernel.
- Problèmes de stabilité des calculs dans la résolution de certain programme quadratique a contraintes.

Conclusion

Dans ce chapitre, nous avons tenté de présenter de manière simple et complète le concept de système d'apprentissage introduit par Vladimir Vapnik, les « Support Vecteur Machine » Nous avons donné une vision générale et une vision purement mathématiques des SVM. Cette méthode de classification est basée sur la recherche d'un hyperplan qui permet de séparer au mieux des ensembles de données. Nous avons exposé les cas linéairement séparable et les cas non linéairement séparables qui nécessitent l'utilisation de fonction noyau (kernel) pour changer d'espace. Cette méthode est applicable pour des tâches de classification à deux classes, mais il existe des extensions pour la classification multi classe.

Nous sommes ensuite intéressés aux différents domaines d'application. Il existe des extensions que nous n'avons pas présentées, parmi lesquelles l'utilisation des SVM pour des tâches de régression, c'est-à-dire de prédiction d'une variable continue en fonction d'autres variables, comme c'est le cas par exemple dans la prédiction de consommation électrique en fonction de la période de l'année, de la température, etc. Le champ d'application des SVM est donc large et représente une méthode de classification intéressante.

Chapitre 3

Conception et implémentation du système

- **Introduction**
- **Différents étapes du système**
- **Conclusion**

Introduction

Dans le chapitre précédent nous avons présentés la méthode de classification binaire SVM, inspirées de la théorie statistique de l'apprentissage de Vladimir Vapnik introduite en 1995. Dans ce chapitre nous allons présentés une conception par affinement successif du système en donnant son architecture générale, puis nous détaillons en étudiant séparément chacun de ses composants.

1. Différentes étapes du système

L'objectif de notre système est la réalisation d'un système vocale (saisir les note d'étudiants à partir les nombres d'inscrit) avec ce nom (XLManager) pour ce faire, on utilise un ensemble de commandes vocales où chaque commande passe par une succession d'opérations : acquisition, prétraitement, segmentation et extraction des vecteurs acoustiques, apprentissage et classification, post-traitement et finalement modifier le fichier Excel xls.

Le système peut être vus ou décomposé en modules (composants):

- Acquisition.
- Prétraitement.
- Segmentation parole/ silence (méthode structurelle).
- Extraction des caractéristiques.
- Apprentissage et classification.
- Post-traitement.

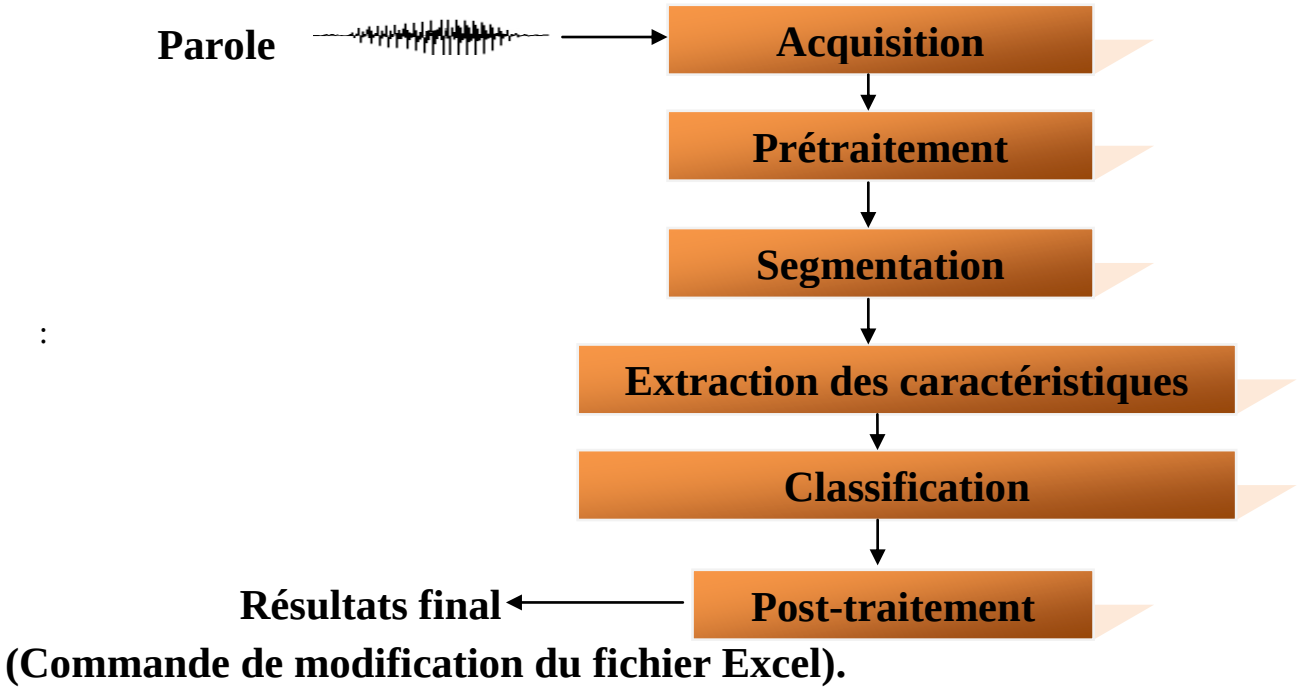


Figure 3.1 : Les différents composants du système.

2. Description des étapes

2.1. Acquisition

D'un sens, acquérir quelque chose c'est devenir propriétaire d'un bien. De ce sens l'acquisition du signal de parole (information) revient à l'appropriation des informations à un micro-ordinateur afin d'exécuter une tâche précise. L'acquisition est la première étape du processus de reconnaissance. Dans notre système nous allons utiliser le microphone comme outil d'acquisition à l'extérieur du PC ainsi que la carte son comme périphérique interne. Après la phase d'acquisition des voix des locuteurs, ces dernières seront automatiquement numérisées sous forme de tampon **[3]**.

2.1.1. Le capteur (microphone)

Le capteur représente le premier élément de l'acquisition. Il est considéré comme un transducteur, dispositif transformant une grandeur physique en une autre grandeur dépendante de la première. Bien qu'un microphone soit obstacle à la propagation des ondes sonores, pour l'acquisition du signal de parole, ce microphone est un capteur comportant un organe sensible aux variations de pression dues à l'onde sonore **[3]**.

Ces variations de pression sont utilisées pour exercer une force sur un système ne pouvant pratiquement pas se déplacer sans cette condition (existence de la force).Il existe plusieurs types de microphone (Microphone : à charbon, à condensateur, à magnétostriction, électrodynamique, électronique, thermique, ionique).On prend le microphone à condensateur comme exemple .Ce dernier se trouve dans un circuit comprenant une résistance et un générateur. L'intensité du courant dans le circuit dépend de ces variations .Ce genre de microphone est le plus performant parmi les microphones disponibles, en plus son avantage majeur est sa petite taille ainsi que sa simple construction **[3]**.

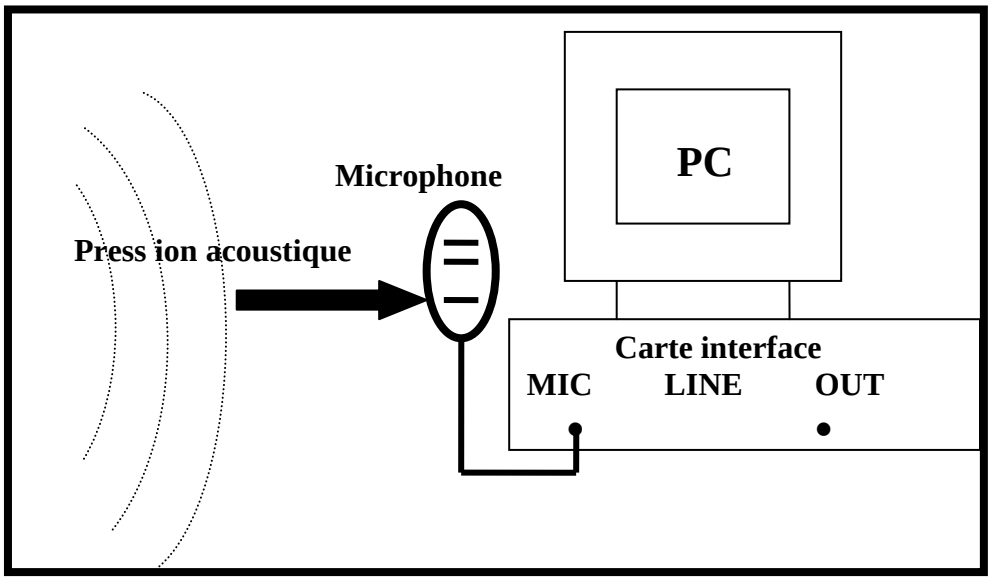


Figure 3.2: schéma synoptique de l'acquisition d'un signal de parole.

2.1.2. Carte interface (carte son)

Une fois le signal analogique, issu du microphone arrive à l'entrée MIC de la carte son, il doit passer par un circuit de conditionnement, qui permet l'amplification et le filtrage de ce signal, après quoi la conversion Analogique-Numérique est effectuée, dans le but de rendre l'information récupérée, traitable par le système numérique (micro-ordinateur).Cette conversion comprend l'échantillonnage, la quantification et le codage **[3]**.

Après la conversion Analogique-Numérique, la carte son passe à la mémorisation des données numérisées dans un espace mémoire ou tampon (buffer) sous forme de valeurs numérique. Ces donnés seront présentés par des vecteurs comportant une série de chiffre .On utilise ce genre de mémorisation plusieurs fois pour un même mot prononcé selon le choix de la taille du dictionnaire voulu, attribuée à l'apprentissage des données **[3].**

Il est à remarquer que la phase de près-traitement n'est pas incluse dans notre système parce que cette tache est gérée par le module d'acquisition du langage choisit et la carte son utilisée.

De ce fait nous allons passer directement à la phase suivante (phase de segmentation).

2.2. Segmentation

Dans ce composant, nous allons faire une analyse temporelle du signal. Une inspection minutieuse de la structure temporelle (forme d'onde), selon un certain nombre de critères, permet une segmentation primaire fiable et précise du signal en deux catégories : parole et silence.

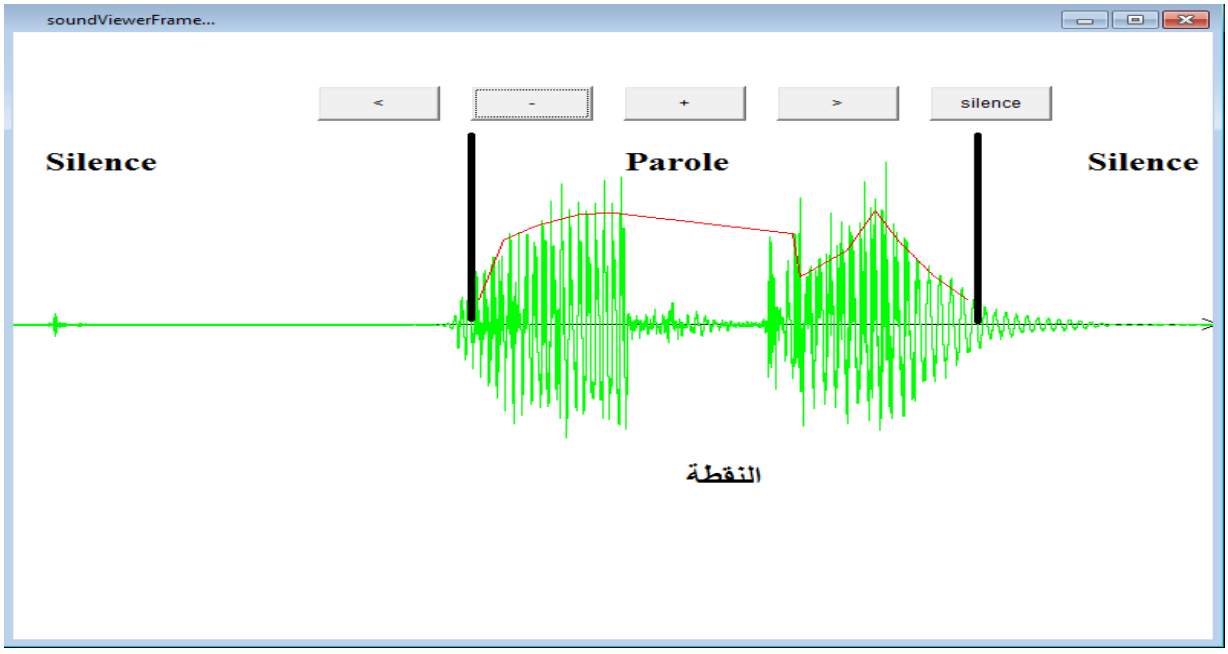


Figure 3.3 : segmentation d'un signale parole avec le mot **"النقطة"** .

Le signal est recodé premièrement selon le passage par zéro de sa dérivée. On obtient ainsi les extremums du signal à partir desquels les silences et les paroles peuvent être détectés. Un silence est une succession de valeurs inférieurs au niveau moyen du silence (par exemple 15) tandis qu'une parole est une succession de piques supérieures à la valeur moyenne du silence. Si une parole est détectée.

Pour segmenter un signale parole nous allons étudier le signale dans chaque petit intervalle et vérifier les valeurs maximales est supérieure à la valeur de silence ou non, si oui alors vérifier les autres valeurs dans le même intervalle, si oui alors partie parole, si non partie silence.

La fonction rouge dans la figure (3.3) est représenter l'intervalle de la signale parole alors le début de cette fonction est le début de mot et la fin est la fin de cette mot.

L'algorithme de segmentation est le suivant :

```
Début
    Signale : tableau de valeurs réel.
    Silence : entier. (donnée).
    Parole  : tableau de valeurs entier.
    Psilence : booléen.

    Pour  i=1  jusqu'à  N  faire
        Si (signal[i] >silence) alors
            Si  (Psilence) alors
                Psilence=false ;
                Parole [j]=i ; //l'indice de début partie parole dans le signal.
                j=j+1 ;
            Fin Si
        Si non
            Si  (non Psilence) alors
                Psilence=true
                Parole [j] =i; // l'indice de fin partie parole dans le signal.
                j=j+1 ;
            Fin Si
        Fin Si
    Fin pour
Fin
```

Généralement, la plupart des mots arabe contiennent un segment de parole qui se répète dans la partie de parole segmentée par la première étape (comme exemples voir les figures ci-dessous pour les mots arabe **"النقطة"** et **"ثلاثة"**).

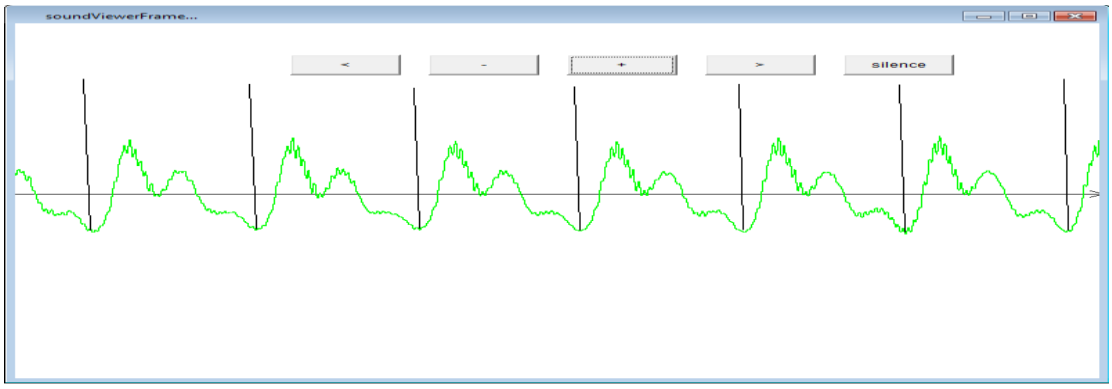


Figure 3.4 : Exemple de répétition du signale parole avec le mot **"النقطة"**.

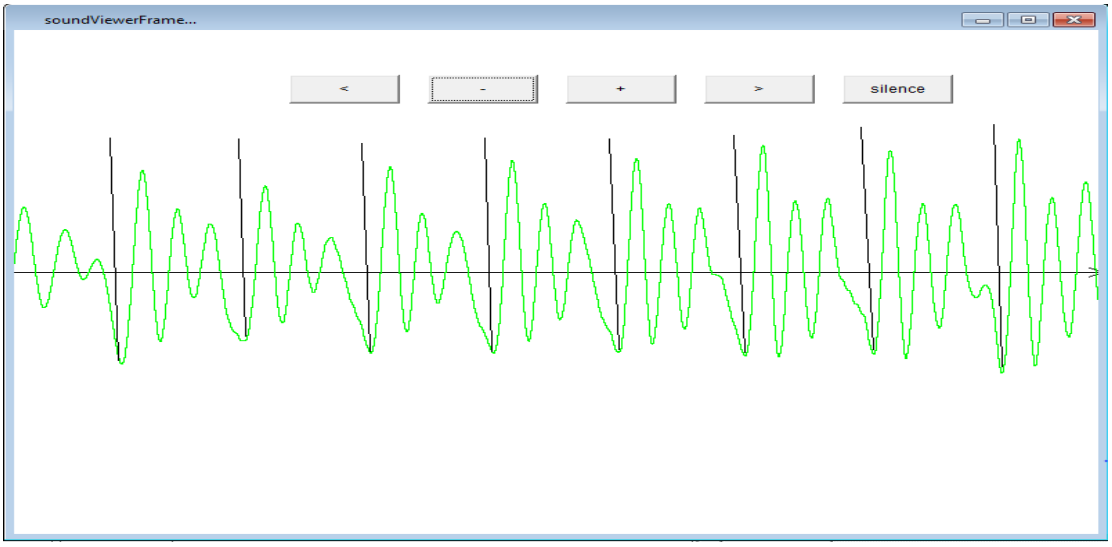


Figure 3.5 : Exemple de répétition du signale parole avec le mot **"ثلاثة"**.

2.3. Extraction des caractéristiques

L'extraction de caractéristiques joue un rôle très important dans les systèmes de reconnaissance la parole ou de la langue. Il existe une diversité de méthodes pour extraire les caractéristiques d'un signal vocal, mais celle dont la fiabilité a été prouvée est bien le codage prédictif linéaire ou LPC (linear predictif coding) car elle extrait l'information d'une petite partie de l'enveloppe spectrale de la parole **[1,19]**.

LPC prend en entrée une parole et fournit des coefficients correspondants à ses caractéristiques statistiques les plus importantes.

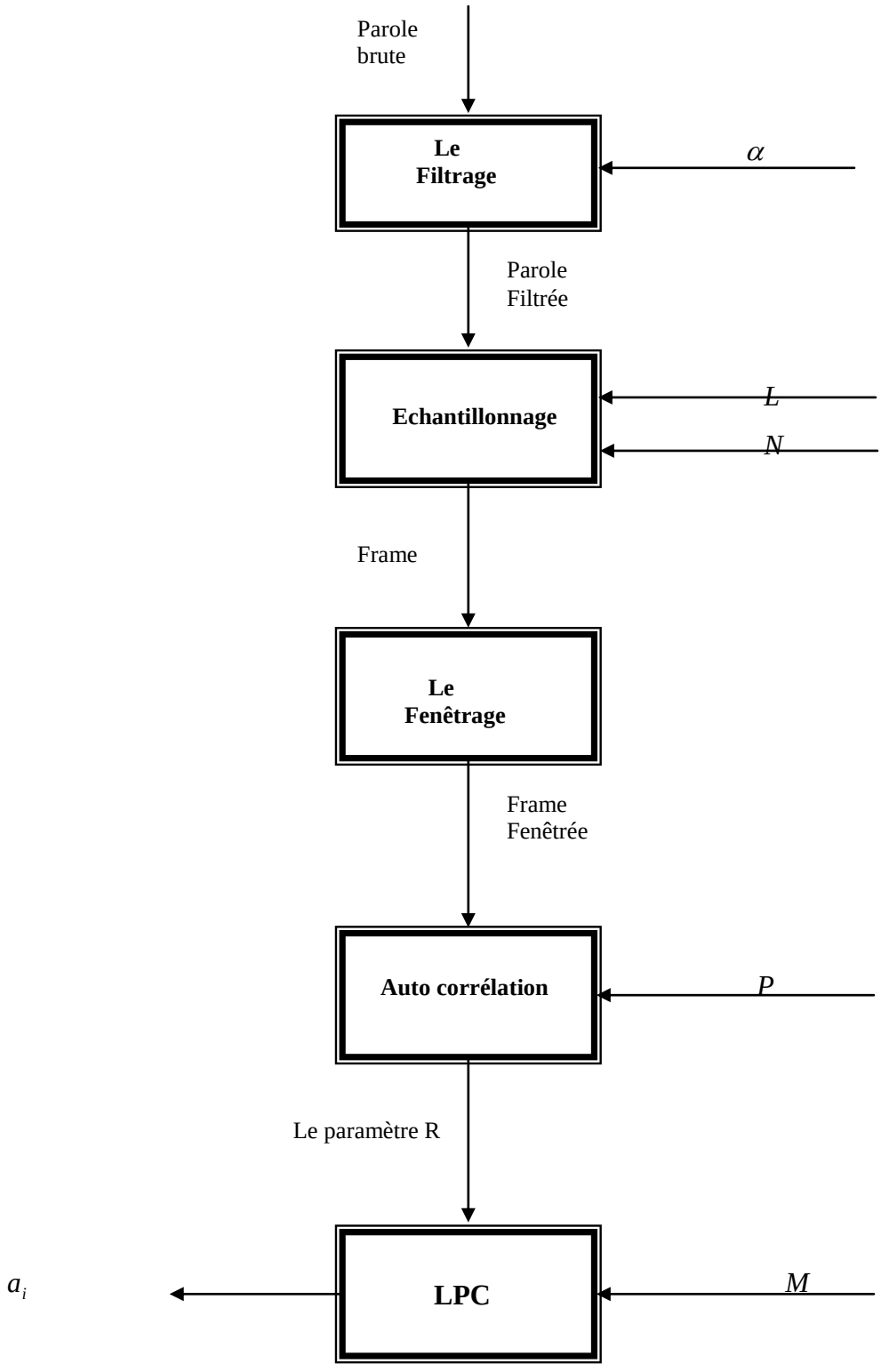


Figure 3.6: L'extraction des paramètres vocaux par LPC

Les coefficients de LPC représentent les caractéristiques les plus importantes**.** Les études faites par Atal montrent l'efficacité des coefficients de LPC dans la reconnaissance et l'identification.

2.3.1. Le principe de la prédiction linéaire

L'analyse par prédiction linéaire est une méthode de modélisation de type essentiellement temporel, bien que son principe repose sur l'hypothèse, selon laquelle un échantillon peut être prédictif comme une combinaison linéaire de N échantillons, qui le précédent d'où le nom de prédiction linéaire, abrégée LPC qui est bien adaptée à l'analyse du signal de parole **[1,19]**.

Cette méthode est fondée sur l'idée qu'un signal qui véhicule un message n'est jamais complètement aléatoire .Il y a une corrélation entre les échantillons successifs. Le codage par LPC utilise cette corrélation pour réduire les données manipulées tout en préservant l'information contenue dans le signal**.** Elle calcule des coefficients sur un échantillon de parole en tenant compte des échantillons précédents **[1,19]**.

$$\hat{y}(n) = -\sum_{i=1}^{P} a(i)y(n-i).$$

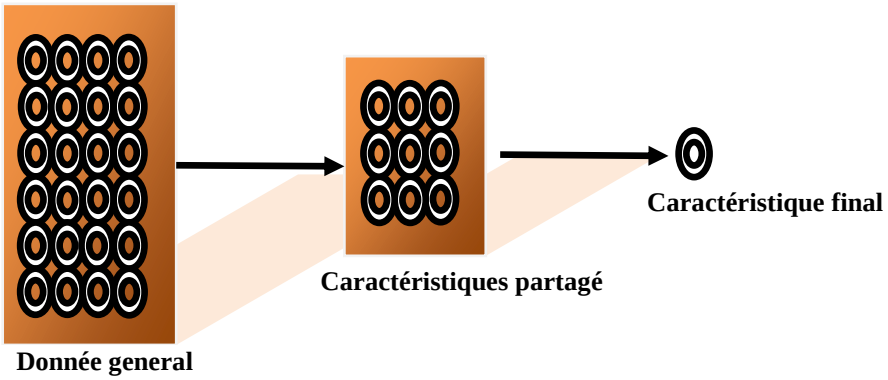


Figure 3.7 : Methode d'extraction de caractiristique.

2.3.2. Avantages de la méthode LPC

On cite les avantages les plus importants qui sont :

- Elle fournit un bon modèle du signal de parole .Cela est spécialement vrai pour la quasi-totalité des régions voisées du signal. Dans lesquelles un de ses modèles (tout pôles) fournit une représentation compacte et précise de l'enveloppe spectrale vocale.

- C'est une méthode d'analyse directe car une fois les critères de l'approximation du signal réel sont fixés, le résultat est obtenu par la résolution d'un système d'équations linéaires.

- C'est un modèle analytique traitable. Il est mathématiquement précis et simple à implémenter soit d'une manière logicielle soit matérielle.

- Le calcul demandé en LPC est considérablement inférieur à celui demandé dans l'implémentation d'autres méthodes comme la méthode de banc de filtre.

- Elle a prouvé son efficacité dans toutes les applications de reconnaissance.

- Les expériences ont montré que les performances des systèmes de reconnaissance basées sur la méthode LPC sont meilleures à celle des systèmes basés sur le banc de filtres**.**

2.3.3. Les phases de LPC

Il existe 5 phases d'extraction de caractéristiques du signal vocal par la méthode LPC :
Le filtrage, l'échantillonnage, le fenêtrage, l'auto corrélation, et le calcul des coefficients. Ces phases ont pour rôle de convertir le signal vocal en coefficients **[1,19].**

a) Le filtrage

Le signal vocal numérique s(n) est capturé par un convertisseur analogique numérique avec une fréquence d'échantillonnage f_s.Le signal est ensuite filtré par un filtre appelé FIR qui a la forme : $H(z)=1-\alpha z^{-1,}$ Où α est dans l'intervalle de 0.9 à 1.0 et reflète le degré du filtrage. La (figure 3.8) montre la réaction fréquentielle du filtrage avec $\alpha=0.95$. Quand $\omega=\pi$ la réaction du filtre est à 32 dB, c'est plus élevé que $\omega=0$.

Le concept de réaction du filtre est sans importance puisque ça n'a pas d'effet sur la perception de la parole. La sortie du filtre peut être liée à l'entrée par l'équation différentielle :

š (n)=s(n)-s(n-1) n=0,1,2,3,…….N-1.

Le filtrage doit généralement être appliqué sur les sons voisés. Quoi que ce soit c'est toujours possible d'appliquer le filtrage en utilisant un échantillon dont la valeur est dépendante de α.
$H(z)=1-\alpha(n)\, z^{-1}$
Où $\alpha(n)$ est une fonction concernant le nombre d'échantillons. Dans ce cas $\alpha(n)$ est dépendant, dans les coefficients des deux premières valeurs d'auto corrélation de l'échantillon courant **[1,19]**.

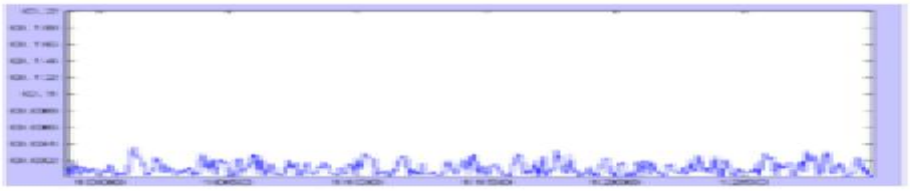


Figure 3.8: LPC de la lettre A.

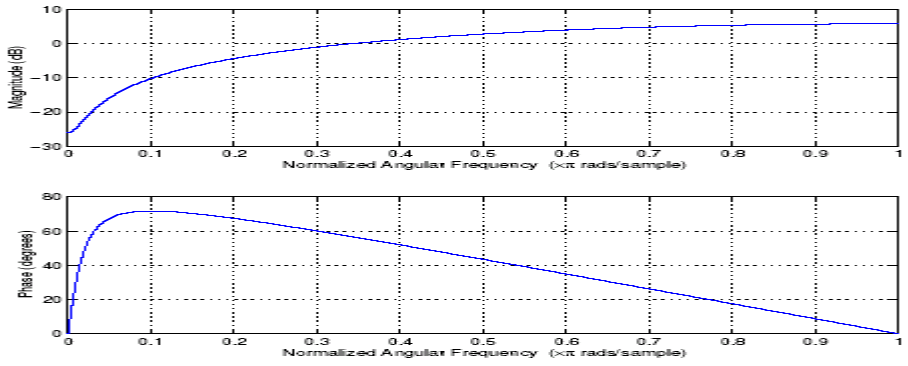


Figure 3.9 : La réaction fréquentielle du filtre.

L'algorithme de filtrage est le suivant :

Début

Signal : Tableau de valeur réel ;
N : réel (longueur de signal) ;
$\alpha = 0.95$: Constante réel ;
Pour i=1 **jusqu'à** N **faire**

Signal[i] =Signal[i]- $\alpha \times$ Signal [i-1];

FinPour

Fin

b) L'échantillonnage

Une fois le signal vocal filtré, il sera ensuite échantillonné en plusieurs échantillons équivalents de même longueur N.

Le début de chaque échantillon est à L samples du début de l'échantillon précédent. Le second échantillon commence à L, le 3ème doit commencer à 2L et ainsi de suite.

Si L<=N, les échantillons se chevauchent et l'estimation des coefficients calculés par LPC montrera un haut niveau de corrélation.

Dans un système où la fréquence d'échantillonnage est 8 KHZ., les valeurs de L et N seront respectivement 80 et 160 **[1,19]**.
Si on définit x_i comme étant le i^ème segment du signal échantillonné š et I échantillons sont requis donc l'échantillonnage sera décrit comme suit :

x_i (n)=š (L_i+ n). n=0, 1,2,……..N-1.

i=0, 1,2,…….I-1.

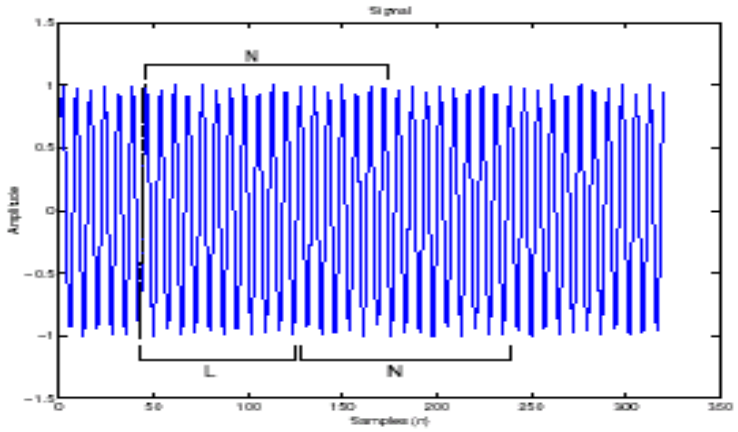


Figure 3.10 : La façon avec laquelle L et N sont utilisés dans

L'algorithme de l'échantillonnage est le suivant :

```
Début
    Signal : Tableau de valeur réel ;
    Frame : une Matrice de valeur réelle ;
    N : entier; (longueur de signal)
    L : entier ;
    I : entier ;
   Pour i=1   jusqu'à I faire
      Pour j=1   jusqu'à N faire
            Frame[i] [j] =Signal [L*i+j];
      FinPour
   FinPour
Fin
```

c) Le fenêtrage

Dans l'analyse du signal vocal, une fenêtre rectangulaire est implicitement employée. La raison d'être du concept de fenêtrage est que la fenêtre rectangulaire possède une discontinuité au début et à la fin de l'échantillon. La distorsion peut être réduite en utilisant une fonction de fenêtrage ω(n) .Il existe plusieurs fonctions de fenêtrage. Le résultat du fenêtrage des segments est défini comme suit:

$$x(n) = x_i(n)\omega(n) \qquad n=0,1,\ldots\ldots\ldots\ldots\ldots N-1.$$

L'algorithme de fenêtrage est le suivant :

```
Début
    W : Tableau de valeur réel ;
    Frame : une Matrice de valeur réelle ;
    N : entier ; // longueur de signal
    I : entier ;
   Pour i=1   jusqu'à I faire
      Pour j=1   jusqu'à N faire
            Frame[i] [j] =Frame[i] [j]*w [j];
      FinPour
   FinPour
Fin
```

d) Analyse et auto corrélation

L'analyse d'auto corrélation a pour rôle d'extraire une harmonie importante et les propriétés des formants à partir d'un signal vocal.

La fonction d'auto corrélation est définie comme suit :

$$R(i) = \frac{1}{N \sum_{n=0}^{N-i-1} x(n)x(n+i)} \quad . \qquad i=0,\ldots..p.$$

Où p est l'ordre de LPC., Ses valeurs sont inclues entre 8 et 16.
R (0) représente la mesure de l'énergie des segments de la parole.

Et peut être utilisé pour éliminer le silence. L'algorithme d'auto corrélation est le suivant :

```
Début

    Frame : Tableau de valeur réel ;
    R : Tableau de valeur réel ;
    N : entier; (longueur de Frame)
    P : entier ; (nombre de coefficients)
    Somme : réel ;

  Pour i=1   jusqu'à P faire

     Somme=0 ;
    Pour j=1   jusqu'à N-i-1faire

          Somme=somme + Frame[j]*Frame [i+j];
    FinPour

              R=Somme ;
  FinPour
Fin
```

e) Le calcul du LPC

Le prochain block fonctionnel converti le (p+1) coefficient d'auto corrélation en coefficient du LPC.

Si on suppose que l'appareil vocal est excité par un signal de bruit blanc ayant une unité de variance et sans sens, nous pouvons alors représenter ce système par un $p^{ème}$ ordre autorégressif.

$$H(z) = \frac{G_p^2}{1 + \sum_{k=1}^{p} a_{pk} Z^{-k}} .$$

Les inconnues dans cette équation :

(G_p^2 , a_{pk} Où k=1,2,……p) sont résolues en utilisant les deux équations suivantes :

$$R(0) = G_p^2 + \sum_{k=1}^{p} a_{pk} R(k).$$

Et

$$R(j) = -\sum_{k=1}^{p} a_{pk} R(j-k). \quad j=1,2,\ldots, p.$$

Ces équations sont communément référenciés comme les équations de Yule Walker.C'est possible de résoudre ces équations en utilisant une méthode récursive. La plus populaire des méthodes récursives est l'algorithme de Levinson Durlin. Cet algorithme est initialisé comme suit :

$$a_{i,i} = -\frac{R(1)}{R(0)}.$$

$$p_i = R(0)(1 - a_{1,1}^2).$$

Et récursivement implémenté pour m=2,…..,p. par :

$$a_{m,m} = -\frac{R(m) + \sum_{i=1}^{m-1} a_{m-1,i} R(m-i)}{p_{m-1}} .$$

$$a_{m,i} = a_{m-1,i} + a_{m,m}\ a_{m-1,m-i}.$$

$$p_m = p_{m-1}(1 - a_{m,m}^2).$$

La solution finale pour les coefficients LPC est donnée comme suit :

$$a_i = a_{p,i} \qquad 1 \le i \le p.$$

$$G_p^2 = p_p .$$

La dérivée de l'algorithme de Levinson est bien représentée sous forme de coefficients K_m (m=1,2,…...p) définis comme suit :

$K_i = a_{i\,i}$ $\quad 1 \le i \le p.$

Les coefficients de réflexion ou PARROR sont directement liés aux croisements de sections non uniformes du conduit vocal tout en formant ainsi un modèle de cet appareil vocal. Le système vocal peut être considéré comme une cascade de P cylindres de longueurs équivalentes **[1,19]**.

Quand l'air traverse l'appareil vocal, la différence au niveau des croisements cause une réflexion sur les frontières où les coefficients sont indiqués par k_m. Ces coefficients sont liés aux ceux de LPC dans la non linéarité. Ils ont été découverts pour être utilisés dans le domaine du codage de la parole **[1,19]**.

Parmi toutes les représentations du LPC, nous trouvons les coefficients cepstraux qui ont été inventés pour assurer une meilleure performance dans la reconnaissance de la parole et celle du locuteur.

L'algorithme de calcul LPC est le suivant :

```
Début
        PC : Tableau de P+1 valeur réel ;
        R : Tableau de P valeur réel ;
        CoefLpc : Tableau de P valeur réel ;
        P : entier ;
        a : Matrice de P+1*P+1 valeur réel ;
        Somme : réel ;

        a [ 1 ] [ 1 ] = - R [ 1 ]  / R [ 0 ] ;    // Calcul du premier coefficient
        PC [ 1 ] = R [ 0 ] * ( 1 – a [ 1 ]  [ 1 ] * a [ 1 ] [ 1 ] ) ;
        Pour  m = 2 jusqu'à  P+1 faire      // Calcul des autres coefficients


                Somme = 0 ;
                Pour i = 1 jusqu'à  m-1 faire
                        Somme = Somme + a [m-1] [i] * R [m-i];
                FinPour
                a [ m ] [ m ] = - ( R [ m ] + Somme  ) / PC [ m-1 ] ;
                Pour i = 1 jusqu'à m faire
                    a [ m ] [ i ] = a [ m-1 ] [ i ] + a [ m ] [ m ] * a [ m-1 ] [ m-i ] ;
                FinPour


                PC [ m ] = PC [ m-1 ] * ( 1 – a [ m ] [ m ] * a [ m ] [ m ] ;
        FinPour

        Pour  m = 1  jusqu'à  P  faire

            CoefLpc[m] =a[i] [i];

        FinPour
Fin
```

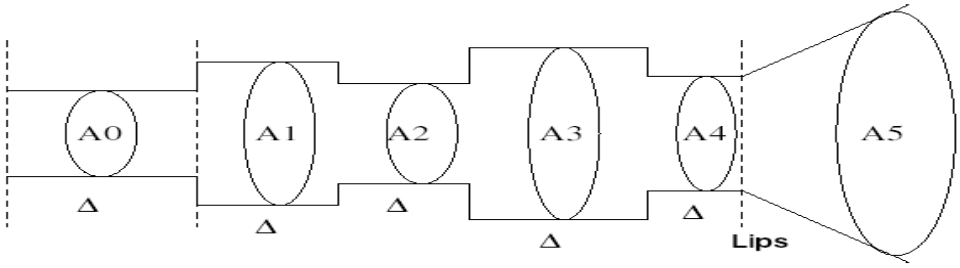


Figure 3.11 : Modèle du tube acoustique de production de la parole.

La Section.	Les Paramètres Utilisés.	Les valeurs Utilisés.
Le filtrage« preemphasis ».	**α.**	**0.95.**
L'échantillonnage« Frameblocker ».	**N.**	**480 (IISC-Microphone).**
	L.	**160 (IISC-Microphone).**
Le fenêtrage« Windowing ».	**ω (n) .**	**0.54-0.46 cos** $\frac{2\pi n}{M-1}$ **.**
L'autocorrélation et l'analyse LPC.	**P.**	**8, 10, 12,40.**
La conversion cepstrale.	**M.**	**P.**

Tableau 3.1: Les paramètres opérationnels utilisés dans l'extraction des caractéristiques avec LPC.

2.4. Classification (SVM et la parole)

Dans le chapitre précédent nous avons parlés des SVM et son utilisation générale, mais dans cette partie nous allons parler de l'utilisation des SVM pour la reconnaissance de la parole. Donc, après l'extraction des paramètres du signal vocal par la méthode LPC, ces paramètre sont utilisés comme une donnée d'entrée pour le composant de classification (SVM), qui va rechercher un hyperplan séparateur qui sépare les exemples dans la phase d'apprentissage et prend une décision de classification dans la phase d'identification.

Dans le module SVM, il y a deux phases : une pour l'apprentissage et l'autre pour la classification. La figure suivante représente la relation entre ces deux phases.

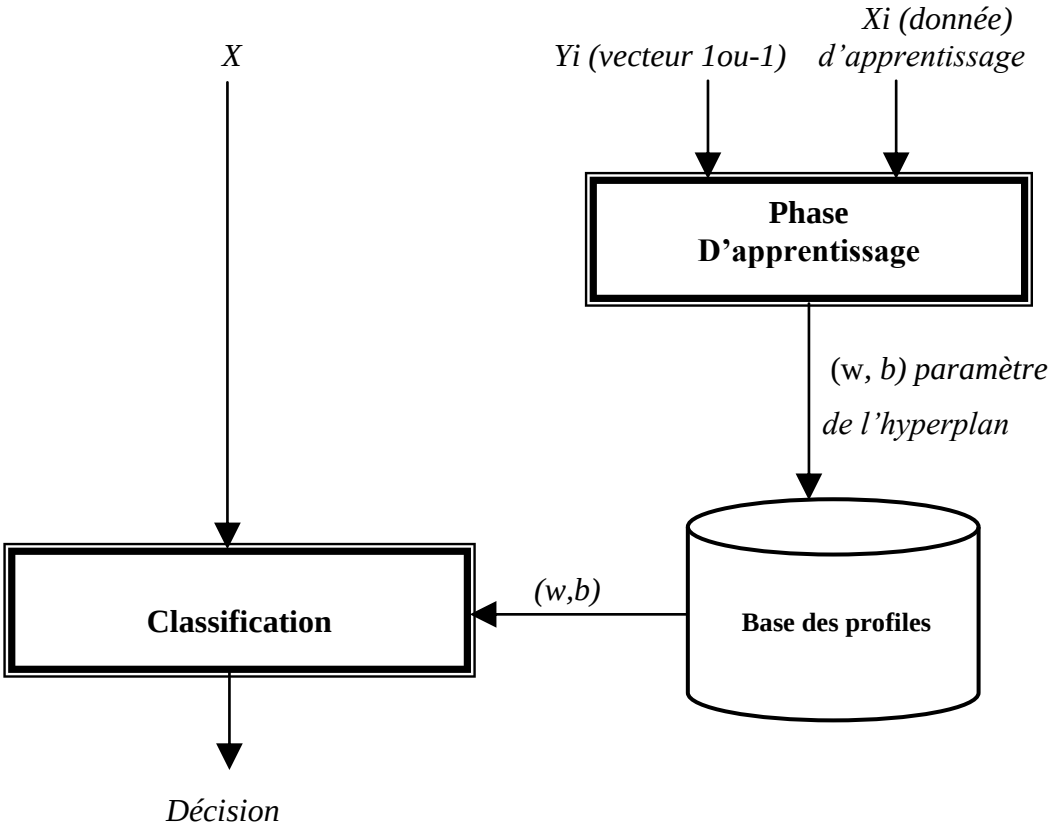


Figure 3.12 : les phases de SVM

Où *X*, *Xi* représentent le vecteur caractéristique des enregistrements du son ; et *Yi* les étiquettes de chaque classe.

Dans la phase précédente nous résolvons un problème dual. Nous connaissons alors *w et b* et nous pouvons définir la fonction de décision pour une nouvelle donnée *x* :

$$f(x) = signe(\langle w.\phi(x)\rangle - b\,.$$

2.5. Post-traitement

Dans la réalité, un nombre important des mots arabe sont composés de deux parties de parole et non pas une seule partie. Dans la phase de segmentation, nous avons parlé de la segmentation du signal en parties de parole ce qui fait le même mot peut être segmenté en deux, et par conséquent, une phase composition des résultats de classification de ces parties est nécessaire.

Par exemple le mot arabe **"النقطة"** est compose de deux parties .

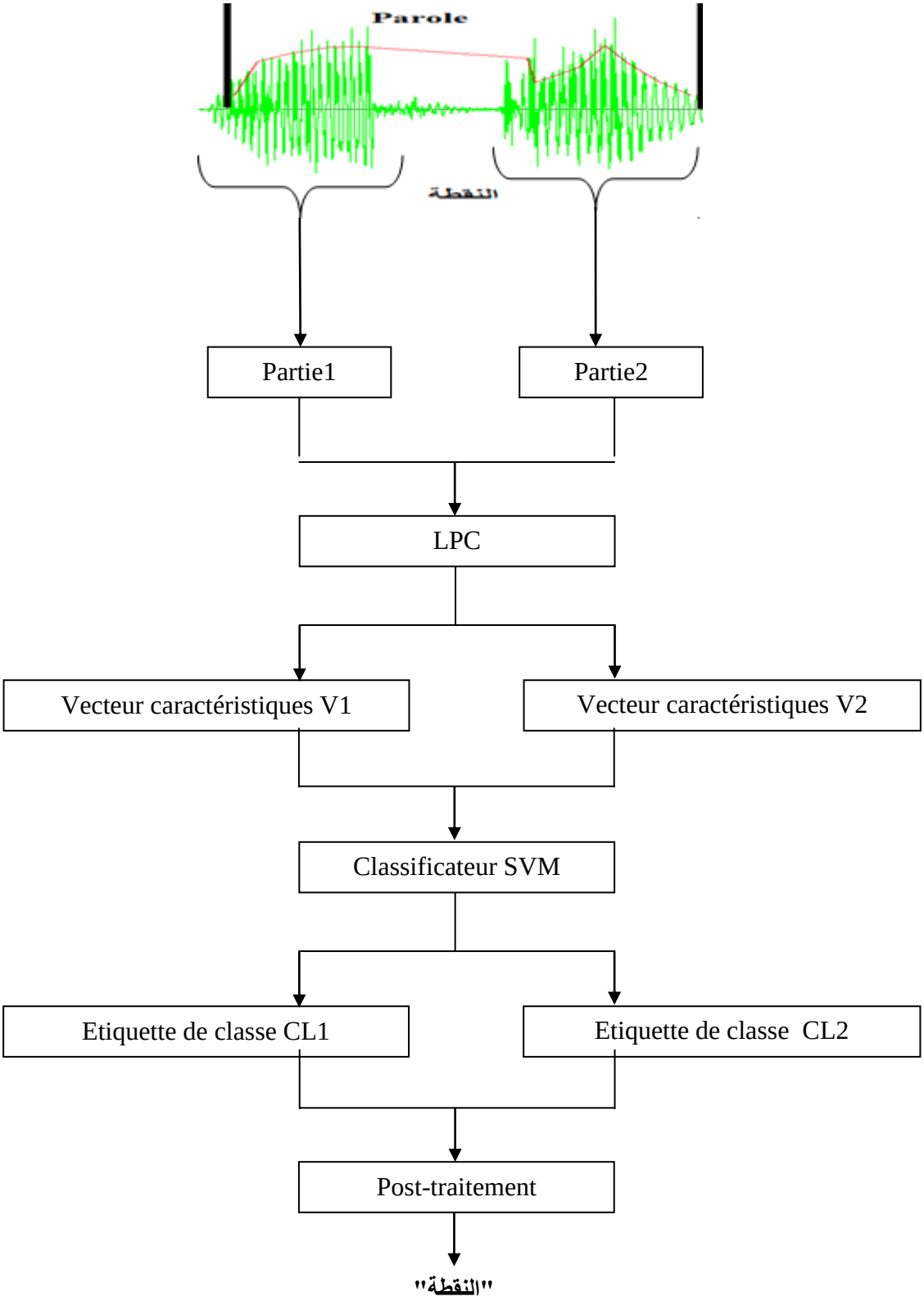


Figure 3.13 : La phase du post-traitement.

Le résultat obtenu par la phase de post-traitement représente une commande qui va engendrer une modification dans le fichier Excel cible.

Conclusion

Nous avons présenté dans ce chapitre les différentes étapes qui peuvent conduire à une conception convenable d'un système de reconnaissance de la parole. Notre système est mono locuteur, on a choisi la méthode de codage LPC pour la paramétrisation du signal acoustique et la méthode SVM pour la reconnaissance des mots isolés. Par la suite, une phase de post-traitement a été utilisée pour améliorer et contrôler les résultats de classification.

Chapitre 4

Réalisation et test de résultats

- **Introduction**
- **Choix du langage de programmation**
- **Interface et fenêtre**
- **Test et résultats**
- **Conclusion**

Introduction

Dans le chapitre précédent nous avons présenté une conception du système en donnant une vue globale du système, en suite nous avons détaillé chaque module composant le système séparément. Dans ce chapitre nous allons voir la réalisation du système : le choix du langage, l'implémentation des différents modules, quelques tests et enfin quelques résultats concernant le taux de reconnaissance.

1. Choix du langage de programmation

Nous avons choisis comme environnement de programmation le langage JAVA qui offre une grande simplicité de manipulation du son, soit en enregistrement (acquisition) ou en génération des fichiers son (sortie). Ce langage possède avantages très intéressants tel que :

- La portabilité des logiciels ;
- La réutilisation de certaines classes déjà développées ;
- La possibilité d'ajouter à l'environnement de base des composants fournis par l'environnement soit même ;
- La quasi-totalité de contrôle de windows (boutons, boites de saisies, listes déroulantes, menus …etc.) qui sont représentés par classes;

2. Interface et fenêtres

En lançant l'application nous allons voir l'interface présentée dans la figure 4.1. L'application peut être utilisée en deux modes :

- Développement ;
- Reconnaissance.

2.1. Mode développement

Le mode développement est chois pour utiliser l'application en mode apprentissage ou mode test.

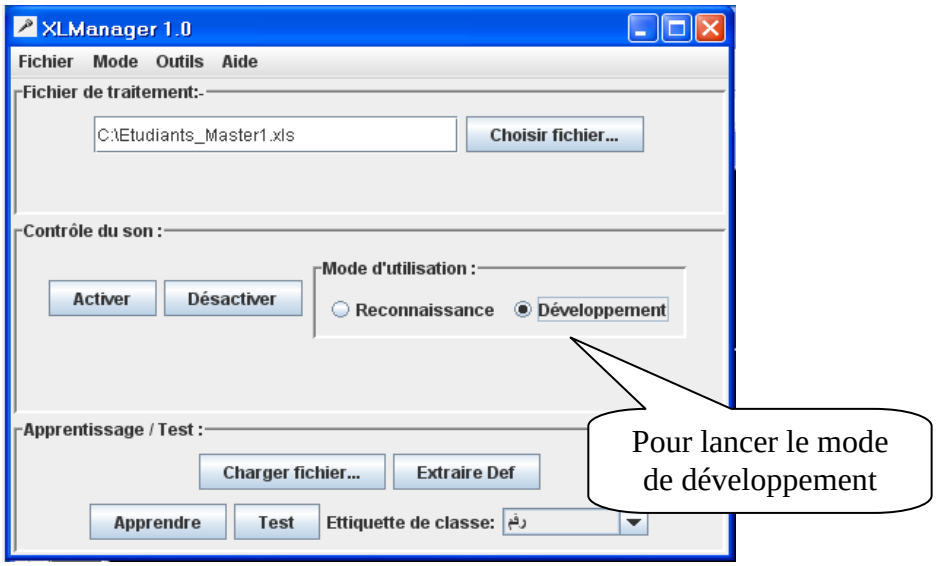


Figure 4.1 : Mode de développement.

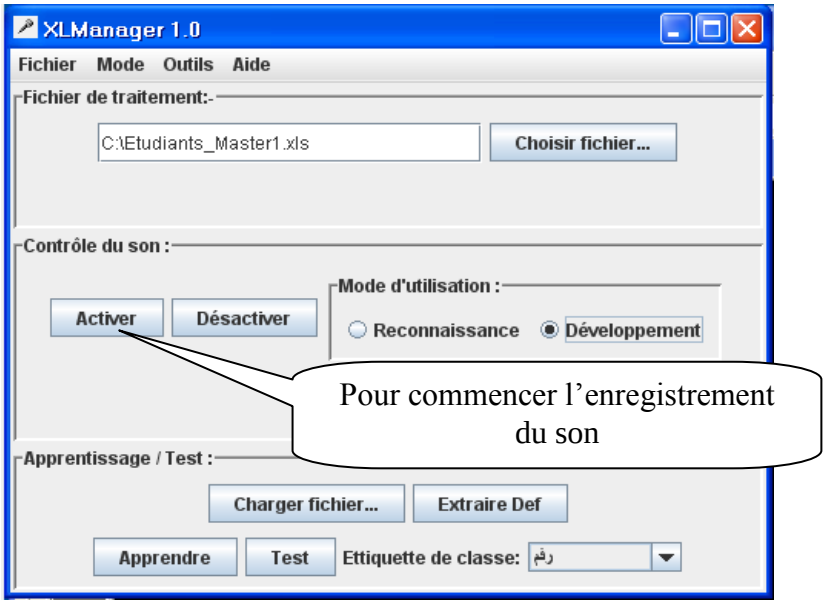


Figure 4.2 : Démarrage d'enregistrement.

Dans ce mode l'application peut être utilisée selon deux sous-modes:

- Apprentissage: comme phase initiale pour aider le système à apprendre les différentes classes.
- Test: pour tester et calculer le taux de reconnaissance, et éventuellement utiliser XLManager.

2.1.1. Mode apprentissage

Ce mode peut être vu comme phase initiale ou d'initialisation de la base de connaissance du système, pour le faire on procède comme suit:

1) choisir une classe cible (l'un des formes primitives), à l'aide du combo box comme le montre la figure suivante:

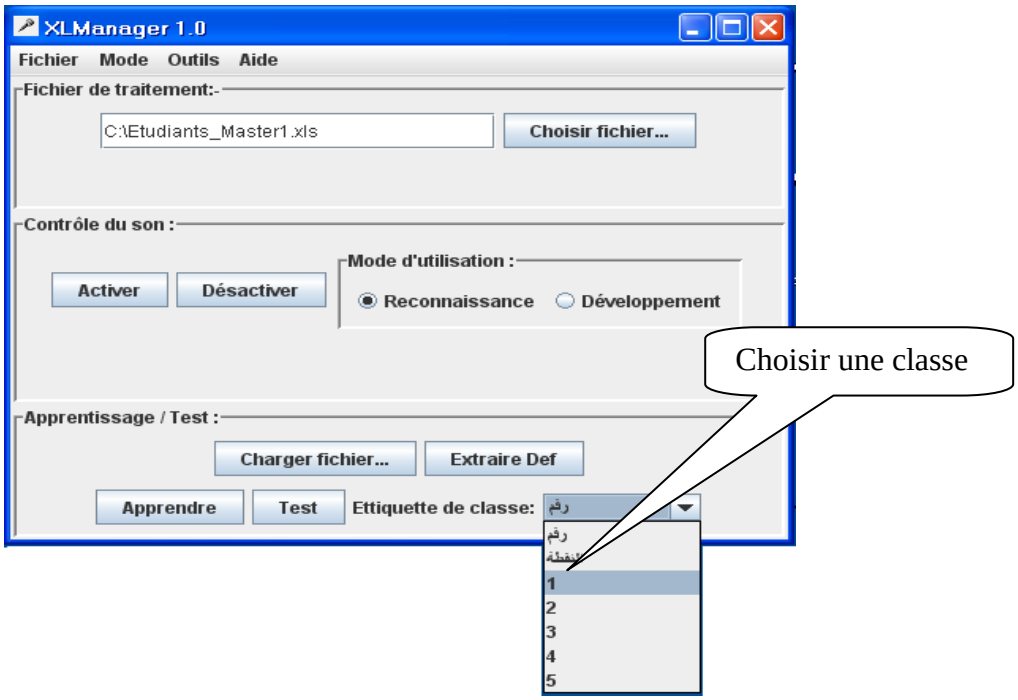


Figure 4.3 : Illustration du choix d'une classe.

2) commencer l'enregistrement de l'exemple de la classe cible, en appuyant sur le bouton *Activer*:

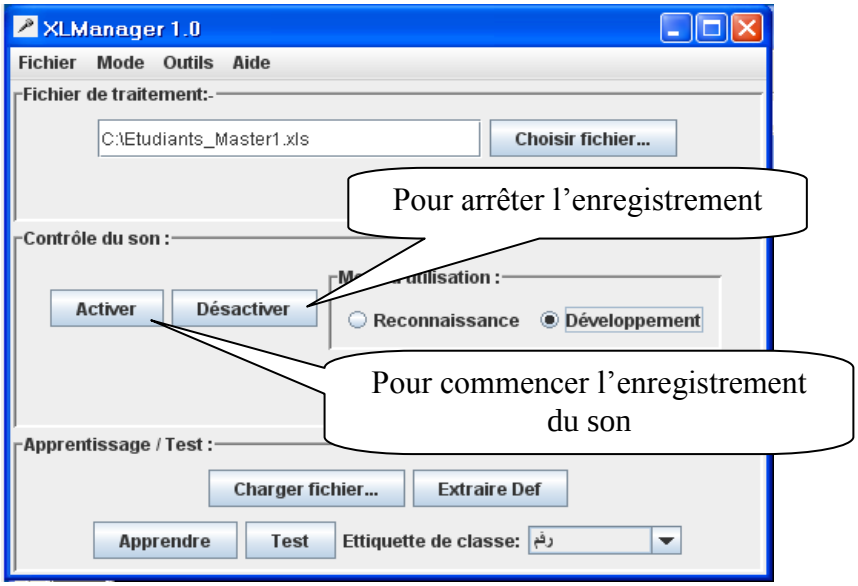


Figure 4.4 : Enregistrer l'exemple de la classe.

En terminant le dicté de l'exemple en question, il faut appuyer sur le bouton *Désactiver*.

3) Cliquer sur le bouton *Extraire Def* pour extraire les caractéristiques de l'exemple enregistré, et les écrire dans le block note du Classifieur.

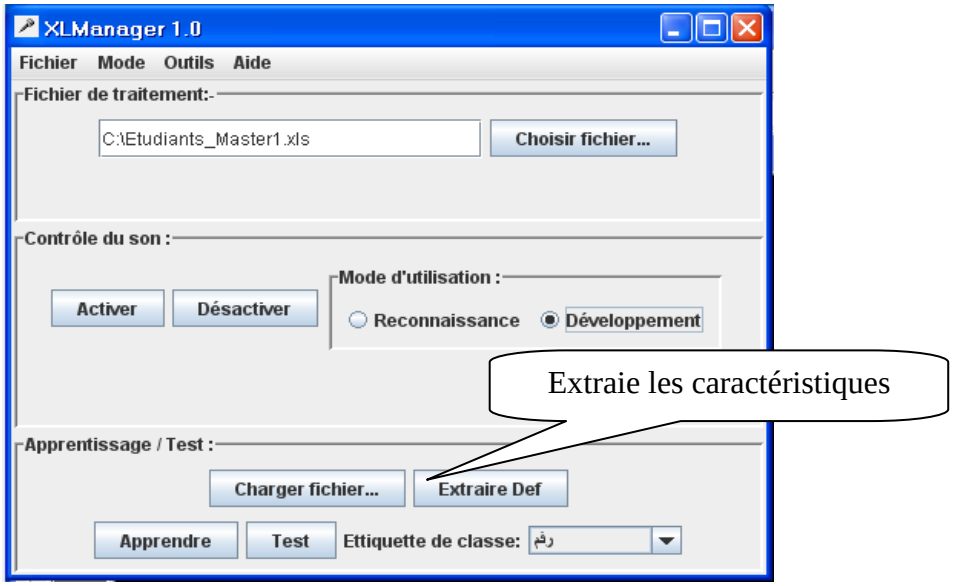


Figure 4.5 : Extraction des caractéristiques de l'exemple enregistré.

Ces étapes décrites ci-dessus sont répétées autant de fois qu'on veut enregistrer d'exemples pour chaque classe. A la fin, nous pouvons appuyer sur le bouton *Apprendre* pour lancer le processus d'apprentissage.

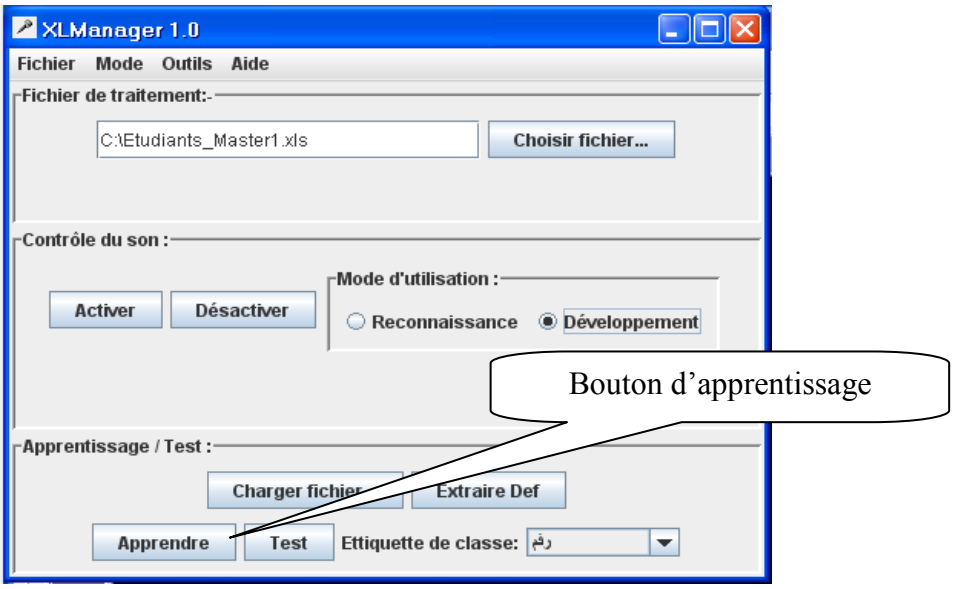


Figure 4.6 : Illustration du mode apprentissage.

2.1.2. Mode test

Ce mode ne peut être exploitable qu'après avoir terminé la phase d'apprentissage et il suit presque les mêmes étapes avec des différences légères. Les étapes 1 et 2 sont les mêmes, pour l'étape 3, on clique sur le bouton de test *Test*.

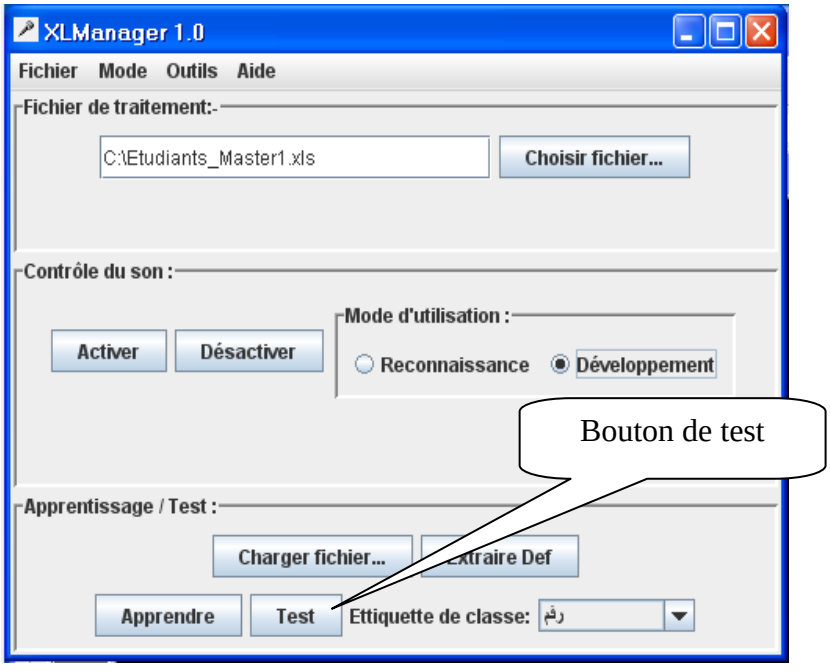


Figure 4.7 : Illustration du mode de test.

Dans la phase de test, l'identité de la commande vocale est d´exterminée à partir d'une comparaison entre les caractéristiques actuelle et les caractéristiques de référence.

2.2. Mode Reconnaissance

Ce mode peut être choisit en cliquant le bouton radio *Reconnaissance* (voir la figure ci-dessous). Après avoir choisir ce mode, il aussi faire le choix d'un fichier Excel cible en appuyant sur le bouton *Choisir fichier*.

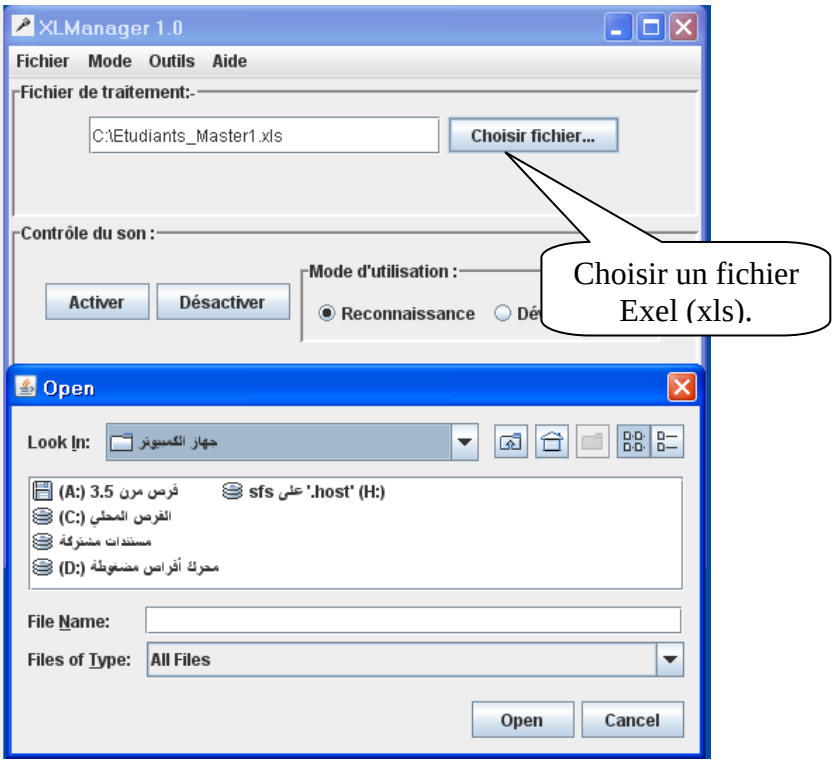


Figure 4.8 : Choisir un fichier Excel (xls).

C'est le mode d'utilisation normale de l'application, où on lance l'enregistrement continu par le bouton *Activer* et le système commence à faire la reconnaissance en temps réel la parole dictée la traduire en commandes qui les appliquent par la suite en modifiant le fichier Excel choisit.

Dans ce travail, nous avons choisis d'utiliser la bibliothèque java JExcelApi qui permet de lire, écrire et modifier des fichiers Excel sous le format 97-2003 (xls).

3. Test et résultats

Dans cette partie nous allons tester la méthode d'apprentissage SVM, et précisément comment déterminer les paramètres qui influent sur l'efficacité de la classification par SVM. Ces paramètres sont le paramètre de pénalisation d'erreur C et pour le noyau choisit (noyau gaussien) on doit déterminer un certain nombre de paramètres pour ajuster sa forme quand à la distribution des données d'apprentissage.

Le choix des paramètres adaptés est une étape cruciale, un ensemble trop encadré peut ne pas parvenir à séparer les données initiales, et au contraire un ensemble trop libre peut aboutir à l'incapacité de généralisation. La méthode de validation croisée est utilisée pour trouver les valeurs les plus adaptées.

Les paramètres C et gamma (sigma) ne sont pas les seuls facteurs qui influent sur le résultat de classification par SVM mais en plus le type de jeu de données a une influence directe, pour bien montrer ça, on a choisi deux classes qui sont facilement séparable « 1 ,2 » et deux autres qui ne le sont pas « 5,7 ».

La configuration de quelques paramètres peut être modifiée par l'utilisateur mais avec prudence, ces paramètres sont les suivant :

- Le paramètre P de LPC =40 ;
- La fréquence d'échantillonnage =44100 ;
- Les paramètres C =100, et gamma =1 de SVM.

On a pris pour chaque classe des exemples dans la phase d'apprentissage, les résultats de test sont résumés dans le tableau suivant :

Classe	Nombre d'exemples	Taux de reconnaissance
1	60	75.66 %
2	60	98.33 %
5	50	97.33 %
Total	170	90.44%

Table 4.1 : Illustration du taux de reconnaissance pour les classes « 1, 2 et 5».

Pour mieux comprendre ces résultats, nous les avons présentés sur l'histogramme et le cercle de secteurs suivants :

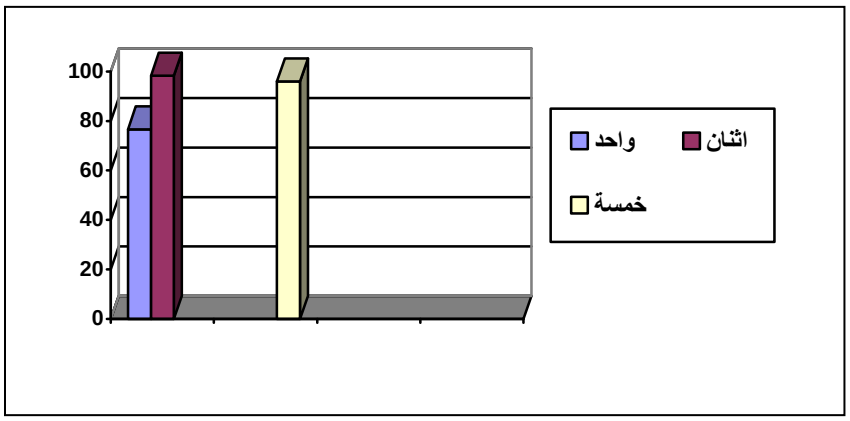


Figure 4.9: Le taux de reconnaissance des classe« 1, 2 et 5».

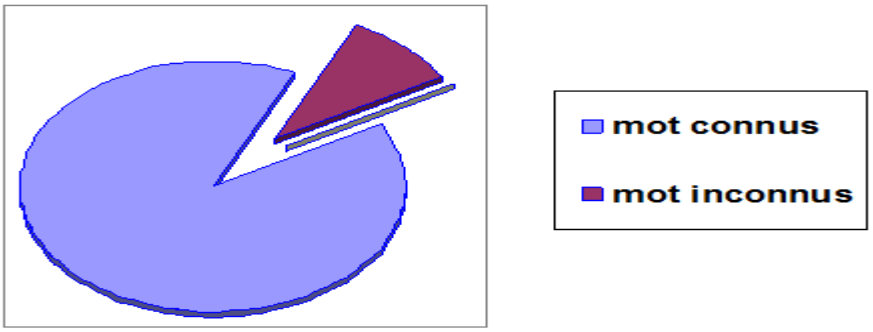


Figure 4.10 : Le taux de reconnaissance globale.

Où mot connus veut dire classes facilement séparables et mot inconnus classes non facilement séparables.

Conclusion

Les résultats de la classification par SVM obtenus peuvent être améliorés en optimisant les paramètres suivants :

- Les paramètres C et gamma.
- Le nombre d'exemples.
- Qualité de matérielle.
- Le bruit et La qualité de Sound traité
- L'environnement d'enregistrement.
- L'état du locuteur.
- …etc.

Comme solution au problème des classes qui ne sont pas facilement séparables, nous pouvons augmenter le nombre d'exemples de l'apprentissage et l'extraction de caractéristiques de la parole.

Conclusion et perspectives

Dans ce mémoire nous présentons les technologies basiques intervenants dans la réalisation d'un système de dicter basée sur un système de reconnaissance de la parole en temps réelle de mots arabe isolées ; l'analyse acoustique du signal par le traitement du signal (LPC) et une méthode de classification SVM (Support Vector Machine), méthode de classification binaire inspiré de la théorie d'apprentissage statistique.

Plusieurs ambiguïtés ont été rencontrées durant notre étude, parmi les quelles nous citons :

- Les conditions d'enregistrement ne répondent pas aux contraintes d'application (bruit, le matérielle, position et sensibilité du microphone…)
- Les états variés des locuteurs (le tempérament du locuteur, état émotif, état de fatigue…). Ces conditions ont une influence sur les résultats obtenus.
- L'outil capteur utilisé n'est pas vraiment fiable.
- La diversité des notions liées au concept de la parole(la reconnaissance de mots prononcés, La dictée vocale , La différentiation entre locuteur masculin, féminin et enfant , la dépendance ou non dépendance du texte…. etc..).
- Le choix de la méthode d'extraction des caractéristiques vocales qui convient au modèle de classification choisi.
- La difficulté d'obtention de caractéristiques vocales différenciables.
- La méthode SVM nécessite des paramètres obtenus par expérience.

Les résultats obtenus sont acceptables, il faut appliquer cette étude dans des circonstances meilleures.

Comme perspective, il est préférable d'utiliser la méthode d'extraction de paramètres MFCC (Mel Frequency Cepstrum Coefficients) au lieu de LPC parcequ'elle est plus robuste dans les environnement bruités. L'utilisation du modèle phonétique au lieu du modèle de mots isolés est beaucoup plus mieux pour minimiser le temps d'apprentissage et le nombre d'exemples.

Bibliographie

[1] **René Boite et Murat Kunt**
« *Traitement de la parole* ».
Presses polytechniques romandes, Lausanne.
1987.

[2] **Maurice Bellanger**
« *Traitement numérique du signal théorie et pratique* ».
Dunod, Paris.
1998-2002.

[3] **Othmani.C et Mazouzi.M**
« *Conception et réalisation d'un système de reconnaissance de locuteur par réseau de neurones artificiels* **».**
Mémoire de fin d'étude en vue de l'obtention du diplôme d'ingénieur en Informatique, Biskra.
Session 2005.

[4] **Rodolphe BATTAULT**
« *La reconnaissance vocale, techniques utilisées, applications actuelles et Futures* »**.**
Examen probatoire pour l'obtention du diplôme d'ingénieur du C.N.A.M, Paris.
1998.

.

[5] http://perso.orange.fr/xcotton/electron/coursetdocs.htm

[6] **José Anibal ARIAS AGUILAR**
« *Méthodes à vecteurs de support et Indexation Sonore* ».
Laboratoire IRIT (**I**nstitut de **R**echerche en **I**nformatique de **T**oulouse).
Année 2003-2004.

[7] http://membres.lycos.fr/guillaumerey/reconnaissance_principes.htm

[8] **Marc Sebban et Gilles Venturini**
« *Apprentissage automatique* ».
Hermes Science Publications, Paris.
1999.

[9] **Jeremy Mary**
« *Méthodes d'apprentissage avancées* ».
***C**entre **N**ational de la **R**echerche **S**cientifique*.
janvier2006.

[10] **Pierre Mahé et Laure Ait-Ali**
« *Projet d'apprentissage statistique SVM pour l'apprentissage non Supervisé* ».
DEA MVA.
Février 2003.

[11] **Antoine Cornuéjols**
« *Une nouvelle méthode d'apprentissage : les SVM. Séparateur à vaste marge* ».
Juin 2002**.**

[12] **Pierre Mahé**
« *Noyaux pour graphes et Support Vector Machines pour le criblage virtuel de molécules* ».
Septembre 2003.

[13] **Pascal Vincent**
« *Modèles à noyau à structure Locale* ».
Thèse présentée à la faculté des études supérieures en de l'obtention du grade de Philosophiæ.
Octobre 2003.

[14] **Olivier Bousquet**
« *Introduction aux Support Vector Machines* ».
Centre de Mathématiques Appliquées, Ecole Polytechniques, Palaiseau.
Novembre 2001.

[15] **Jérôme CALLUT**
« *Implémentation efficace des Support Vector Machines pour la Classification* ».
Mémoire présenté en vue de l'obtention du grade de maître en informatique.
2002-2003.

[16] **Mohamadally Hasan** et **Fomani Boris**
« *SVM : Machines à Vecteurs de Support ou Séparateurs à Vastes Marges* ».
janvier 2006.

[17] **Anderzej Drigajlo**
Traitement de la parole. *«Speech Processing and Biometrics Group (GTPB)»*.
juin 2006.

[18] **Thomas Styger, Bernard Gabioud, Eric Keller.**
« *Méthodes informatiques pour l'analyse de paramètres primaires en parole pathologique* ».
1993.

[19] **Guy Almouzni**
« *Traitement de la parole* ».
Cours et Tps.
2006-2007.

[20] **BENAMMAR Ryadh**
«Traitement Automatique De La Parole Arabe Par Les HMMs : Calculatrice Vocale».
Septembre 2012.

[21] **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 sciences spécialité Informatique
2011-2012.