N° d'ordre : 04 / 2010-M / EL

**République Algérienne Démocratique Et Populaire**
**Ministère de l'Enseignement Supérieur et de la Recherche Scientifique**

**Université des Sciences et de la Technologie Houari Boumediene**

Faculté d'Electronique et d'Informatique

جامعة هواري بومدين
للعلوم والتكنولوجيا
U S T H B

MEMOIRE
Présenté pour l'obtention du diplôme de MAGISTER
En : ELECTRONIQUE
Spécialité : **Communication Parlée**
Par
**ASBAI Nassim**

# Thème

# Reconnaissance Automatique de Locuteurs en Environnement Bruité par les Méthodes à Noyaux

Soutenu publiquement le 09 / 12 / 2010, devant le jury composé de :

| | | | |
|---|---|---|---|
| Mme R.TOUHAMI | Professeur | à l'USTHB | Présidente |
| Mr A. AMROUCHE | Maitre de Conférences-A | à l'USTHB | Directeur de mémoire |
| Mr Y.CHIBANI | Professeur | à l'USTHB | Examinateur |
| Mr B.FERGANI | Maitre de conférences-A | à l'USTHB | Examinateur |
| Mr H.TEFFAHI | Maitre de conférences-A | à l'USTHB | Examinateur |

# *Remerciements*

*Je tiens d'abord à adresser mes plus vifs remerciements à Monsieur AMROUCHE Abderahmane Maitre de conférences classe A à la faculté d'électronique et d'informatique (FEI), USTHB, pour son aide et l'intérêt avec lequel il a suivi mon travail.*

*Je tiens à remercier Mme TOUHAMI Rachida, Professeur à la Faculté d'Electronique et d'Informatique, USTHB, pour l'honneur qu'elle me fait de présider ce Jury.*

*Je voudrai également exprimer mes remerciements à Mr CHIBANI Youcef, Professeur à la Faculté d'Electronique et d'Informatique, USTHB, Mr FERGANI Belkaceme, Maître de Conférences à la Faculté d'Electronique et d'Informatique, USTHB et Mr TEFFAHI Houcine, Maître de Conférences à la Faculté d'Electronique et d'Informatique, USTHB pour avoir bien voulu accepter de faire partie ce Jury.*

*Mes remerciements vont aussi à Monsieur DEBYECHE Mohamed, Directeur du Laboratoire de Communication Parlée et Traitement du Signal et Maitre de conférences à la Faculté d'Electronique et d'Informatique pour m'avoir bien accueilli au sein du laboratoire LCPTS.*

*Je ne saurais oublier ma famille qui m'a été d'un grand soutien et qui n'a ménagé ni sa patience ni ses encouragements pour que ce travail puisse un jour aboutir.*

*Je tiens à exprimer toute ma gratitude à tous ceux qui ont contribué, de près ou de loin, à la concrétisation de ce travail*

# Résumé

Dans le domaine de la reconnaissance vocale, la biométrie vocale est une technique émergente utilisant les techniques de reconnaissance automatique de locuteurs (RAL). Cette dernière exploite la variabilité interlocuteurs et s'intéresse aux informations extralinguistiques du signal vocal.

Les applications sont nombreuses, notamment dans le contrôle d'accès, dans les institutions financières ou le donneur d'ordre doit être identifié et dans le domaine sécuritaire ou judiciaire. En effet, les applications en sciences criminalistiques sont de plus en plus évoquées, à tels point que de nombreux travaux trouvent leur prolongement dans les sciences forensiques.

Les systèmes de RAL résultent de la combinaison des techniques de traitement du signal nécessaires à l'extraction des paramètres acoustiques et de modèles issus de la reconnaissance des formes pour la discrimination entre locuteurs. Les méthodes actuellement utilisées sont fondées sur les modèles statistiques : mélange de gaussiennes GMM, HMM, SOSM ou neuronaux tels que les ANNs. Ces systèmes ont montré leur efficacité en environnement calme ou peu perturbé (absence de bruits de fond). Cependant , leurs performances se dégradent fortement en environnement réel, notamment pour certains bruits audio de large spectre.

Notre travail consiste à introduire les méthodes à noyaux dans le développement d'un Système de Reconnaissance Automatique de locuteurs, en particulier les supports à vecteur de machines SVMs. Les Support Vector Machines (SVM) permettent de projeter les données de l'espace d'entrée (appartenant à deux classes différentes) non-linéairement séparables dans un espace de plus grande dimension appelé espace de caractéristiques de façon à ce que les données deviennent linéairement séparables.

A travers une étude comparative, incluant les GMM et les ANNs, portant sur la reconnaissance de locuteurs en mode indépendant du texte, en ensemble ouvert (vérification) et fermé (identification) en environnement bruité, nous montrons que les systèmes de RAL basés sur les SVMs présentent de meilleurs résultats. Les vecteurs acoustiques d'entrées, extraits de la base de données ARADIGITS, sont constitués par les coefficients MFCC et leurs dérivées secondes et premières, les LSF et la fusion de ces paramètres pour évaluer l'apport de la coopération de connaissance.

L'extension de notre étude à la reconnaissance en milieu fortement bruité, faisant appel à quatre types de bruits additifs (produits par le chahut dans une cantine, par un véhicule militaire, dans une usine de production de véhicule, par un avion de combat) atteignant de forts niveaux SPL, tirés de la base de données NOISEX-92 (NATO : AC 243/RSG 10) confirme la supériorité de l'approche retenue basée sur les méthodes à noyaux de type SVM proposée dans ce mémoire.

**Mots clés :**Reconnaissance Automatique du Locuteur, IAL, VAL, LSF, MFCC, SVM, ANN, GRNN, GMM.

# Abstract

In the field of speech recognition, voice biometrics is an emerging technology using techniques of Automatic Recognition of Speakers (ARS). The latter, exploits the variability information and interested to extralinguistic speech signal: it is therefore to recognize a person from his voice. The Automatic Speaker Identification (ASI) and Automatic Speaker Verification (ASV) are the two most common tasks in the field of ARS.

The applications are numerous, including access control, in financial institutions or the customer must be identified and in the security field or court. Indeed, applications in forensic science are increasingly referred such an extent that many works found their continuation in the forensic sciences.

ARS systems resulting from the combination of signal processing techniques for extracting the acoustic parameters and models from pattern recognition to discriminate between speakers. The methods currently used are based on statistical models: Gaussian mixture GMM, HMM, neural such as ANNs or SOSM. These systems have proven effective in quiet or undisturbed (no background noise). However, their performances degrade significantly in real environments, including some audio noise spectrum.

Our work is to introduce the kernel methods in the development of a system of automatic speakers recognition, especially the Support Vector Machines SVMs. Support Vector Machines (SVM) are used to project the data from the input space (belonging to two different classes) non-linearly separable in a larger space called feature space so that the data become linearly separable.

Through a comparative study, including the GMM and the ANNs, on the recognition of speakers in the text independent mode, in open set (verification) and closed (identification) in quiet and noisy environment, we show that ARS systems based on SVMs show better results.

The input acoustic vectors extracted from the database ARADIGITS, consist of the MFCC coefficients and their first and second derivatives, the LSFs and the fusion of these parameters to evaluate the contribution of knowledge cooperation.

Extending our study to the recognition in very noisy environment, using four types of additive noise (produced by the ruckus in a canteen, a military vehicle in a factory production vehicle, a jet Fighter) reaching high levels of SPL, derived from the database Noisex-92 (NATO: AC 243/RSG 10) confirms the superiority of the approach based on kernel methods like SVM proposed in this paper.

**Keywords:** Automatic speaker recognition, ASI, ASV, LSF, MFCC, SVM, ANN, GRNN, GMM.

# Sommaire :

# Liste des Figures

# Liste des tableaux

# Abréviations

| | |
|---|---|
| DTW | Dynamic Time Warping |
| GMM | Gaussian Mixture Models |
| HMM | Hidden Markov Models |
| HO | Hyperplan optimal |
| IAL | Identification Automatique du Locuteur |
| LPCC | Linear Prediction Cepstral Coefficients |
| MAP | Maximum A Posteriori |
| MFCC | Mel Frequency Cepstral Coefficients |
| LSF | Line Spectral Frequency |
| MLP | Multi-Layer Perceptrons |
| RAL | Reconnaissance Automatique du Locuteur |
| RAP | Reconnaissance Automatique de la Parole |
| RdF | Reconnaissance de Formes |
| RN | Réseaux de Neurones |
| SVM | Support Vector Machines |
| VAL | Vérification Automatique du Locuteur |
| VQ | Vector Quantization |
| KNN | K-plus proches voisins |
| GRNN | General Regression Neural Networks |

# Introduction générale

La reconnaissance automatique du locuteur (RAL) est une tache de reconnaissance de formes. Ce domaine regroupe les problèmes relatifs à L'identification Automatique du Locuteur (IAL) et la Vérification Automatique du Locuteur (VAL) sur la base de l'information contenue dans le signal acoustique : il s'agit de reconnaitre une personne à partir de sa voix.

Le champ d'application de la reconnaissance vocale, qui pourrait constituer un élément important de la biométrie humaine à travers la signature vocale et ses applications, est très large, allant des applications domestiques aux applications militaires, en passant par les applications judiciaires.

En effet, la biométrie vocale, technique émergeante, est appelée à jouer un rôle important dans les sciences criminalistiques, en complément des autres modalités reposant sur l'image telle que les empreintes digitales ou l'iris, pour apporter la matérialisation des faits pouvant aider la justice. C'est ce qui explique l'intérêt grandissant manifesté par toutes les sphères chargées du domaine sécuritaire de par le monde.

Par ailleurs, l'émergence de la reconnaissance vocale dans les réseaux de communication est née du besoin d'éviter les impostures dans certains domaines sensibles. Les personnalités occupant des responsabilités stratégiques, les transactions boursières, l'accès sécurisé restrictif, etc., nécessitent l'authentification ou la vérification du donneur d'ordre distant.

Aussi, depuis plusieurs années, de nombreux laboratoires internationaux ont mené des recherches intensives dans ce domaine et des progrès importants ont été réalisés, notamment grâce au développement d'algorithmes puissants, alliés aux technologies de traitement numérique du signal. De nombreux systèmes de reconnaissance du locuteur sont maintenant disponibles ou peuvent être développés.

Sur le plan méthodologique et scientifique, de nombreuses questions peuvent être abordées ;

1- Le choix des techniques d'extraction des paramètres au niveau du front –end.

2- Choix de la technique de reconnaissance .

3- Efficacité des reconnaisseurs, notamment en environnement acoustique adverse (bruité).

4- Comment évaluer les systèmes mis en œuvre?

Enfin, la problématique de la langue Arabe, usitée par plus de 300 millions de personnes à travers la planète, qui demeure peu dotée dans les technologies de la langue, nous interpelle et nous dicte le choix du corpus d'étude dans ce travail.

Dans ce mémoire, nous discuterons donc des outils théoriques qui sont à la base de la plupart des systèmes de reconnaissance du locuteur. Ceux-ci résultent de modèles statistiques puissants dont les paramètres peuvent être estimés automatiquement sur la base d'un grand ensemble d'entraînement.

Beaucoup d'outils et de connaissances relatives au mécanisme de reconnaissance du locuteur sont maintenant disponibles. Les systèmes actuels sont basés sur l'approche statistique utilisant les mélanges de gaussiennes (GMMs) sont dominants, alors que d'autres utilisent les modèles de Markov cachés (HMM) ou l'approche neuronale avec les ANNs . Une expérience importante a été acquise concernant la mise en œuvre de ces algorithmes, leur comportement dans les milieux réels, qui montre notamment leur manque de robustesse aux environnements bruités.

Dans notre travail, nous nous focaliseront sur les outils théoriques des méthodes dites à noyaux qui feront l'objet de ce mémoire telles que les mélanges de gaussiennes GMM [**1**], les réseaux de neurones [**2**],[**3**] et les machines à vecteurs de support SVMs [**4**]. Chaque algorithme sera étudié dans ce mémoire afin de pouvoir le mettre en œuvre pour la reconnaissance automatique du locuteur en mode indépendant du texte et en milieu bruité.

Dans la première partie de ce manuscrit, nous présenterons l'historique de la RAL ainsi que les motivations qui nous ont amené à aborder ce thème. Ensuite, nous feront une description du signal de parole produit par l'appareil phonatoire de l'être humain avec un rappel sur ses caractérisations.

Une connaissance approfondie des techniques d'analyse du signal de parole aidera à la paramétrisation de ce dernier. Dans le deusième chapitre, nous nous intéresserons aux méthodes de reconnaissance automatique du locuteur alors que dans le troisième chapitre nous détaillerons les méthodes à noyaux que nous avons utilisées dans la phase d'apprentissage ou de classification utilisées dans la tache de reconnaissance automatique du locuteur. Dans le dernier chapitre, nous présenterons le système de reconnaissance de locuteurs que nous avons élaboré et qui repose sur un classificateur basé sur les méthodes à noyaux de type SVM, ainsi qu'une étude comparative avec les systèmes similaires mis en œuvre basés sur les GMM et ANNs de type MLP et GRNN.

L'application de tels systèmes n'a d'intérêt que s'ils sont utilisés dans un milieu naturel réel, donc forcement pollué par les nuisances sonores. Dans notre étude, l'efficacité de la reconnaissance a été évaluée dans des environnements acoustiques hostiles à l'aide de la base de données NOISE'92 NATO… . Des discussions porteront sur l'interprétation des résultats obtenus.

Enfin une conclusion incluant les perspectives ouvertes par ce travail, ainsi que les principales références bibliographiques utilisées et des annexes termineront ce mémoire.

# I Caractéristiques et Analyse du Signal de Parole

# Chapitre I :

# Caractéristiques et Analyse du Signal de Parole

### I.1 Eta de l'art sur la RAL :

La reconnaissance automatique du locuteur s'inscrit dans le domaine plus général du traitement de la parole. Elle exploite la variabilité interlocuteurs et s'intéresse aux informations extralinguistiques du signal vocal. Les variations individuelles entre locuteurs ont deux origines essentielles. D'abord, les caractéristiques morphologiques de l'appareil phonatoire sont différentes pour chaque locuteur, indépendamment de la phrase prononcée. Ensuite, une même phrase n'est pas prononcée de la même façon par deux locuteurs, en regard de la variabilité inter et intra locuteur. Cette variabilité est l'essence même de la reconnaissance automatique du locuteur. Comme dans le cas de la reconnaissance de la parole, le problème de reconnaissance du locuteur peut se formuler selon un problème de classification.

Les travaux sur la reconnaissance vocale (parole et locuteur) datent du début du XX$^{e}$ siècle. Le premier système pouvant être considéré comme faisant de la reconnaissance vocale a été développé par Davis, Biddulph, and Balashek aux laboratoires Bell Labs en 1952. Ce système électronique était essentiellement composé de relais et ses performances se limitaient à reconnaître des chiffres isolés. La recherche s'est ensuite considérablement accrue durant les années 1970 avec les travaux de Jelinek chez IBM (1972-1993). Aujourd'hui, la reconnaissance vocale en particulier la RAL, est un domaine à forte croissance grâce à l'émergence d'applications notamment en biométrie vocale avec son corolaire les sciences forensiques , les contrôles d'accès sécurisés nécessitant la signature vocale, les applications dans les communications mobiles et filaires a déferlante des systèmes embarqués.

Depuis les premier travaux dédiés à la RAL, de nombreuses approches ont été proposées dans la littératures à savoir les approches vectorielle, statistique, prédictive et connexionniste.

Vers les années 60 (1967) fut apparue une approche de classification utilisée dans la RAL, dite K-plus proches voisins (KNN) publiée par Cover et Hart. Cette méthode discriminante de base appartient à la catégorie des algorithmes graphiques et ne comporte pas d'étape d'apprentissage à proprement parler.

Sur une mesure de distance arbitraire entre les vecteurs. En phase de test, les distances entre le vecteur à classer et tous les vecteurs d'apprentissage sont estimées et rangées en ordre décroissant. Pour la décision, on procède par vote majoritaire parmi les k vecteurs d'apprentissage les plus proches. Il peut arriver que deux classes majoritaires aient le même nombre de plus proches voisins. Pour résoudre ce conflit, plusieurs stratégies sont envisageables, comme par exemple choisir la classe ayant la distance moyenne la plus faible. Notons enfin que la capacité de généralisation de la modélisation est réglée via le paramètre k. Outre sa simplicité, l'avantage de cette méthode est qu'elle peut naturellement s'appliquer au cas multi-classes même avec un nombre élevé de classes [**5**]. Mais les inconvénients sont de taille :

1. Un volume important de données d'apprentissage implique une capacité des ressources mémoire nécessaires d'autant plus élevée, ainsi qu'une forte complexité calculatoire en phase de test.
2. Le renvoi d'une mesure de confiance de la décision (score) ne peut se faire que de manière arbitraire, par exemple en calculant une moyenne des distances au k-plus proches voisins à partir des distances calculées. A la base, la méthode est conçue pour renvoyer une décision binaire.

A partir du milieu des années 1970,une des premières applications de la MMC a été la reconnaissance vocale. Le Modèle de Markov Caché (Hidden Markov Model) est une méthode statistique puissante pour caractériser les échantillons de données observés d'un processus à temps discret [6]. Elle apporte un moyen efficace de construction de modèles paramétriques. Dans la modélisation d'un processus par un HMM, les échantillons peuvent être caractérisés par un processus paramétrique aléatoire dont les paramètres peuvent être estimés suivant un modèle à plusieurs états d'après L. Baum [**7**]. Les HMMs sont devenus la méthode la plus couramment utilisée pour la modélisation des signaux de parole dans les applications suivantes :

reconnaissance automatique de la parole, suivi de la fréquence fondamentale et des formants, synthèse vocale, traduction automatique, étiquetage syntaxique,
compréhension du langage oral, traduction automatique et reconnaissance du locuteur. Dans une chaîne de Markov, chaque état correspond à un événement à observation déterministe [**8**][**9**].

Une extension naturelle à la chaîne de Markov introduit un processus non déterministe qui génère des symboles de sortie pour chaque état. L'observation est donc une fonction probabiliste de l'état [**10**].

Le nouveau modèle est appelé HMM, pouvant être vu comme deux processus stochastiques imbriqués dont l'un (la séquence d'états) est non observable directement. Ce processus sous-jacent est donc associé de façon probabiliste à un autre processus produisant la séquence de trames, qui elle, est observable.

Peu après et dans les années 80 apparue une autre méthode dite Dynamic Time Warping dans le domaine du traitement de la parole et encore utilisée dans des systèmes de reconnaissance de locuteurs disposant de ressources matérielles limitées. Dans les systèmes de reconnaissance basés sur la DTW, chaque locuteur est représenté par une réalisation de référence. Le processus de reconnaissance consiste à évaluer la distance d'une observation à chacune des références. Toute la difficulté du décodage réside dans cette mesure d'un degré de similarité entre des formes acoustiques variables à la fois au niveau spectral et temporel.

En effet, les réalisations acoustiques représentant un locuteur subissent des déformations spectrales liées à divers paramètres (locuteurs, contextes, conditions d'acquisition, etc.) mais aussi des déformations temporelles globales (vitesse d'élocution) ou plus locales (accent, dynamique des organes phonatoires, etc.). Pour comparer deux segments de parole soumis à cette double déformation, il faut préalablement leur appliquer un processus d'alignement temporel. L'algorithme DTW (Dynamic Time Warping) réalise cet alignement en recherchant, parmi tous les alignements possibles, celui qui minimise une fonction de coût intégrant l'écart spectral des données alignées et un coût de distorsion temporelle [**11**], [**12**], [**13**]. La distance retenue est celle correspondant à l'alignement de coût minimal.

Dans les années 90, un engouement pour les méthodes connexionnistes a débouché sur leurs applications dans le domaine de la parole.

Depuis, les réseaux neuromimétiques constituent une technique utilisée dans les systèmes de reconnaissance automatique de la Parole et de locuteurs [**2**] [**3**]. Ils sont basés sur une modélisation mathématique du neurone biologique ou neurone formel. Dans ce modèle, le neurone formel calcule son activation en fonction des signaux qu'il reçoit d'autres neurones, pondérés par des « poids synaptiques » et une fonction d'activation plus ou moins complexe. L'ensemble de ces neurones est organisé selon des architectures diverses suivant la complexité de problème à modéliser.

Quelques années après, furent apparus ce qu'on a appelé Les séparateurs à vastes marges qui reposent sur deux idées clés : la notion de marge maximale et la notion de fonction noyau. Ces deux notions existaient depuis plusieurs années avant qu'elles ne soient mises en commun pour construire les SVM. L'idée des hyperplans à marge maximale a été explorée dès 1963 par Vladimir Vapnik et A. Lerner[**14**], et en 1973 par Richard Duda et Peter Hart dans leur livre Pattern Classification[**15**] . Les fondations théoriques des SVM ont été explorés par Vapnik et ses collègues dans les années 70 avec le développement de la Théorie de Vapnik-Chervonenkis, et par Valiant et la théorie de l'apprentissage . L'idée des fonctions noyaux n'est pas non plus nouvelle: le théorème de Mercer date de 1909 [**16**], et l'utilité des fonctions noyaux dans le contexte de l'apprentissage artificiel a été montré dès 1964 par Aizermann, Bravermann et Rozoener[**17**]. Ce n'est toutefois qu'en 1992 que ces idées seront bien comprises et rassemblées par Boser, Guyon et Vapnik dans un article, qui est l'article fondateur des séparateurs à vaste marge[**18**]. L'idée des variables ressorts, qui permet de résoudre certaines limitations pratiques importantes, ne sera introduite qu'en 1995. À partir de cette date, qui correspond à la publication du livre de Vapnik [**19**], les SVM gagnent en popularité et sont utilisés dans de nombreuses applications.

Et enfin, l'utilisation des GMMs pour la modélisation des locuteurs a été initiée par les travaux de thèse de Douglas Reynolds [**1**], cette approche a donné, depuis plus de10 ans maintenant, les meilleures performances pour les systèmes de reconnaissance du locuteur en mode indépendant du texte basé sur l'approche probabiliste. La plupart des systèmes actuels utilisent une modélisation de locuteurs par GMM[**20**].

Les modèles de Mélange de lois Gaussiennes (GMM : Gaussian Mixture Models, en anglais) ont étés utilisés dans de nombreux domaines, par exemple pour le traitement et la reconnaissance des images ou de la parole. Dans le cadre de la reconnaissance du locuteur, un GMM modélise un locuteur donné par une somme pondérée de gaussiennes. On peut assimiler un modèle de GMM à un modèle de Markov cachés (HMM : Hidden Markove Model, en anglais) à un seul état. On ne modélise donc pas les aspects temporels du signal. Cette méthode est plus utilisée en ce qui concerne la reconnaissance du locuteur en mode indépendant du texte.

**Limitations des méthodes actuelles et motivations :**

Parmi les limitations pratiques des méthodes citées ci-dessus : la complexité calculatoire pour les KNN , la lenteur en phase d'apprentissage pour le réseau de neurones (RN) ,restriction aux modèles de Marcov d'ordre 1 rendant la modélisation d'apprentissage de dépendances à long terme (Calcul de vecteur moyenne et matrice de covariance) difficile pour les HMM et restriction aux estimations de distributions seulement de formes gaussiennes ou multi-gaussienne pour les GMMs . Cependant, ces méthodes ont montré leurs efficacités en un milieu calme, par contre leurs performances se dégradent fortement dans des milieux adverses (milieux réels ou bruités). C'est la raison pour la quelle, les travaux de recherche actuellement en cours s'orientent vers l'utilisation de méthodes de reconnaissance robustes et plus rapides. C'est dans ce cadre que nous avons mené ce travail basé sur l'utilisation de méthodes à noyaux en particulier les SVMs [**20**][**21**][**22**][**23**].

## I.2 Signal de parole :

La parole est en effet produite par l'appareil phonatoire, décrit par la figure Fig.I. 1, contrôlé en permanence par le cortex moteur. L'étude des mécanismes de phonation permettra donc de déterminer, dans une certaine mesure la parole. Dans le processus de communication parlée, incluant un locuteur et un auditeur, les propriétés de production et de perception sont réunies grâce un feed-back (boucle de retour) chez le locuteur.

Les techniques de traitement de la parole, et plus particulièrement des langues naturelles, tendent à reproduire des systèmes automatiques qui se substituent à l'une ou l'autre de ces fonctions :

- L'analyse de parole cherche à mettre en évidence les caractéristiques du signal vocal tel qu'il est produit, ou parfois tel qu'il est perçu (on parle alors d'analyse perceptuel)..

Les analyseurs de parole sont utilisés soit comme composant de base de systèmes de codage, de reconnaissance ou de synthèse, soit en tant que tels pour des applications spécialisées, comme l'aide au diagnostic médical (par exemple les pathologies du larynx) ou l'étude des langues.

- La reconnaissance a pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse. On distingue fondamentalement deux types de reconnaissance, en fonction de l'information que l'on cherche à extraire du signal vocal : la reconnaissance du locuteur, dont l'objectif est reconnaître la personne qui parle (individu) indépendamment de message produit, et la reconnaissance de la parole, ou l'on s'attache plutôt à reconnaître ce qui est dit (le message).

## I.3 La production de la parole :

Au contraire des acousticiens, ce n'est pas tant le signal qui intéresse les phonéticiens que la façon dont il est produit par le système articulatoire (présenté a la figure Fig.I. 1) et aperçu par le système auditif.

### I.3.1 La phonation :

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations.

L'appareil respiratoire fournit l'énergie nécessaire à la production de son, en poussant l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le larynx ou la pression de l'air est modulée avant d'être appliquée au conduit vocal.

Le **larynx** est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée. **Les cordes vocales** sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et en s'écartant progressivement, déterminer une ouverture triangulaire appelée **glotte**.

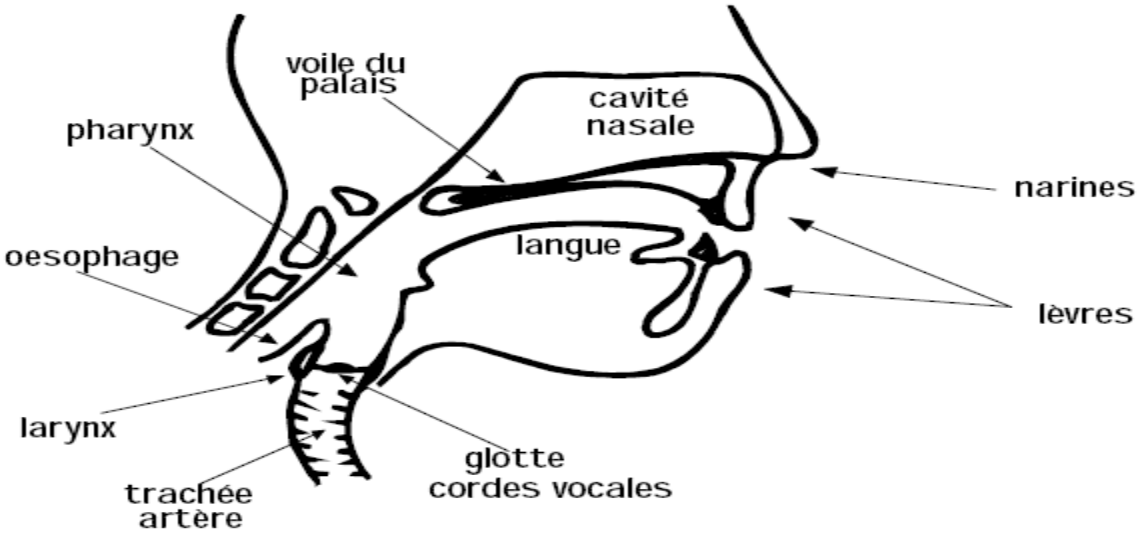


**Fig. I. 1 schéma général de l'appareil phonatoire (les poumons jouent le rôle de soufflerie alimentant le conduit vocal à travers la trachée artère).**

### I.3.2 Les différents sons produits par le système phonatoire :

#### I.3.2.1 Son voisé :

Un son voisé résulte de la vibration des cordes vocales suite à un passage de l'air phonatoire. Dans ce cas les cordes vocales sont tendues.

- ✓ Différents muscles et mécanismes (mâchoire, langue, luette, lèvres, bouche) modifient la configuration des cavités pour produire les différents types de sons voisés.
- ✓ Le flux d'air est découpé en un train d'impulsion quasi périodique qui résonne dans les différentes cavités : pharynx, bouche et optionnellement le nez (conduit nasal).
- ✓ Physiquement, le train d'impulsion quasi périodique subit une modulation en fréquence en passant par les différentes cavités.

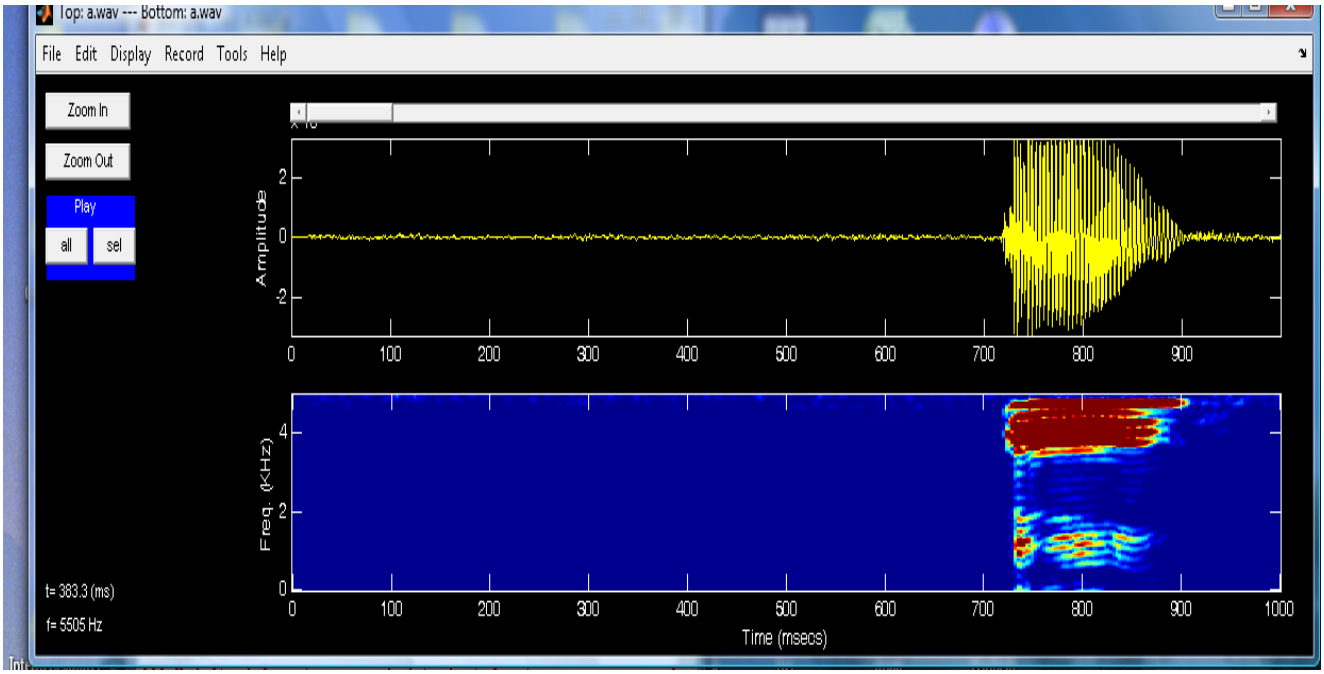


**Fig. I. 2 son voisé [a]**

### I.3.2.2 Son non voisé :

Lorsque les cordes vocales sont relâchées, l'air passe librement au niveau du larynx. La production des sons non voisés résulte du passage de cet air à travers le conduit vocal en fonction de la position des différent articulateurs : mâchoire, langue, luette, lèvres, bouche.

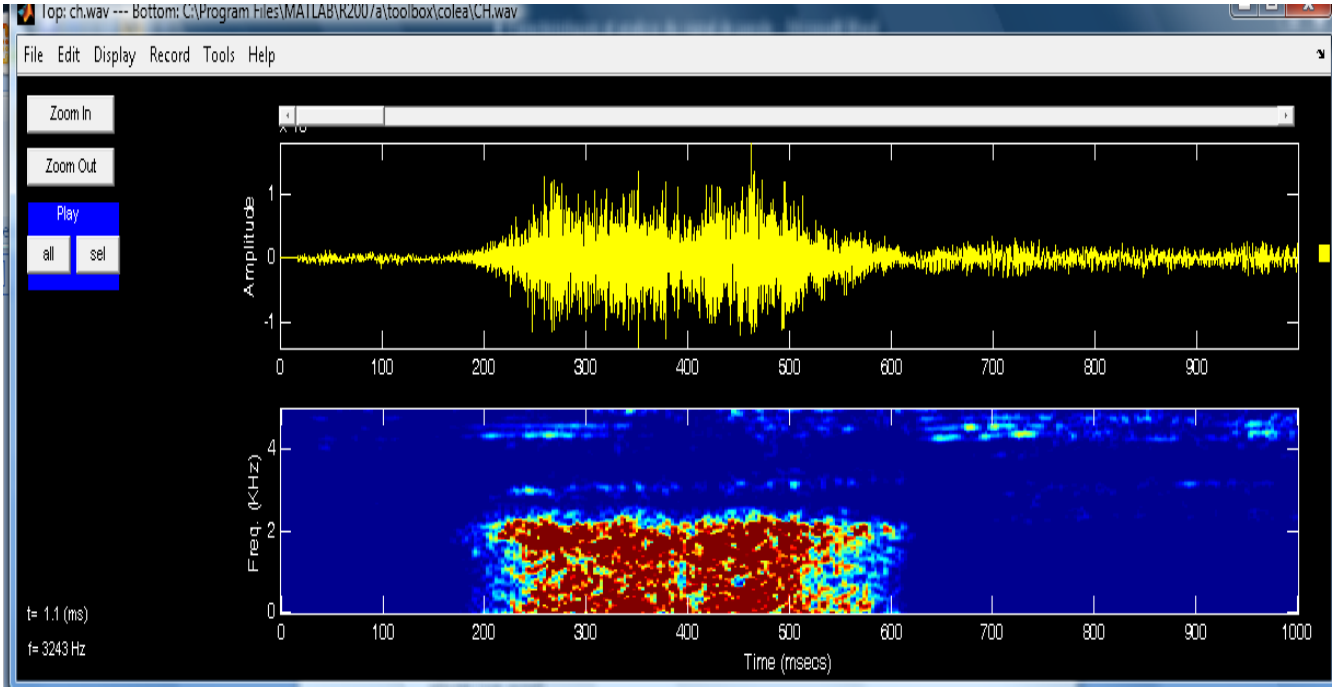


**Fig. I. 3 son non voisé [**CH**]**

### I. 3.2.3 Classification des sons :

Les états de l'appareil phonatoire déterminent les natures des sons produits, on distingue :

❖ **Voyelles :**

Les voyelles sont produites lorsque le conduit vocal est ouvert et les cordes vocales vibrent (son voisé).

Les voyelles sont **orales** ou **nasales** selon que la cavités nasale n'est pas ou est mise en parallèle à la cavité buccale.

❖ **consonnes :**

Des consonnes sont produites lorsqu'un rétrécissement apparaît dans l'appareil phonatoire :

- ✓ les cordes vocales peuvent vibrer ou laisser librement l'air (sons voisés et non voisés)
- ✓ les consonnes sont fricatives si le rétrécissement est partiel ou occlusives si une occlusion totale apparaît dans l'appareil phonatoire, causant une augmentation de la pression et un relâchement brutal de celle-ci.

- **Fricatives non voisés**
- **Fricatives voisés**
- **Occlusion non voisés**
- **Occlusion voisés**

### I. 4 Acoustique de l'audition :

L'acoustique de l'audition s'attache, entre autres, à la mesure des sensations perçues lors de l'audition. Les découvertes, expérimentales, ont permis d'établir les relations qui existent entre les grandeurs acoustiques (l'intensité, la fréquence, le spectre, le temps) et les sensations produites sur l'oreille lors de l'écoute (hauteur, intensité sonore subjective, timbre, durée).
La hauteur (ou tonie) d'un son pur est liée à la fréquence de l'onde sonore. Mais, au-delà de 1000 Hz, il faut plus que doubler la fréquence pour percevoir un doublement de la hauteur (il faut un signal de 3120 Hz pour avoir une sensation de 2000 Hz).
L'échelle de tonie est graduée en Mels. Un écart constant en Mels est perçu comme un écart constant en hauteur. Dans le domaine de l'audition, l'oreille est capable de discerner 1400 hauteurs distinctes.

L'unité permettant de mesurer le niveau de l'intensité subjective est le phone. Les courbes d'isophonies (courbes de Fletcher et Munson) montrent que la courbe de réponse de l'oreille dépend de la fréquence et de la pression sonore. Pour qu'un son pur de 100 Hz soit perçu avec la même intensité subjective qu'un son de 1000 Hz, sa pression sonore doit être plus élevée.
De même, à fréquence fixe, la hauteur perçue augmente avec l'intensité subjective. L'oreille peut distinguer 280 niveaux différents d'intensité.
Le timbre dépend de la répartition spectrale d'un son complexe: timbre clair pour une prédominance des fréquences hautes, timbre sombre pour une prédominance des basses fréquences.

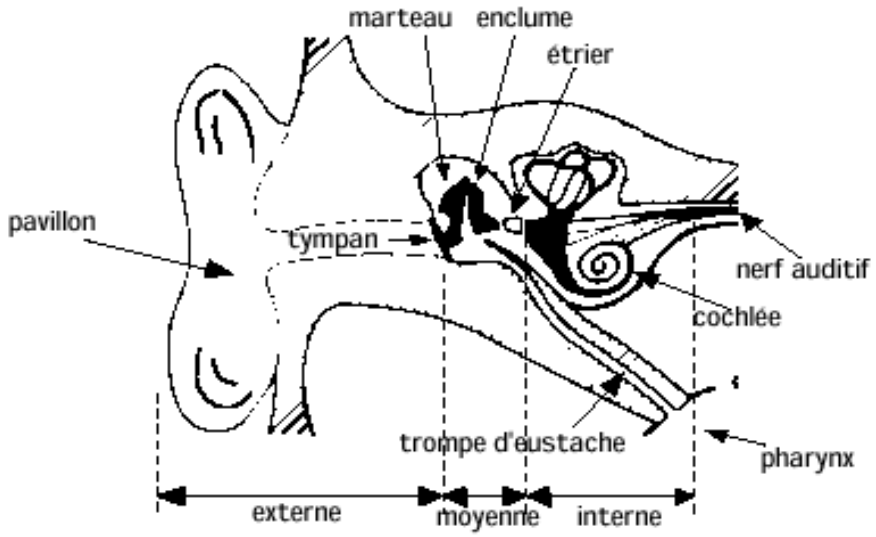


**Fig. I. 4 Schéma de l'appareil auditif**

**I. 5 Parametrisation du signal vocal :**

**I. 5.1 Pré-traitement :**

L'information portée par le signal de parole peut être analysée de plusieurs manières**.** On en rappelle simplement ici les aspects acoustiques. La parole, émise par le système articulatoire, apparaît physiquement comme une variation de la pression de l'air autour de la pression ambiante. Le signal acoustique est transformé dans un premier temps en signal électrique grâce à un transducteur approprié : le microphone, (le plus souvent associé à un préamplificateur). Le signal électrique résultant est ensuite numérisé. La figure Fig. I. 5 résume la chaine de traitement acoustique du signal de parole.

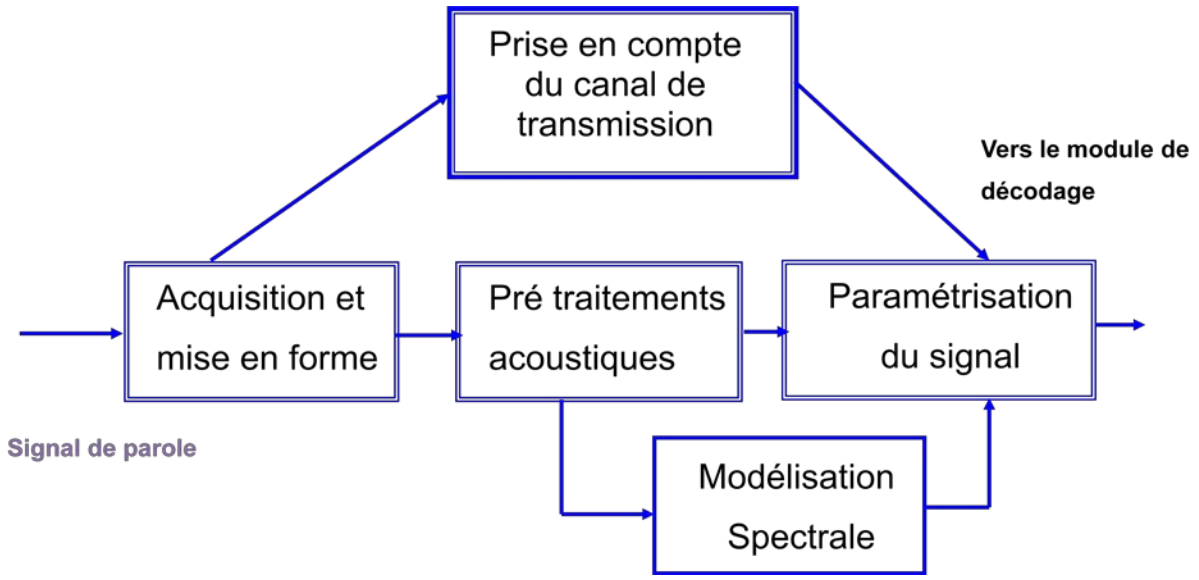


**Fig. I. 5 Chaine de traitement acoustique du signal de parole.**

### I. 5.1.2 Acquisition :

Le signal de parole est continu, ce qui rend son traitement par la machine difficile, on procède à une opération simple appelée : échantillonnage.

Il s'agit tout simplement de relever à chaque instant « T » le niveau énergétique du signal acoustique tout en respectant le théorème de Shannon.

- **Théorème de Shannon :**

La perte d'information entre le signal continu et le signal discret doit être nulle si et seulement si la fréquence d'échantillonnage, notée « $f_e$ », est supérieure ou égale à la fréquence maximum du spectre du signal notée « $f_{max}$. »

$$f_e \geq 2 f_{max} \qquad \text{avec } f_e = \frac{1}{T_e}$$

**I. 5.1.3. Préaccentuation :**

On remarque qu'au niveau du spectre de la parole, les basses fréquences sont favorisées par rapport aux hautes fréquences, car ce signal se caractérise par une pente globale négative de 6 dB/octave due aux influences de la source d'excitation et du rayonnement des lèvres.

Pour cela, on compense cette perte par un filtre appelé pré -accentuation (Preemphasis) qui a pour fonction de transfert :

$$H(z) = 1 - a \cdot z^{-1} \qquad \text{avec} \quad 0.95 < a < 1 \qquad \text{(I. 1)}$$

**I. 5.1.4 Fenêtrage :**

Il est difficile voire impossible de traiter un signal non stationnaire tel celui de la parole sans le fragmenter en trames. Une analyse à court terme montre que le signal vocal est quasi stationnaire sur des tranches temporelles de durées de 10 à 30 ms. Cette analyse est effectuée à l'aide de fenêtres telles que :

**Fenêtre Hamming** $$w_n = 0{,}54 - 0{,}46.\cos(2\pi \frac{n}{N-1}), \quad 0 \le n \le N-1 \qquad \text{(I. 2)}$$

avec : n : valeur d'échantillon à l'instant nTe.

**N :** la taille de la fenêtre.

Cette fenêtre de Hamming est souvent utilisée, vu que son spectre n'introduit pas trop de distorsion sur le signal vocal : l'atténuation du lobe principal par rapport aux lobes secondaires est de - 41db et la concentration de l'énergie du principal est de 99.96% .

**I. 6 Analyse spectrale :**

L'analyse spectrale présente des avantages au niveau de la perception car l'oreille humaine effectue une discrimination fréquentielle des sons. De plus, cette analyse fait apparaître des propriétés et des paramètres pertinent pour la suite du traitement. Les principaux outils utilisés sont les suivants :

**I. 6.1 La transformée de Fourier discrète:**

Pour effectuer cette analyse on utilise :

$$X(n) = \sum_{k=0}^{N-1} x(k) \times e^{-j\pi \frac{nk}{N}} \qquad \text{(I. 3 )}$$

Avec X(n ) le spectre du signal numérique x(k).

N : Le nombre d'échantillons de la trame.

n : Valeur d'un échantillon à l'instant nTe.

Ce qui nous donne le spectre fréquentiel du signal analysé.

**I. 7 Analyse temporelle :**

**I. 7.1 Energie totale :**

L'amplitude du signal de la parole varie au cours du temps selon le type de son, en particulier, l'amplitude des segments non voisés est généralement plus faible que celle des segments voisés. L'énergie à court terme du signal de la parole fournit une représentation convenable qui reflète ces variations d'amplitude.

Elle est calculée à partir de la relation suivante :

$$E=\frac{1}{N}\sum_{k=0}^{N-1}x^2(k)\,. \qquad \text{(I. 4)}$$

Avec *E :* la valeur à évaluer.

*N :* la largeur de la fenêtre d'analyse.

*x(k) :* le signal numérique.

La courbe d'énergie permet la distinction entre son voisé et non voisé.

**I. 7.2 Taux de passage par zéro :**

Le taux de passage par zéro (TPZ) est donné par l'expression suivante:

$$TPZ=\frac{1}{2}\sum_{k=0}^{k-1}\left|sign(x(k+1))-sign(x(k))\right| \qquad \text{(I. 5)}$$

Souvent le mot est constitué de segments voisés et d'autres non voisés, ces derniers sont caractérisés par une faible énergie.

Quand l'énergie du signal est faible, la TPZ permet de déceler l'existence d'une émission haute fréquence peu énergétique mais porteuse d'informations importantes, caractérisant par exemple les fricatives non voisées telles que les phonèmes /s/,/f/,/ch/..

**I. 8 Analyse homomorphique (Cepstre) :**

Le signal vocal x(n) est produit par un signal excitateur g(n), qui est la source glottique, traversant un système linéaire passif de réponse impulsionnelle h(n) qui représente leconduit vocal **[11].**

D'après cette hypothèse, tirée du concept source_filtre de G.Fant, on aura le système suivant:

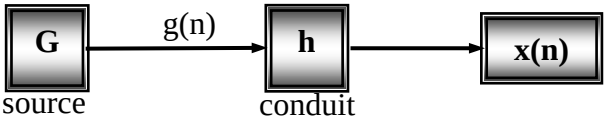


**Fig. I.6 Modèle source-filtre**

Donc on peut écrire pour tout n > 0 :

$$x(n) = g(n) * h(n) \tag{I. 6}$$

Pour déconvoluer x(n), c'est à dire pour retrouver les deux composantes g(n) et h(n), avec g(n) une séquence d'impulsions périodique pour les sons voisés, il suffit de transposer le problème par homomorphisme dans un espace où l'opérateur de convolution « * » correspond à un opérateur d'addition « + ».

Soit $D_*^+$ cet homomorphisme.

$D_*^+$ est un homomorphisme (application) qui applique l'espace vectoriel des signaux d'entrées muni de la loi de convolution « * », sur l'espace vectoriel des signaux de sortie muni de la loi d'addition « + » .

$$x(n)=g(n)*h(n) \longrightarrow \hat{x}(n)=\hat{g}(n)+\hat{h}(n)$$

L'intérêt de la méthode réside dans le fait que $\hat{g}(n)$ et $\hat{h}(n)$ sont facilement séparables par un filtrage temporel et ceci grâce à l'hypothèse simplificatrice sur g(n). Ce qui donne le système schématisé dans la figure suivante :

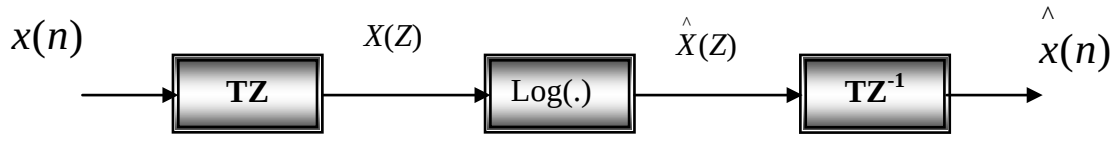


**Fig. I.7 Calcul du cepstre complexe**

TZ est la transformée en Z ($TZ^{-1}$ sa transformée inverse).

La fonction log est utilisée pour le passage du domaine de la loi «. »(La multiplication) au domaine de la loi « + » (l'addition), cette fonction n'est valable que pour les signaux positifs, toutefois, étant donné que la majorité des signaux courants sont bipolaires (positifs et négatifs), donc il faut faire appel à fonction log complexe.

Soit :

$$X(z)=\left|X(z)\right|\times\exp\left[jArg(X(z))\right] \qquad \text{(I. 7)}$$

donc :

$$\hat{X}(z)=\log\left[X(z)\right]=\log\left|X(z)\right|+jArg\left[X(z)\right]. \qquad \text{(I. 8)}$$

- La fonction exp doit être aussi la fonction exponentielle complexe.

D'après le schéma de la Fig. I. 7 on a :

$$X(Z)=TZ\left[x(n)\right] \qquad \text{(I. 9)}$$

$$\hat{X}(z)=\log\left[X(z)\right] \quad \text{(la fonction log est complexe)} \qquad \text{(I. 10)}$$

$$\hat{x}(n)=TZ^{-1}\left[X(Z)\right] \qquad \text{(I. 11)}$$

tous ceci peut être résumé par la notation suivante :

$$\hat{x}(n)=D_{*}^{+}\left[x(n)\right] \qquad \text{(I. 12)}$$

Le signal $\hat{x}(n)$ est appelé cepstre complexe associé au signal x(n).

**I. 8.1 Extraction des paramètres :**

Le signal de parole présente de la redondance et contient des informations jugées trop redondantes pour la reconnaissance, ce qui justifie la recherche d'une représentation spécifiquement pertinente.

L'extraction des paramètres du signal consiste à associer au signal de parole une série de vecteurs de paramètres acoustiques en suivant les étapes données dans la Fig. I. 8 .

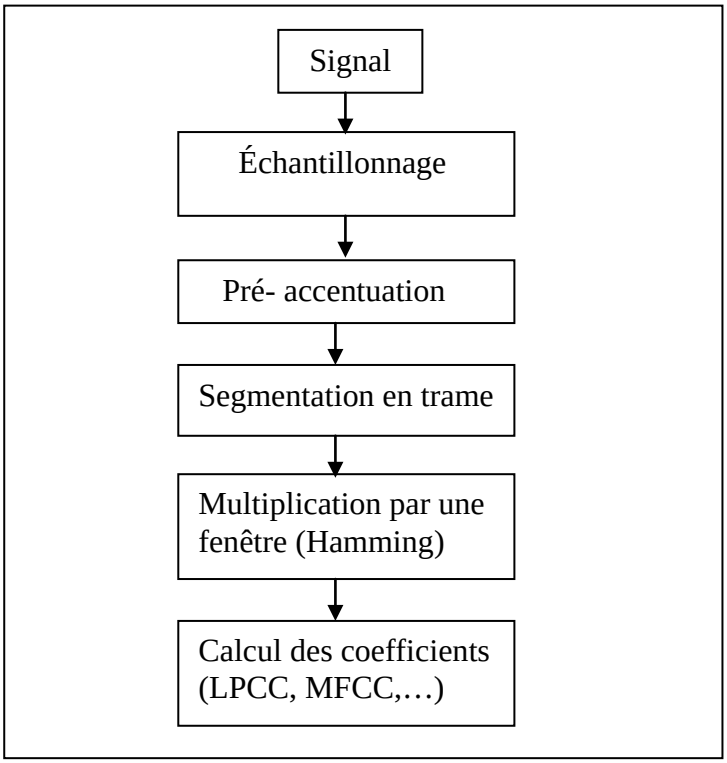


**Fig. I. 8 : Les étapes d'extraction des paramètres acoustiques.**

**- le calcul des coefficients :** il existe plusieurs types de coefficients avec lesquels le signal de parole est paramétré.
Les plus utilisés sont la coefficients LPC, LPCC (Linear Predictive Cepstral Coefficients), les coefficients PLP (Perceptual Linear Predictive) et les coeffficients MFCC (Mel Frequency Cepstral Coefficients)..

En reconnaissance du locuteur, les paramètres extraits doivent être :

- **pertinents** : extraits de mesures suffisamment fines, ils doivent être précis mais leur nombre doit rester raisonnable afin de ne pas avoir de coût de calcul trop important dans le module de décodage.

- **discriminants** : ils doivent donner une représentation caractéristique des sons de base et les rendre facilement séparables.

- **robustes** : ils ne doivent pas être trop sensibles à des variations de niveau sonore ou à un bruit de fond. Il existe dans la littérature différentes méthodes de paramétrisation du signal vocal.

– **Paramétrisation basée sur un modèle de production de la parole :** Cette méthode est basée sur les connaissances en production de la parole. La plus connue est l'analyse LPC (Linear Predictive Coding) dans laquelle le système de production de la parole est modélisé par un filtre Auto Régressif (AR).

### I. 8.1.1 Paramètres MFCC :

Les coefficients MFCC sont des coefficients cepstraux très souvent utilisés en reconnaissance automatique de la parole et du locuteur. Le codage MFCC utilise une échelle fréquentielle non-linéaire ou échelle Mel.
La fréquence Mel-échelle est définie par:

$$B(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad \text{(I. 13)}$$

Où $f$ est la fréquence en Hz, $B(f)$ est la fréquence Mel-échelle de $f$.
L'intérêt de l'échelle Mel est d'être assez proche d'échelles issues d'études sur la perception sonore et sur les bandes passantes critiques de l'oreille .Le calcul des paramètres MFCC se réalise de la façon suivante (Fig. I. 9):

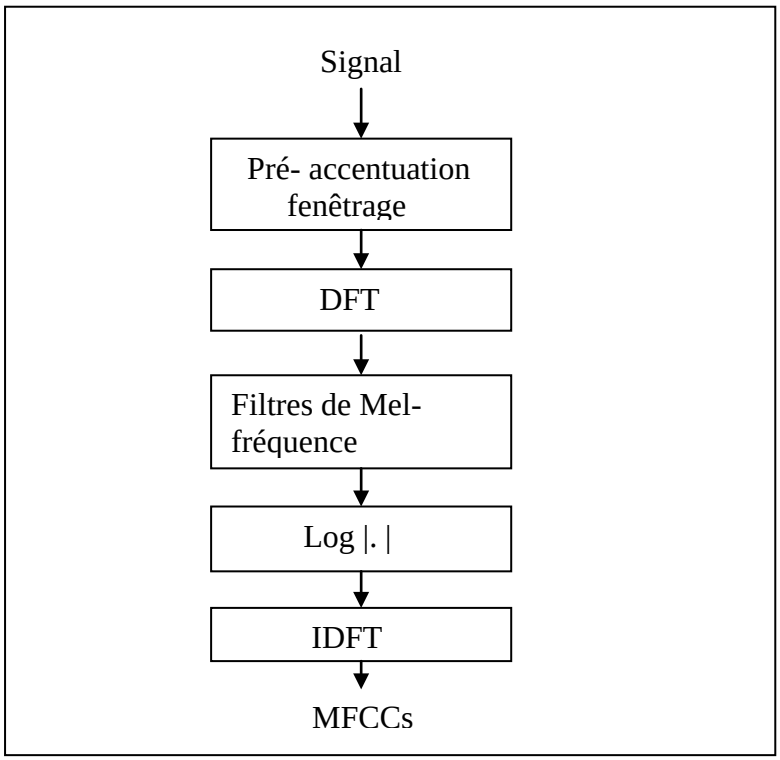


**Fig. I. 9: calcul des MFCCs**

Après le filtre de pré- accentuation et la segmentation du signal en trames, une transformée de Fourier discrète (DFT) est calculée pour faire passer le signal de parole dans le domaine spectral :

Pour un signal discret {*x[n]*} avec 0 <*n* <*N, où N* est le nombre d'échantillons d'une fenêtre d'analyse, *Fs* est la fréquence d'échantillonnage, la transformée de Fourier discrète (DFT)

*S[k]* est obtenue par:

$$s[k] = \sum_{n=0}^{N-1} x[n] e^{-j 2\Pi nk / N} \qquad \text{(I. 14)}$$

Le spectre du signal est multiplié avec des filtres triangulaires (Fig. I. 10) dont les bandes passantes sont équivalentes en domaine Mel-fréquence. Les points frontières *B[m]* des filtres en mel-fréquence sont calculés ainsi :

$$B[m] = B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1} \qquad 0 \le m \le M+1 \qquad \text{(I. 15)}$$

Où *M* est le nombre de filtres, $f_h$ est la fréquence la plus haute et *fl* est la fréquence la plus basse pour le traitement du signal.

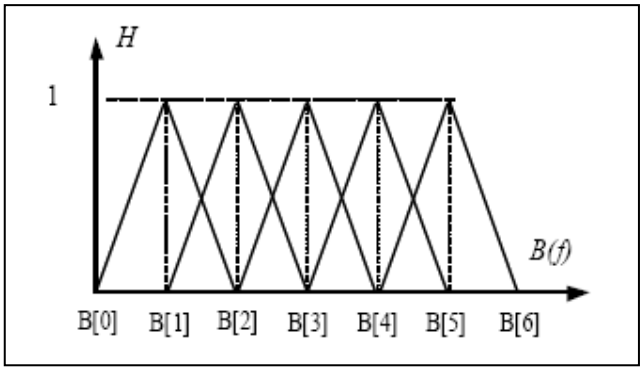


**Fig. I. 10 Les filtres triangulaires passe-bande en Mel-Fréquence (B(f)).**

Dans le domaine fréquentiel, les points *f[m]* discrets correspondants sont calculés par l'équation :

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \qquad \text{(I. 16)}$$

Où $B^{-1}$ est la transformée de mel-fréquence en fréquence. $B^{-1}(m) = 700 * (10^{m/2595} - 1)$.

Le coefficient *Hm[k]* de chaque filtre est déterminé par le système suivant :

$$H_m[k] = \begin{cases} 0 & si \quad k \le f[m-1] \\ \dfrac{k - f[m-1]}{f[m] - f[m-1]} & si \quad f[m-1] \le f[m] \\ \dfrac{f[m+1] - k}{f[m+1] - f[m]} & si \quad f[m] \le k \le f[m+1] \\ 0 & si \quad k \ge f[m+1] \end{cases} \qquad \text{(I. 17)}$$

Pour un spectre lissé et stable, à la sortie des filtres un logarithme de spectre d'amplitude est calculé : $E[m] = \log\left[\sum_{k=0}^{N-1} |S[k]|^2 H_m[k]\right] \quad 0 \le m \le M$ (I. 18)

Les coefficients cepstraux de mel-fréquence (MFCCs) seront obtenus par une transformée de cosinus discrète (permet d'obtenir des coefficients peu corrélés) à partir des coefficients aux sorties des filtres :

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos\left( \frac{\pi n (m + \frac{1}{2})}{M} \right) \qquad 0 \le n \le M \qquad \text{(I. 19)}$$

Une douzaine de coefficient MFCCs sont généralement considérés comme suffisants pour les expériences de reconnaissance de la parole .

Afin de prendre en compte la dynamique du signal, nous ajoutons aux paramètres MFCC les coefficients différentiels (ou coefficients delta) du premier et du second ordre (Fig. I. 11 ).

Soit le vecteur acoustique à N composantes MFCCs $C_t = \{c_t^1, c_t^2, ...., c_t^N\}$. Les coefficients delta de premier ordre sont alors estimés par :

$$\Delta C_t = \frac{\sum_{K=-L}^{L} K C_t}{\sum_{K=-L}^{L} K^2} \qquad \text{(I. 20)}$$

Les coefficients du second ordre sont calculés en itérant deux fois l'expression (I. 19)

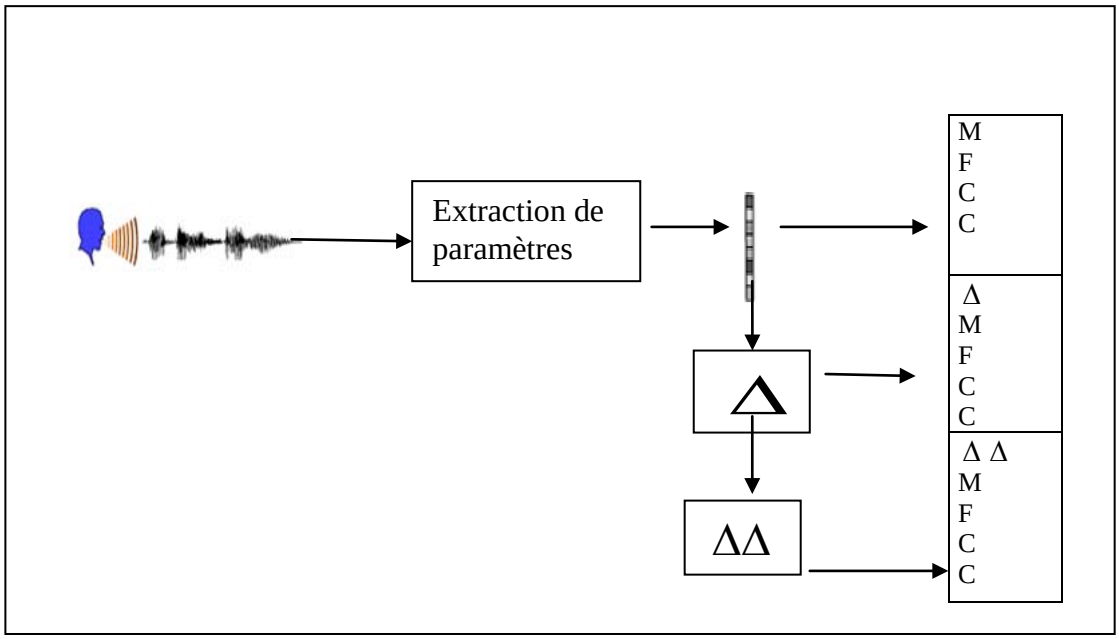


**Fig. I. 11 Calcul des dérivées premières et secondes de coefficients MFCCs.**

### I. 8.1.2 Les paramètres LSFs :

Les paramètres LSF sont une variante des coefficients LPC reconnue comme ayant de bonnes propriétés d'interpolation. Pour les calculer, il faut d'abord calculer les coefficients LPC. Pour cela, Le logarithme de la densité spectrale de puissance est représenté par le log des amplitudes *Al* [**24**].

Puis, les coefficients du filtre LPC sont estimés par application de l'algorithme de Levinson-Durbin sur les coefficients d'autocorrélation obtenus par une FFT inverse du carré des amplitudes.

Les coefficients $a_k$ du filtre LPC

$A_p(Z) = 1 + \sum_{k=1}^{p} a_k Z^{-k}$ sont donc convertis en coefficients LSFs. Dans cette représentation

$$A_p(Z) = \frac{1}{2}(P_{p+1}(Z) + Q_{p+1}(Z)) \qquad \text{(I. 21)}$$

Avec

$$P_{p+1}(z) = A(z) + z^{-(p+1)} A(z^{-1}) \qquad \text{(I. 22)}$$

$$Q_{p+1}(z) = A(z) - z^{-(p+1)} A(z^{-1}) \qquad \text{(I. 23)}$$

Les coefficients LSF sont extraits à partir des racines complexes des polynômes $P_{p+1}(z)$ et $Q_{p+1}(z)$.

Ces fonctions de transfert possèdent deux propriétés très intéressantes:

Pour un filtre stable $1/A_p(z)$, toutes les racines de $P_{p+1}(z)$ et $Q_{p+1}(z)$ sont sur le cercle unité et s'alternent deux à deux. D'autre part, ces racines sont conjuguées.

En ignorant les racines réelles (1 et -1 selon que $p$ est pair ou impair), le filtre $Ap(z)$ peut être représenté par la séquence $w1$, $w2$; …, $wp$ des arguments des racines complexe des filtres $P_{p+1}(z)$ et $Q_{p+1}(z)$ se trouvant sur le demi cercle entre 0 et $\pi$.

Les paramètres *wi* présentent plusieurs propriétés intéressantes. Tout d'abord, ils sont ordonnés : $0 < w1 < w2 < \ldots < wp < \pi$. Cette relation d'ordre est une condition nécessaire et suffisante pour la stabilité du filtre de synthèse $1/Ap(z)$.

### I. 8.1.3 Paramètres PLP (Perceptual Linear Prediction) :

L ’étude expérimentale a conduit à la notion de bande critique: des signaux dont la fréquence se situe à l ’intérieur d ’une bande critique influent sur la perception de signaux situés dans la même bande, mais n’influent pas à l extérieur de cette bande.

Une bande critique peut être considérée comme un filtre passe-bande dont la réponse en fréquence correspond approximativement à une courbe d ’accord d ’une fibre nerveuse auditive.

La méthode LP identifie uniformément le spectre sur toutes les fréquences de la bande audible. Or cette propriété est loin d’être vérifiée pour l’oreille humaine, car il a été établi que celle-ci est plus sensible aux fréquences situées au milieu de la bande d’analyse du spectre.
Ainsi, il est possible que certains détails spectraux importants du spectre ne soient pas pris en compte par l’analyse LP ou encore qu’ils prennent une importance majeure sans qu’ils soient physiologiquement pris en compte par l’oreille.
L’analyse PLP permet de résoudre ce problème. Elle permet d’estimer les paramètres du filtre auto-régressif tout pôle, modélisant au mieux le spectre auditif.
Le processus de calcul des coefficients PLP peut être décrit par la figure suivante:

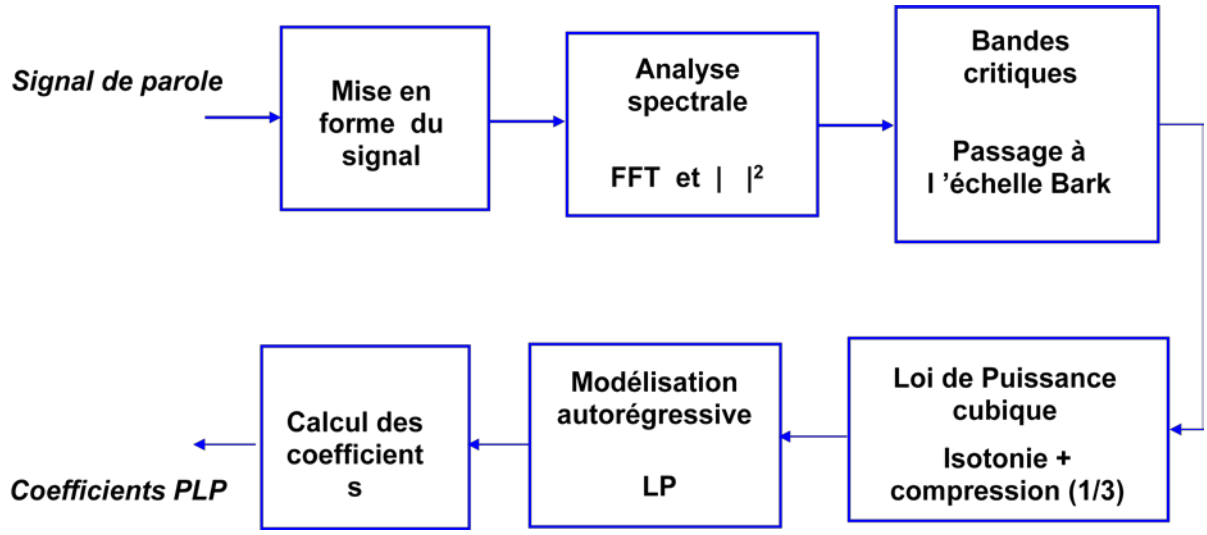


**Fig. I. 12 Le processus de calcul des coefficients PLP**

Après une mise en forme du signal de parole, le spectre de puissance P(ω) est calculé. Ensuite, un passage de l’échelle de fréquence usuelle à l’échelle de Bark est effectué en utilisant la relation (I. 24).

$$\Omega(\omega) = 6\ln\left( \frac{\omega}{1200\pi} + \left( \left( \frac{\omega}{1200\pi} \right)^2 + 1 \right)^{0.5} \right) \qquad \text{(I. 24)}$$

$\boldsymbol{\omega}$ représentant la fréquence angulaire exprimée en rd/s et Ω la fréquence de Bark.

Ce passage à l'échelle Bark, permet d'approximer de manière grossière ce que nous savons de la forme des filtres auditifs. Elle est approximativement constante le long de l'échelle de Bark. Le spectre de puissance dans l'échelle de Bark est convolué avec le spectre de puissance de la courbe de bande critique en utilisant l'équation suivante:

$$\psi(\Omega) = \begin{cases} 0 & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & -1.3 \le \Omega \le -0.5 \\ 1 \quad pour & -0.5 \le \Omega \le 2.5 \\ 10^{-1.0(\Omega-0.5)} & 0.5 \le \Omega \le 2.5 \\ 0 & \Omega > 2.5 \end{cases} \qquad \text{(I .25)}$$

Cette courbe de masquage est une approximation de la courbe de masquage asymétrique de Schroeder.

On essaye ensuite d'approximer la sensibilité de l'oreille humaine à différentes fréquences par l'intermédiaire d'une fonction de transfert E(ω). Le spectre de puissance est multiplié par cette fonction de transfert.

$$E(\Omega) = E(\omega).\Theta(\Omega) \qquad \text{(I. 26)}$$

$$\Theta(\Omega_t) = \sum_{\Omega=-1.3}^{\Omega=2.3} P(\Omega - \Omega_t).\Psi(\Omega) \qquad \text{(I. 27)}$$

La non linéarité entre l'intensité d'un son et sa force de perception par l'oreille est ensuite approximée par une loi de puissance :

$$\Phi(\Omega) = E(\Omega)^{0.33} \quad \text{(I. 28)}$$

L'étape finale consiste en une modélisation autorégressive classique du spectre du modèle auditif tout pôle, en calculant les coefficients autorégressifs du filtre.

L'analyse PLP est très similaire à l'analyse MFCC. La différence est que l'analyse PLP utilise l'échelle Bark au lieu de l'échelle Mel et un modèle autorégressif tout pôle au lieu de la transformée en cosinus discrète (DCT) pour le calcul des coefficients.
Cette méthode PLP a été par la suite améliorée pour résister à certaines conditions de bruit. C'est ainsi que l'analyse RASTA-PLP a été développée, RASTA étant l'acronyme de RelAtive SpecTrAl.

La méthode PLP, dont l'algorithme repose sur des spectres à court terme de la parole, résiste difficilement aux contraintes qui peuvent lui être imposées par la réponse fréquentielle d'un canal de communication. Pour atténuer les effets de distorsion spectrales linéaires, Hermansky, propose de modifier l'algorithme PLP en remplaçant le spectre à court terme par un spectre estimé où chaque canal fréquentiel est modifié par passage à travers un filtre. Cette modification est à la base de la méthode RASTA PLP. La mise en œuvre de ce filtrage (RASTA) permet, lorsqu'il est effectué dans le domaine spectral logarithmique, de supprimer les composantes spectrales constantes, supprimant ainsi les effets de convolution du canal de communication.

## I. 9 Conclusion :

Dans ce chapitre nous avons décrit brièvement le principe de production de la parole et quelques organes intervenant dans ce processus. Ensuite, nous avons étudié les techniques d'analyse du signal de parole, afin d'extraire des paramètres (LPC,MFFC,PLP et LSF) pertinents pour les utiliser dans la tache de reconnaissance de locuteurs (RAL) qui fera l'objet de chapitre IV.

# II Méthodes de Reconnaissance Automatique de Locuteurs.

# Chapitre II :

# Méthodes de Reconnaissance Automatique de Locuteurs

## II.1 Introduction :

Dans ce chapitre , nous nous intéressons au problème de la reconnaissance du locuteur et plus particulièrement au problème de l'identification et la vérification. La Reconnaissance Automatique du Locuteur (RAL) a pour objectif d'extraire les informations pertinentes concernant le locuteur à partir de son signal vocal, afin de pouvoir le reconnaître ultérieurement. On peut distinguer deux modes de RAL :

RAL en mode dépendant du texte : la reconnaissance du locuteur est réalisée à l'aide d'un message connu a priori par le système que le locuteur doit prononcer (mot de passe, code PIN, phrase,...).

Ce message peut être choisi par le locuteur comme dans les systèmes qui utilisent des mots de passe personnalisés [**25**], ou imposé par le système lui-même comme dans les systèmes utilisant des codes PIN [**26**].Ce message peut également être imposé et présenté par le système sous forme visuelle ou auditive.

RAL en mode indépendant du texte : dans ce cas, il n'existe aucune contrainte sur le message que le locuteur doit prononcer ni sur la langue qu'il peut utiliser.

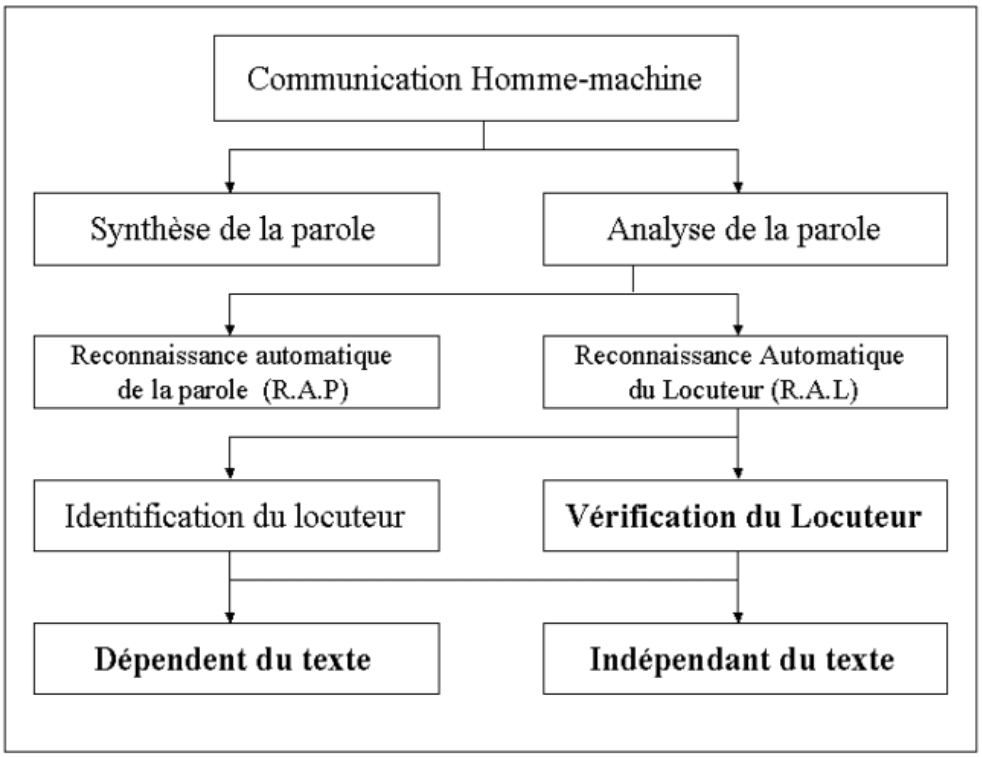


**Fig.II.1 Place de la RAL dans la Communication Homme-machine.**

### II.2 Les différentes tâches en RAL :

L'identification Automatique du Locuteur (IAL) et la Vérification Automatique du Locuteur (VAL) sont les deux tâches les plus répandues dans le domaine de la RAL. Récemment, pour des applications plus spécifiques, d'autre tâches ont vu le jour comme l'indexation du locuteur qui consiste à indiquer à quel moment chaque locuteur, intervenant dans une conversation, a pris la parole. Une application connexe est la détection d'un locuteur lors d'une conversation multiple. Dans cette section, nous allons décrire principalement les deux tâches principales de la RAL objet de notre étude : IAL et VAL.

#### II.2.1 Identification Automatique du locuteur (IAL) :

Dans cette tâche, le système doit fournir l'ensemble des locuteurs de la base d’apprentissage les plus proches du locuteur qui a produit le signal de parole de test (locuteur à identifier). Pour cela, le système calcule des mesures de similarités entre ce signal et tous les modèles des locuteurs de la base.

Le système n'ajoute un locuteur dans l'ensemble des locuteurs les plus proches que si le score de test du signal de parole présenté sur le modèle de ce locuteur est supérieur à un seuil défini à priori. En pratique, la plupart des systèmes d'IAL fournissent un ensemble d'un seul locuteur qui représente le locuteur le plus proche.

Dans le cas où le système doit fournir un ensemble d'au moins un locuteur, on parle alors d'une identification dans un ensemble fermé. Mais dans certaines applications, le système peut être amené à évoluer dans un ensemble ouvert. Il peut fournir dans ce cas un ensemble vide excluant de ce fait les éventuels imposteurs.

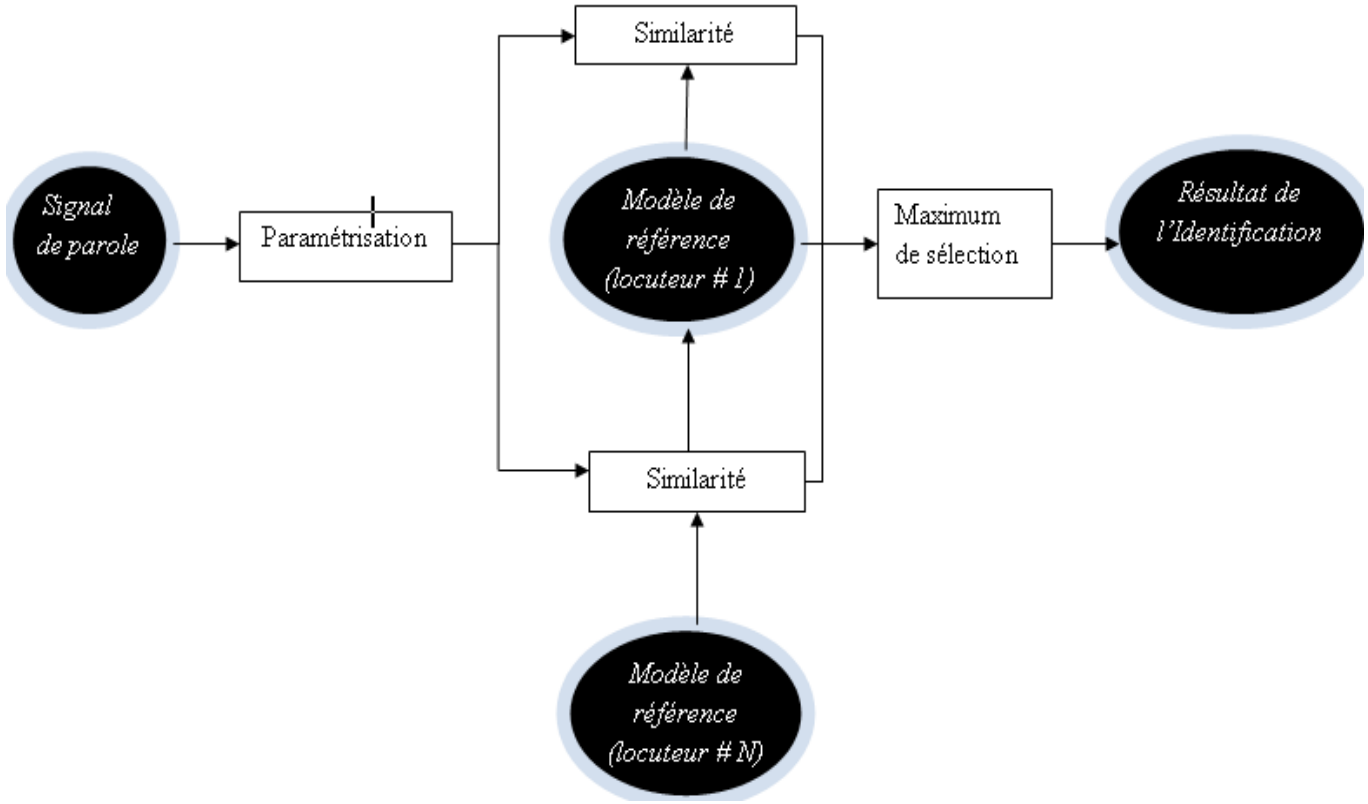


**Fig.II.2 Schéma modulaire d'un système d'IAL.**

### II.2.2 Vérification Automatique du locuteur (VAL) :

Un système de VAL doit vérifier à partir d'un signal de parole et d'une identité proclamée qui appartient à la base de données si le signal présenté provient de l'identité proclamée ou non. Pour cela, le système calcule une mesure de similarité entre le signal de test produit (identité prétendue) et une forme particulière de la base d'apprentissage (identité réelle).

En cas de concordance entre l'identité prétendue et l' identité réelle, nous pouvons dire que l'identité du locuteur a été vérifiée. Dans le cas contraire, le locuteur candidat du test est imposteur.

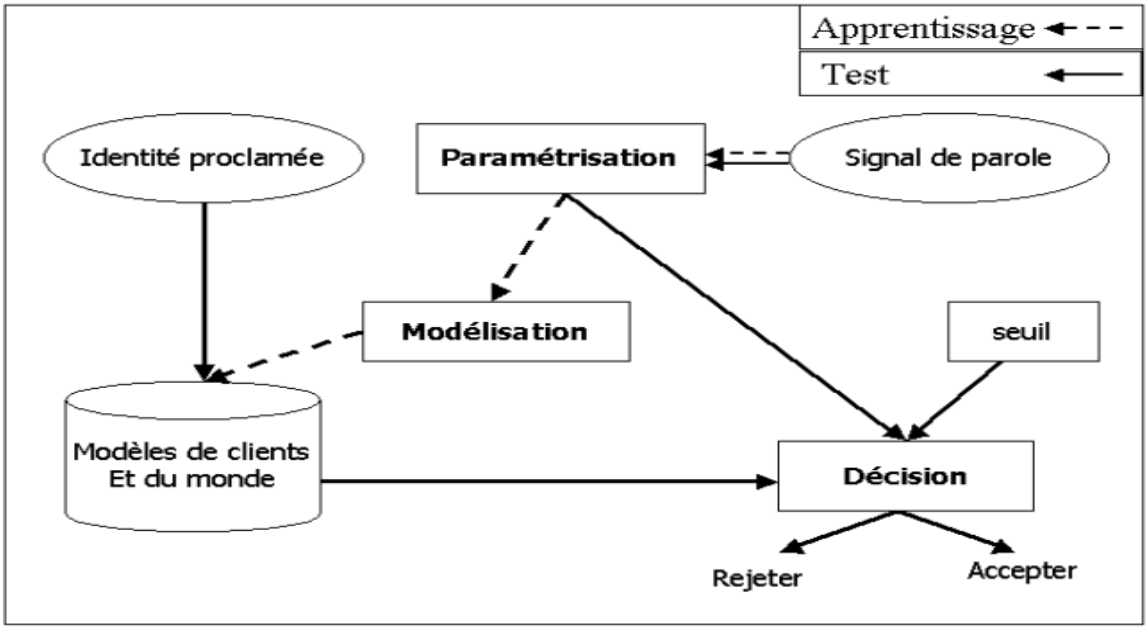


**Fig.II.3 Schéma modulaire d'un système de VAL.**

## II. 3 Modes de reconnaissance automatique de locuteurs

### II. 3.1 Reconnaissance du locuteur en mode dépendant du texte :.

Dans les systèmes de RAL opérant en mode dépendant du texte, ce dernier est imposé par le système. Les systèmes de reconnaissance du locuteur sur base de texte présenté ont été développés en premier. Dans ce cas, à chaque accès, l'utilisateur sera invité par le système (par exemple, sous forme de voix synthétique ou texte écrit) à prononcer un vocabulaire de base qui peut être très large ou simplement contenir les 10 chiffres qui seront utilisés pour créer des séquences aléatoires. L'avantage de cette approche est que l'utilisateur ne peut prédire la phrase qu'il sera invité à prononcer, ce qui rend tout enregistrement inutilisable.

### II. 3.2 Reconnaissance du locuteur en mode indépendant du texte :

Dans le cas de la reconnaissance du locuteur indépendante du texte, les mots ou les phrases prononcés pendant l'utilisation ne peuvent pas être prédits.

En général, les systèmes de reconnaissance du locuteur dépendants du texte sont plus robustes que les systèmes indépendants du textes. Malheureusement, dans les deux cas, ceux-ci sont aussi sujets à fraude étant donné que pour toutes les applications typiques de contrôle d'accès sur ligne téléphonique, la voix du locuteur (ainsi que son mot de passe dans le cas de systèmes dépendant du texte) pourrait être saisi, enregistré et reproduit frauduleusement

## II. 4 Techniques de reconnaissance automatique de locuteurs

### II. 4 .1 Méthodes basées sur les statistiques à long terme telles que la moyenne et la variance :

Celles-ci sont alors calculées sur une séquence (aussi longue que possible pour obtenir de bons estimateurs) de vecteurs acoustiques (généralement les vecteurs LPC, étant donné que ceux-ci sont plus caractéristiques du conduit vocal). Malheureusement, ces statistiques à long terme ne sont qu'une représentation très grossière des caractéristiques spectrales du locuteur et sont très sensibles aux variations de fonction de transfert.

Plus récemment, une approche a également été proposée dans laquelle les statistiques de variables dynamiques (par exemple, dans le domaine cepstral) sont utilisées et modélisée par un système autorégressif multidimensionnel. Dans, différentes mesures de distances pour cette approche sont comparées et on montre que des performances similaires à l'approche HMM [**7**][**8**]sont atteintes. On montre aussi que l'ordre optimal du processus auto régressif est d'ordre 2 ou 3. De plus, la normalisation des distances selon un critère de probabilité a posteriori semble ici essentielle à l'obtention de bons résultats.

**II. 4 .2 Méthodes basées sur la quantification vectorielle:**

dans ce cas, l'idée générale est de représenter les caractéristiques spectrales de chaque locuteur sur base de quelques vecteurs acoustiques les plus représentatifs et obtenus par quantification vectorielle.

Dans ce cas, le score d'une phrase d'entrée est défini comme la somme des distances de chacun des vecteurs acoustiques de la séquence par rapport au vecteur prototype le plus proche dans l'ensemble de vecteurs prototypes associés au locuteur considéré [**28**](proclamé ou faisant partie de la cohorte dans le cas de la normalisation).

Une variante de cette approche consiste à définir deux ensembles de prototypes par locuteur, respectivement pour les parties voisées et non-voisées (ainsi qu'un détecteur automatique de voisement). Pour les parties voisées, le pitch peut alors également être utilisé dans la détermination des prototypes et des distances. Evidemment, différentes méthodes de pondération des paramètres acoustiques intervenant dans le calcul des distances ont été largement testées.

Finalement, une alternative à l'utilisation de quantification vectorielle "sans mémoire" (étant donné que chaque vecteur acoustique est quantifié indépendamment du précédent) a été proposée dans où des algorithmes de codage de source à mémoire ont été testés.

**II. 4 .3 Méthodes basées sur HMM entièrement connectés (ergodiques):**

Dans ce cas, un modèle HMM entièrement connecté est entraîné pour chaque locuteur. Les états peuvent alors être définis de façon arbitraire et non supervisée ou être associés à des classes bien spécifiques (typiquement, des classes phonétiques ou, mieux, des classes phonétiques grossières de façon a réduire le nombre de paramètres). Finalement, quelques contraintes temporelles seront généralement introduites dans le modèle, en imposant une durée minimum pour chaque état. Plusieurs solutions relatives au nombre d'états de ces modèles, ainsi que des densités de probabilités associées à chaque état, ont été proposées:

– Modèles HMM basés sur le critère de vraisemblance[**10**] et ayant plusieurs états à une seule gaussienne, plusieurs états à multi-gaussiennes,ou un seul état à multi gaussiennes. Certaines méthodes discriminantes typiquement utilisées en reconnaissance du locuteur ont également été testées pour augmenter la discrimination entre locuteurs.

### II.5 Décision en identification et vérification:

En identification (vérification), un signal de test est comparé à toutes les références des locuteurs connus du système, résultant en un ensemble de mesure de similarité (ou un ensemble de mesure de distance) à l'entrée du processus de décision.

Aussi, la règle de décision consiste à choisir le locuteur dont la mesure de similarité est maximale (ou minimale dans le cas de mesure de distance).

Pour l'évaluation des performances du système d'identification(vérification) du locuteur, le taux de classification correcte est souvent utilisé. Ce taux est le rapport entre le nombre des segments correctement identifie(vérifiés) et le nombre total des segments de test.

$$\textbf{Taux d' identification (vérification) correct \%} = \frac{\textbf{tests correctement identifiés(vérifiés)}}{\textbf{test total}}$$

### II.6 Domaines d'applications :

Dans ce paragraphe on donne quelques exemples d'applications en RAL et que l'on peut regrouper en trois catégories principales : applications en contrôle d'accès sur sites sensibles, application dans le domaine sécuritaire et juridiques (notamment en sciences forensiques), applications dans les systèmes de communication.

#### II.6.1 Applications sur l'accès restreint sécurisé à des sites sensibles :

Cette catégorie concerne les applications qui se trouvent sur un site géographique particulier, elles sont utilisées principalement pour limiter l'accès à des lieux privés . Voici quelques exemples de ce type d'applications :

Verrouillage automatique: ces applications sont utilisées comme une sorte de verrous électroniques comme par exemple la protection de domicile, garage, bâtiment, etc.

Validation des transactions sur site (comme contrôle supplémentaire au niveau des distributeurs bancaires) .

Accès aux lieux de production des usines : qui sont en général réservés aux employés, ouvriers et inspecteurs afin de protéger le secret de la production et du matériel.

L'intérêt de ce type d'application est :

D'abord l'environnement est facilement contrôlable.

La vérification du locuteur a un rôle dissuasif.

La reconnaissance vocale peut être associée à d'autre techniques de reconnaissance d'identité (ex : analyse du visage, des empreintes digitales, iris etc.). L'utilisateur peut avoir son modèle sur lui (ex : sur la puce d'une carte).

**II.6.2 Applications dans les systèmes de communication :**

Ce type d'applications utilise par exemple le téléphone comme un moyen matériel de communication entre l'homme et la machine. C'est la catégorie la plus importante parce qu'elle permet de vérifier ou identifier le locuteur à longue distance.

Il existe plusieurs applications dans cette catégorie et parmi elles :

Validation de transactions bancaires par téléphone (pour améliorer le service bancaire, ainsi que pour valider légalement la transaction effectuée ) –

Accès à des bases de données pour plus de sécurité et pour plus de protection (ex : consultation d'émail, consultation de répondeur, etc.).

Accès à des services téléphoniques (ex : téléphoner sur son compte de facturation personnelle de n'importe quelle ligne téléphonique).

Les inconvénients de ce type d'applications sont principalement :

L'environnement est difficilement contrôlable parce que la qualité des lignes téléphoniques peut varier considérablement d'un appel à un autre, ainsi que le bruit de fond produit par le lieu d'appel (bar, restaurant, bureau, etc.). Les applications exigent le stockage des données de manière centralisée.

**II.6.3 Applications juridiques :**

Enfin on trouve le domaine d'applications qui pose actuellement le plus de problèmes, c'est le domaine juridique. La reconnaissance de locuteur est utilisée par exemple pour :

-L'orientation des enquêtes.

-La constitution des éléments de preuves au cours d'un procès.

Dans ces applications on trouve beaucoup plus d'inconvénients que d'avantages :

-La quantité de la parole à disposition est en général très limitée .

-Les conditions d'environnement sont très mauvaises .

-Les locuteurs impliqués sont très rarement coopératifs .

## II.7 Conclusion :

Dans ce chapitre nous avons présenté les méthodes de reconnaissance automatique de locuteurs ainsi que quelques domaines d'application de la RAL. On s'est focalisé sur les deux modes de l'identification, dépendant et indépendant du texte, ainsi que les techniques de reconnaissance qui seront développées dans le chapitre suivant par la suite le sujet de nos expériences au chapitre IV.

# III Les Méthodes à Noyaux

# Chapitre III :

# Les Méthodes à Noyaux

## III.1 Introduction :

Dans cette partie, nous introduisons les méthodes statistiques pour la classification automatique avec apprentissage supervisé. Qui consiste à réunir les méthodes dites "discriminantes", qui ont connu un succès croissant pour de nombreuses taches de classification. Ces approches incluent la méthode des GMM, les Réseaux de Neurones et les Machines à Vecteurs de Support(SVM).

## III.2 Qu'est ce-qu' un noyau reproduisant?

### III.2.1 Définition du Noyau reproduisant:

Soit X un espace quelconque, et (H, $<.,.>_H$) un espace de Hilbert de fonctions ($H \subset R^X$ ).

une fonction K : X × X⟶R est appelée un noyau reproduisant (noté n.r.) seulement si

- H contient toutes les fonctions de la forme : $\forall x \in \chi \ , K_X : t \to = K(x,t)$
- Pour tout x ∈ X et f ∈H, on a: f (x) =$<f, K_X>_H$,

Si un **n.r**. existe, H est appelé un espace de Hilbert à noyau reproduisant (**rkhs**).

### III.2.2 Propriétés des Noyaux reproduisants (n.r) et espace de Hilbert à noyau reproduisant (rkhs) :

- Si un n.r. existe, il est unique.
- Un n.r. existe si et seulement si $\forall x \in \chi$ , la fonctionnelle :f⟶ f (x) (de H dans R) est continue.
- Un n.r. est un noyau défini positif.
- Si K est un n.r., il vérifie la propriété reproduisante :

$$\forall (x,y) \in \chi^2, \langle K_x, K_y \rangle_H = K(x,y).$$

## III.3 Méthodes assimilées aux méthodes à noyaux :

### III.3.1 Les mélanges de gaussiennes (GMM) :

L'utilisation des GMM[**1**], pour la modélisation du locuteur est motivée pour deux raisons. La composante gaussienne fut, en premier lieu, utilisée pour représenter les caractéristiques spectrales des formes phonétiques issues de la voix d'une personne. L'espace acoustique correspondant à la voix d'un locuteur peut être caractérisée par un ensemble de classes acoustiques représentant des événements phonétiques, voyelles, nasals ou fricatives. Ces classes acoustiques reflètent une certaine configuration des cordes vocales dépendante du locuteur qui sont utiles pour caractériser son identité. La forme spectrale de la iième classe acoustique peut être représentée par le vecteur moyen µi de la iième composante et les variations de la forme spectrale moyenne peuvent être représentées par la matrice de covariance σi. En générale,

les classes acoustiques sont cachées vu que la classe d'une observation est inconnue. Supposant que les vecteurs acoustiques soient indépendants, la distribution de ces vecteurs acoustiques tirés de ces classes cachées est un mélange de

gaussiennes.

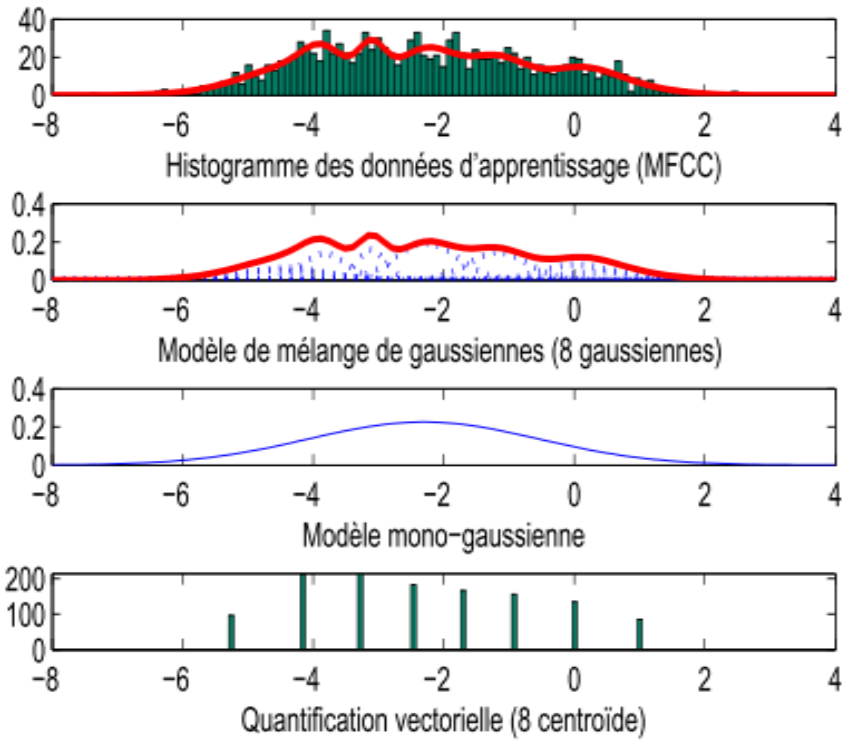


**Fig. .III.1 Histogramme d'un ensemble des paramètres, avec son modèle monogaussienne, son modèle de mélange de gaussiennes et son modèle par quantification vectorielle**

La deuxième motivation pour l'usage de mélange de gaussiennes pour l'identification du locuteur est une observation empirique. Une combinaison linéaire des distributions gaussiennes est capable de représenter une grande classe des paramètres dans une simple distribution. Le modèle mono gaussien classique du locuteur représente la distribution des vecteurs acoustiques d'un locuteur par une position (moyenne)et une forme elliptique (matrice de covariance) et le modèle de QV représente la distribution de ces vecteurs par un ensemble discret de poids de pondération caractéristiques. Dans un certain sens le GMM peut être considéré comme une hybridation entre ces deux modèles en utilisant un ensemble discret de fonctions gaussiennes, chacune avec leur propre vecteur de moyenne et matrice de covariance, ce qui permet une meilleure modélisation pour le locuteur. La densité de probabilité d'une mixture de gaussiennes à N composantes pour une variable aléatoire x s'exprime sous la forme suivante :

$$p\left(x/\Theta\right)=\sum_{i=1}^{N}\gamma_i N\left(x,\mu_i,\sum_i\right) \qquad \text{(III.1)}$$

sous la contrainte :

$\sum_i \gamma i = 1$ et $\forall i : \gamma_i \geq 0$.

$\gamma$ : est le vecteur de poids de la mixture.

$N(x;\mu,\Sigma)$ : est la loi gaussienne de moyenne $\mu$ et de variance $\Sigma$.

$\Theta = [\mu,\Sigma,\gamma]^T$ : est le vecteur de paramètre global du GMM.

Si x est de dimension d alors, une mixture de gaussienne est paramétrée par N.d paramètres de moyennes, $N.d^2$ paramètres de variance, et N paramètres de poids. La densité d'une distribution normale de dimensions d est :

$$N\left(x,\mu,\sum\right)=\frac{1}{\left(2/p_i\right)^{d/2}\left|\sum\right|^{1/2}}\exp\left[\left(x-\mu\right)^T\sum^{-1}\left(x-\mu\right)\right] \qquad \text{(III.2)}$$

Pour calculer la vraisemblance d'une séquence *X = [x1...xT]*, pour un modèle paramétré par $\Theta$, le logarithme est généralement utilisé en considérant l'indépendance des réalisations de la séquence d'apprentissage. Posons la notation $\log(p(.)) = \ell(.)$, alors

$$\log_p\left(x/\Theta\right)=\ell\left(x/\Theta\right)=\sum_{t=1}^{T}\log\sum_{i=1}^{N}\gamma_i N\left(x,\mu_i,\sum_i\right) \qquad \text{(III.3)}$$

L'apprentissage d'un GMM est généralement réalisé avec l'algorithme EM.

**III.3.1.1 L'Algorithme EM :**

L'algorithme EM se compose de deux paliers. Le premier est une initialisation du modèle par Quantification Vectorielle (par exemple). Le second palier est une optimisation des paramètres du mélange par l'algorithme classique Expectation Maximization.

La partie optimisation est un algorithme itératif qui comporte deux étapes :estimation et maximisation.

**Initialisation** Utilisation de l'algorithme Lloyd (Quantification Vectorielle) pour l'initialisation des moyennes des M gaussiennes du modèle.

Initialisation de toutes les matrices de covariance $\sum_{i=1}^{N}$ à la matrice unité I.

Initialisation équiprobable des poids des composantes : $\omega_i = 1/M$.

**Itération** pour i = 1,...,N−1

**Phase d'Estimation :**

Pour tous les vecteurs acoustiques n = 1,.....,T Calcul de la probabilité $P_{ni}$ , probabilité que le vecteur $x_n$ soit généré par la loi gaussienne i.

$$P_{ni} = \frac{\frac{w_i}{(2/p_i)^{d/2}\left|\sum\right|^{1/2}} \exp\left[-\frac{1}{2}(x_n-\mu_i)^T \sum^{-1} (x_n-\mu_i)\right]}{\sum_{i'}^{N-1} \frac{w_{i'}}{(2/p_i)^{d/2}\left|\sum\right|^{1/2}} \exp\left[-\frac{1}{2}(x_n-\mu_{i'})^T \sum^{-1} (x_n-\mu_{i'})\right]} \quad \text{(III.4)}$$

Cette étape est équivalente à avoir un ensemble Q de variables continues cachées, prenant des valeurs dans l'intervalle [0; 1], qui donnent un étiquetage des données (vecteurs acoustiques) en indiquant dans quelle proportion un vecteur $x_n$ appartient à la gaussienne i.

**Phase de Maximisation** Réestimation des paramètres à partir des probabilités $P_{ni}$.

$$w_i^* = \frac{1}{T}\sum_{n=1}^{T} P_{ni} \quad \text{(III.5)}$$

$$\mu_i^* = \frac{\sum_{n=1}^{T} P_{ni} x_n}{\sum_{n=1}^{T} P_{ni}} \qquad \text{(III.6)}$$

Dans le cas présent, tous les vecteurs de données participent `a la mise à jour du modèle, mais leur participation est proportionnelle `a la valeur $P_{ni}$ , Incrémentation de i à i+1 et retour à la phase d'estimation

### III.3.2 Réseaux de Neurones :

#### III.3.2.1 Topologies de réseaux de neurones :

Il existe plusieurs topologies de réseaux de neurones :

- Les réseaux multicouche : Ils sont organisés en couches, chaque neurone prend généralement en entrée tous les neurones de la couche inférieure.

Ils ne possèdent pas de cycles ni de connexions intra-classe. On définit alors une « couche d'entrée », une « couche de sortie », et n « couches cachées ». Ce type de réseaux est très répandu, du fait de son apprentissage aisé.

- Les réseaux à connexions locales : On reprend la même structure en couche que précédemment, mais avec un nombre de connexions limité :

Un neurone n'est pas forcément connecté à tous les neurones de la couche précédente.

- Les réseaux à connexion récurrentes : On a toujours une structure en couche, mais avec des retours ou des connexions possibles entre les neurones d'une même couche (réseau récurrent d'Elman, de Jordan, etc.).

- En fin dans les réseaux à connexions complètes, tous les neurones sont interconnectés.

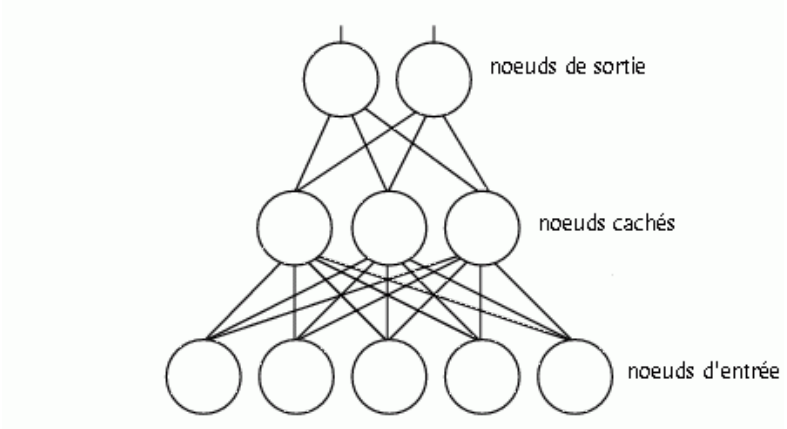


**Fig.III.2 : Exemple de réseau multicouche à une couche cachée**

### III.3.2.2 Le neurone formel :

Le neurone formel est une unité élémentaire. Il effectue la somme pondérée de ses entrées, et la soumet à une fonction non linéaire dérivable :

Pour un neurone formel possédant n entrées, le neurone effectue la somme pondérée :

$$y = \sum_{i=1}^{n} w_i x_i \qquad \text{(III.7)}$$

Puis « active » sa sortie grâce à une fonction non linéaire :

$$Z = f(y) = f\left( \sum_{i=1}^{n} w_i x_i \right) \qquad \text{(III.8)}$$

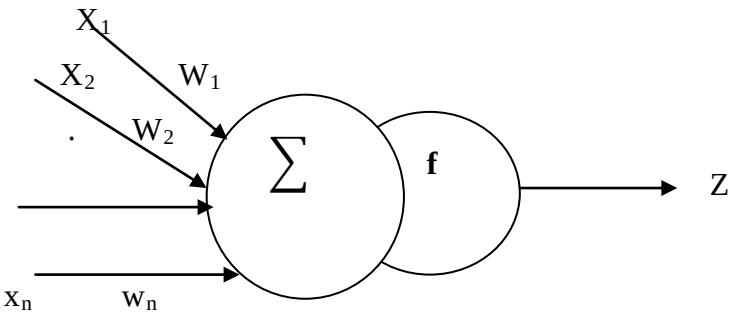


**Fig. III.3 Neurone Formel**

Plusieurs fonctions sont utilisées pour l'activation :

La fonction sigmoïde :

$$g(a) = 1/ (1 + \exp(-a)). \qquad \text{(III.9)}$$

La fonction de Heaviside : g(a) = 0 si a < 0 ; 1 sinon.

Une fonction radiale de type gaussienne :

$$g(a) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(a-\mu)^2}{2\sigma^2}\right). \qquad \text{(III.10)}$$

$\sigma$ : écart type appelé aussi paramètre de lissage.

$\mu$ : moyenne.

### III.3.2.3 Le perceptron multicouches :

Les Multilayer Perceptron (MLP) appartiennent aux réseaux multicouches [**29**] : ils ne possèdent donc pas de boucle de retour, ils sont « Feed-forward ».

Les MLP possèdent une fonction d'activation de type sigmoïde ou de Heaviside.

Le MLP est une extension multicouche du perceptron, qui est un réseau à une couche, assez limitée. Il utilise un algorithme d'apprentissage très répandu car facile à implémenter : la rétro-propagation du gradient, qui utilise une erreur quadratique moyenne.

#### III.3.2.3.1 La rétro-propagation du gradient :

##### III.3.2.3.1.1 Principe :

La rétro-propagation du gradient consiste à propager « à l'envers » (de la couche de sortie vers la couche d entrée) l erreur obtenue sur les exemples de la base d'apprentissage. On utilise pour cela l'erreur quadratique, i.e. le carré de la différence entre ce qu'on obtient et ce qu'on désire.

Si on calcule la dérivée partielle de l'erreur quadratique par rapport aux poids des connexions (d'ou le « gradient »), il est possible de déterminer la contribution des poids à l erreur générale, et de corriger ces poids de manière à se rapprocher du résultat souhaité.

La correction se fait par itération en corrigeant plus ou moins fortement les poids par l'intermédiaire d'un coefficient à l'issue d'un certain nombre d'itérations, lorsque qu'on est satisfait du classement des exemples de notre base d'apprentissage, on fixe les poids qui constituent ainsi des frontières entre les classes.

##### III.3.2.3.1.2 Algorithme :

**Définition du réseau :**

Considérons un réseau à une couche cachée. Le réseau possède :

Une couche d'entrée à m cellules d'entrées $x_i = e_i$ (Il ne s'agit pas de neurones, ces cellules présentent simplement les entrées ei du réseau).

- Une couche cachée à n neurones d'activation $y_j$.
- Une couche de sortie à p neurones d'activation $z_k$.
- n x m connexions entre la couche d'entrée et la couche cachée, chacune pondérée par $V_{ji}$.
- m x p connexions entre la couche cachée et la couche de sortie, chacune pondérée par $w_{kj}$.

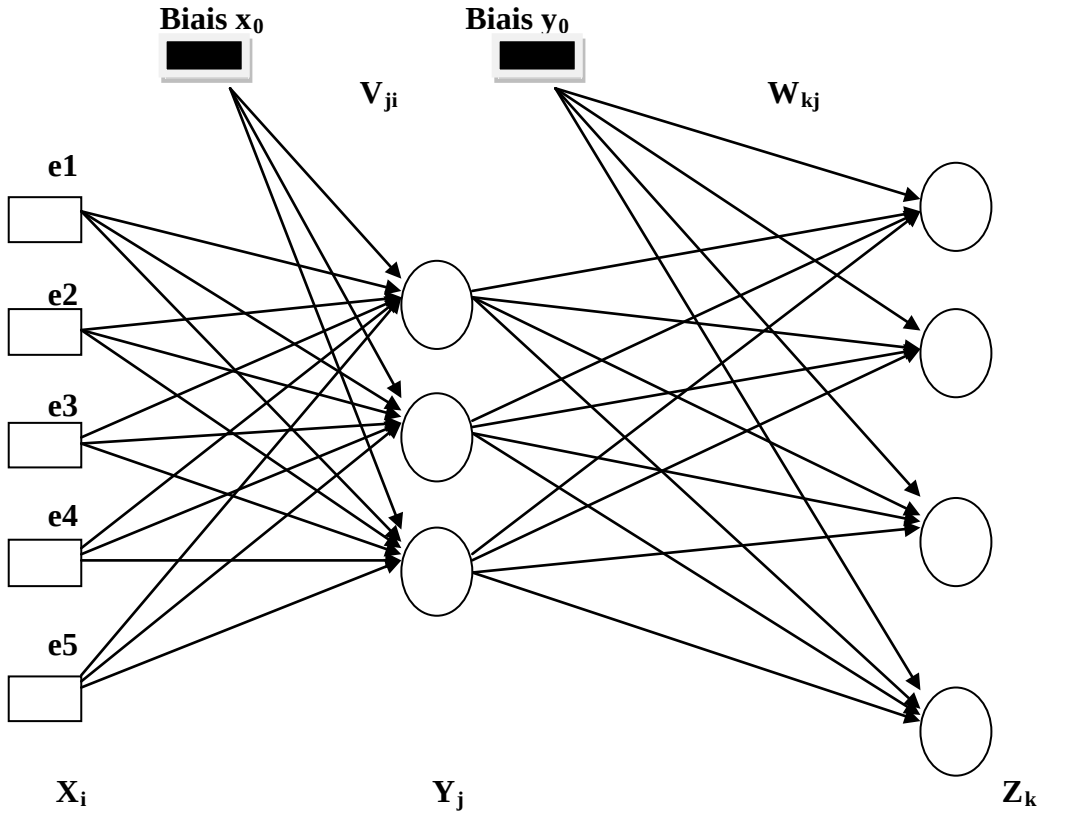


**Fig. III.4 Exemple de réseau MLP à une couche cachée avec 5 entrées, 3 neurones dans la couche cachée, et quatre sorties.**

**ETAPE 1 : Initialisation des poids des connexions**

Ces poids sont choisis au hasard.

**ETAPE 2 : Propagation des entrées**

Les ei sont présentées à la couche d'entrée : $\mathbf{x_i}$ = ei.

On propage vers la couche cachée :

$$y_j = f\left(\sum_{i=1}^{m} x_i v_{ji} + x_0\right) \qquad \text{(III.11)}$$

puis de la couche cachée vers la couche de sortie :

$$Z_k = f\left(\sum_{j=1}^{n} y_j w_{kj} + y_0\right) \qquad \text{(III.12)}$$

Les valeurs $x_0$ et $y_0$ sont des biais : des scalaires et non des sorties de la couches précédente.

**ETAPE 3 : rétro-propagation de l'erreur**

Pour chaque exemple de la base d'apprentissage appliqué en entrée du réseau, on calcule son erreur sur les couches de sorties, c'est à dire la différence entre la sortie désirée $s_k$ et la sortie réelle $z_k$ :

$$E_k = Z_k(1 - Z_k)(S_k - Z_k) \quad \text{(III.13)}$$

On propage cette erreur sur la couche cachée ; l'erreur de chaque neurone de la couche cachée est donnée par :

$$F_j = y_j(1 - y_j)\sum_{k=1}^{p} w_{kj} E_k \quad \text{(III.14)}$$

**ETAPE 4 : Correction des poids des connexions**

Il reste à modifier les poids des connexions :

Entre la couche d'entrée et la couche cachée :

$$\begin{cases} \Delta v_{ij} = \eta x_i F_j \\ \Delta x_0 = \eta F_j \end{cases} \quad \text{(III.15)}$$

Entre la couche cachée et la couche de sortie :

$$\begin{cases} \Delta w_{kj} = \eta y_j E_k \\ \Delta y_0 = \eta E_k \end{cases} \quad \text{(III.16)}$$

$\eta$ étant un paramètre qu'il reste à déterminer.

**BOUCLER à l'étape 2** jusqu'à un critère d'arrêt à définir (seuil).

## III.4 Méthode à noyaux :

### III.4.1 Réseau de neurone à régression généralisée (GRNN: General regression neural networks ) :

Une fonction continue quelconque peut être approchée par une combinaison linéaire des fonctions gaussiennes bien choisies.

Objectif : Régression : construire une bonne approximation d'une fonction qui n'est pas connue par seulement un nombre fini d'échantillons tirés d'expérience [**29**].

Régression locale : les gaussiennes de base n'influent que sur des petites zones autour de leurs valeurs moyennes.

Ce réseau peut être employé pour des problèmes de classification. Quand une entrée est présentée, la première couche calcule les distances entre le vecteur d'entrée est le vecteur de poids et produit un vecteur, multipliée par le biais.

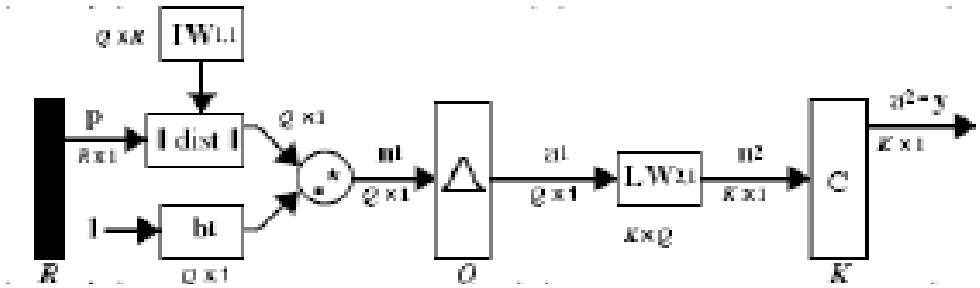


**Schéma général d'un réseau de neurones type GRNN**

### III.4.2 Machines à Vecteurs de Supports :

Les Support Vector Machines (SVM) sont des nouvelles techniques discriminantes dans la théorie de l'apprentissage statistique. Elles ont été proposées en 1995 par V. Vapnik dans son livre « The nature of statistical learning theory »[**19**].

Elles permettent d'aborder plusieurs problèmes divers et variés comme la régression, la classification, la fusion etc.

Les SVM fournissent une approche très intéressante de l'approximation statistique. Souvent, le nombre des exemples pour l'apprentissage est insuffisant pour que les estimateurs fournissent un modèle avec une bonne précision.

D'un autre côté, l'acquisition d'un grand nombre d'exemples s'avère être souvent très coûteuse. Pour ces raisons, il faut arriver à un compromis entre la taille des échantillons et la précision recherchée.

Dans ces cas spécifiques comme la reconnaissance de formes, il serait intéressant de trouver une mesure de la fiabilité de l'apprentissage, et d'avoir une mesure du taux d'erreur qui sera commis durant la phase de test.

Le principe des SVM consiste à projeter les données de l'espace d'entrée (appartenant à deux classes différentes) non-linéairement séparables dans un espace de plus grande dimension appelé espace de caractéristiques de façon à ce que les données deviennent linéairement séparables. Dans cet espace, on construit un hyperplan optimal séparant les classes tel que :

Les vecteurs appartenant aux différentes classes se trouvent de différents côtés de l'hyperplan.

La plus petite distance entre les vecteurs et l'hyperplan (la marge) soit maximale.

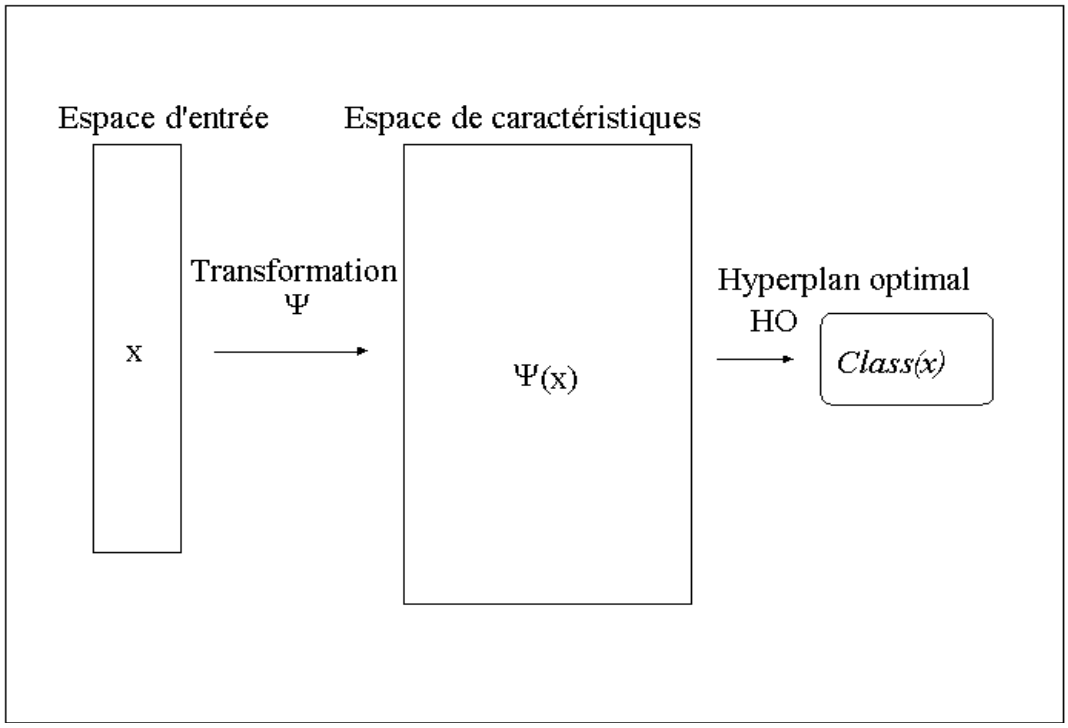


**Fig. III.5 Principe des techniques SVM**

### III.4.2.1 Construction de l'hyperplan optimal :

Pour bien décrire la technique de construction de l'hyperplan optimal séparant des données appartenant à deux classes différentes dans deux cas différents : le cas des données linéairement séparables et le cas des données non linéairement séparables, nous considérons le formalisme suivant :

Soit l'ensemble *D* tel que :

$$D = \{(xi, yi) \in \mathfrak{R}^n x\{-1, 1\} \text{ pour i = 1, .., m}\} \quad \text{(III.17)}$$

#### III.4.2.1.1 Cas des données linéairement séparables :

Dans ce paragraphe nous présentons la méthode générale de construction de l'Hyperplan Optimal (HO) qui sépare des données appartenant à deux classes différentes linéairement séparables. La figure (Fig.III.6) donne une représentation visuelle de l'HO dans le cas des données linéairement séparables.

Soit $H$ : $(w \cdot x + b)$ l'hyperplan qui satisfait les conditions suivantes :

$$\begin{cases} w.x + b \geq 1 \quad si \quad y_i = 1 \\ w.x + b \leq 1 \quad si \quad y_i = -1 \end{cases} \quad \text{(IV.18)}$$

ce qui est équivalent à : $y_i(w \cdot x_i + b) \geq 1$ pour i = 1…..m (IV.19)

Comme nous l'avons déjà mentionné, un HO est un hyperplan qui maximise la marge M qui représente la plus petite distance entre les différentes données des deux classes et l'hyperplan. Maximiser la marge M est équivalent à maximiser la somme des distances des deux classes par rapport à l'hyperplan. Ainsi, la marge a l'expression mathématique suivante :

$$M = \min_{\frac{x_i}{y_i}=1} \frac{w.x + b}{\|w\|} - \max_{\frac{x_i}{y_i}=-1} \frac{w.x + b}{\|w\|} \quad \text{(III.20)}$$

$$=1/||w|| - (-1)/||w||$$

$$=2/||w||$$

Trouver l'hyperplan optimal revient donc à maximiser 2/||w|| Ce qui est équivalent à minimiser $||\mathbf{w}||^2/2$ sous la contrainte (III.19) Ceci est un problème de minimisation d'une fonction objective quadratique avec contraintes linéaires.

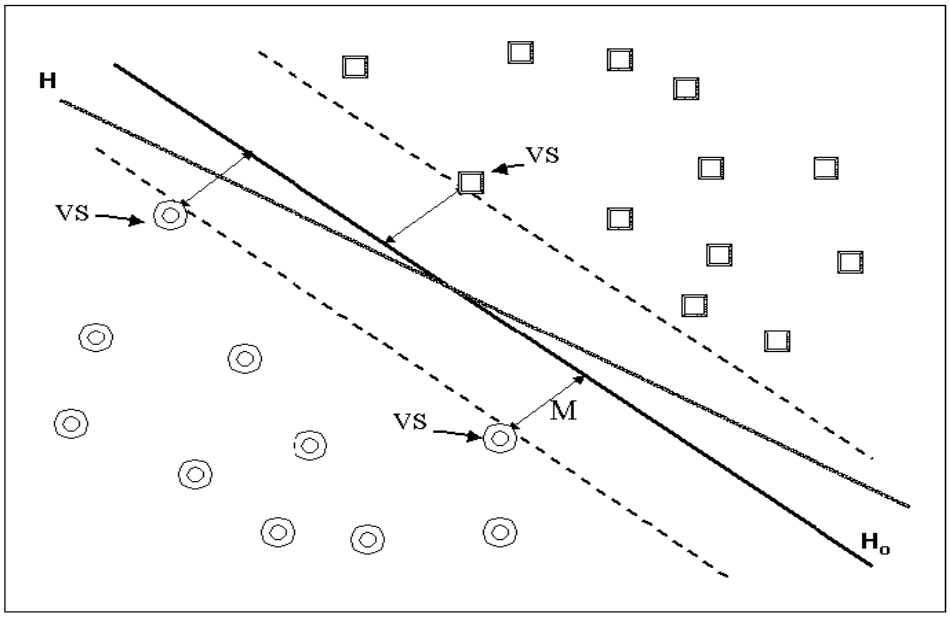


**Fig. III.6 - Hyperplans séparateurs : H est un hyperplan quelconque, Ho est l'hyperplan optimal et M est la marge qui représente la distance entre les différentes classes et Ho (VS sont les Vecteurs Supports).**

**III.4.2.1.2 Principe de Fermat:**

Les points qui minimisent où maximisent une fonction dérivable annule sa dérivée. Ils sont appelés points stationnaires.

**III.4.2.1.3 Principe de Lagrange:**

Pour résoudre un problème d'optimisation sous contrainte, il suffit de rechercher un point stationnaire $z_0$ du lagrangien L (z;α) de la fonction $g$ à optimiser et les fonctions $C_i^g$ exprimant les contraintes

$$L(z;\alpha) = g(z) + \sum_{i=1}^{m} \alpha_i C_i^g(z) \qquad \text{(III.21)}$$

ou les $\alpha_i = (\alpha_1, \alpha_2 \dots \dots \dots \alpha_m)$ sont des constantes appelés coefficients de Lagrange.

#### III.4.2.1.3 Principe de Kuhn-Tucker :

Avec des fonctions *g* et $C_i^g$ convexe, il est toujours possible de trouver un point-selle ($z_0$,$\alpha^*$) qui vérifie

$$\min_{\mathbf{z}} L(z,\alpha^*) = L(z_0,\alpha^*) = \max_{\alpha \geq 0} L(z_0, \alpha) \quad \text{(III.22)}$$

En appliquant le principe de Kuhn-Tucker, on est amené à rechercher un point-selle $(w_0, b_0, \alpha_0)$. Le lagrangien correspondant à notre problème est :

$$L(w; b; \alpha) = 1/2\ w^T . w - \sum_{i=1}^{m} \alpha i\{ y_i [x_i . w + b] - 1\} \quad \text{(III.23)}$$

Le lagrangien doit être minimal par rapport à *w* et *b* et maximal par rapport à $\alpha \geq 0$

- $L(w,b,\alpha)$ est minimal par rapport à *b* :

$$\frac{\partial L(w,b,\alpha)}{\partial b} = 0 \Leftrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \quad \text{(III.24)}$$

-L(w,b, α) est minimal par rapport à w :

$$\frac{\partial L(w,b,\alpha)}{\partial w} = 0 \Leftrightarrow w - \sum_{i=1}^{m} \alpha_i x_i y_i = 0 \quad \text{(III.25)}$$

-L(w,b, α) est maximal par rapport à α >= 0 :

En remplaçant (III.24)et (III.25)dans le lagrangien (III.23) on aura

$$L(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - 1/2 \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i . x_j) \quad \text{(III.26)}$$

Ainsi notre problème est de maximiser L(w,b, α) sous la contrainte :

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \quad ; \qquad \alpha_i \geq 0 \quad \text{(III.27)}$$

Soit la solution $\alpha^0 = (\alpha 1^0, \alpha 2^0, \ldots\ldots\ldots\ldots\ldots\ldots\ldots \alpha m^0)$. D'après le théorème de Kuhn-Tucker une condition nécessaire et suffisante pour que $\alpha^0$soit optimal est :

$$\alpha_i{}^0 \, [yi[(w_0 \, . \, x_0) + b_0] - 1] = 0 \text{ pour } i = 1, \ldots , m$$

ce qui veut dire que : $\alpha_i{}^0 = 0$ où $yi \, [(w_0 \, . \, xi) + b_0 \, ] = 1$ (III.28)

**III.4.2.1.4 Définition :**

On définit les Vecteurs Supports VS tout vecteur xi tel que $yi[(w_0 \, . \, xi) + b_0 \, ] = 1$. Ce qui est équivalent à :

$$V S = \{xi \mid \alpha_i > 0\} \text{ pour } i = 1,\ldots,m$$

Ainsi, on peut facilement calculer $w_0$ et $b_0$ :

$$w_0 = \sum_{VS} \alpha_i{}^0 y_i x_i \quad \text{(III.29)}$$

$$b_0 = -1/2 \, [(w_0 . \, x^*(1) \,)] + [(w_0 \, . \, x^*(-1) \,)] \quad \text{(III.30)}$$

où x*(1) est un vecteur support de la classe1, et x*(-1) un vecteur support de la classe –1.
la fonction de classement *class*(*x*) est définit par :

$$\text{Class}(x) = \text{sign} \, [(w_0 \, . \, x) + b_0]$$

(III.31)

$$= \text{sign}\left[\sum_{x_i \in VS} \alpha_i{}^0 y_i (x_i . x) + b_0\right] \quad \text{(III.32)}$$

Si *class*(*x*) est inférieure à 0, *x* est de la classe -1 sinon il est de la classe 1.

**III.4.2.2 Cas des données non-linéairement séparables :**

Dans ce cas où les données sont non-linéairement séparables Fig.III.7, l'hyperplan optimal est celui qui satisfait les conditions suivantes :

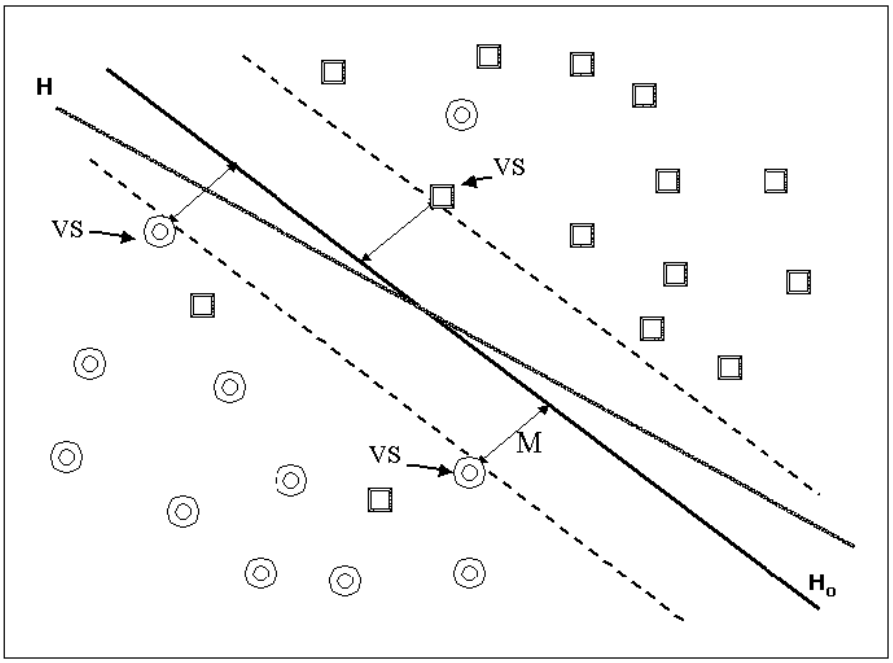


**Fig. III.7 - Hyperplans séparateurs dans le cas de données non-linéairement séparables (VS sont les Vecteurs Supports).**

- La distance entre les vecteurs bien classés et l'hyperplan optimal doit être maximale.

-la distance entre les vecteurs mal classés et l'hyperplan optimal doit être minimale.

Pour formaliser tout cela, on introduit des variables de pénalité non-négatives $\boldsymbol{\xi_i}$ pour i = 1,, m appelées variables d'écart. Ces variables transforment l'inégalité (III.19) comme suit :

$$yi\ (w.\ xi + b) >= 1 - \xi_i \quad \text{pour i = 1,…, m} \qquad \text{(III.33)}$$

L'objectif est de minimiser la fonction suivante :

$$\Psi(w, \Xi) = \frac{1}{2} w.w + C\sum_{i=1}^{m} \xi_i \qquad \text{(III.34)}$$

où C est un paramètre de régularisation. il permet de concéder moins d'importance aux erreurs. Cela mène à un problème dual légèrement différent de celui du cas des données linéairement séparables. Maximiser le lagrangien par rapport à $\alpha_i$ sous les contraintes suivantes :

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \quad \text{avec } 0 \le \alpha_i \ge C \quad \text{pour i=1,…, m.}$$

Le calcul de la normale $w_0$, du biais $b_0$ et de la fonction de classification *class*(*x*)reste exactement le même que pour le cas des données linéairement séparable.

### III.4.3 Principe des SVM :

Les classifieurs SVM utilisent l'idée de l'HO (Hyperplan Optimal) pour calculer une frontière entre des nuages de points. Elles projettent les données dans l'espace de caractéristiques en utilisant des fonctions non-linéaires. Dans cet espace on construit l'HO qui sépare les données transformées. L'idée principale est de construire une surface de séparation linéaire dans l'espace des caractéristiques qui correspond à une surface non-linéaire dans l'espace d'entrée. Le problème principal à relever ici est comment bien manipuler la transformation de tous les vecteurs d'entrée dans l'espace des caractéristiques de façon à éviter une augmentation du coût en nombre de paramètres libres. Soit l'ensemble *D'* l'image de l'ensemble *D*, défini dans la section (IV.5.1) par la transformation $\boldsymbol{\psi}$.

$$D' = \{(\psi(xi), yi) \in \mathfrak{R}^p \mathbf{x}\{-\mathbf{1}, \mathbf{1}\} \text{ pour } i = 1, .., m \mid p>=n\}$$

En construisant un HO dans l'espace des caractéristiques suivant la technique expliquée dans la section (IV.5.1 **)**On aura la fonction de classement suivante :

$$\text{Class(x)} = \text{sign}\left[\sum_{x_i \in VS} \alpha_i^{\,0} y_i \left(\Psi(x_i)\Psi(x)\right) + b_0\right] \qquad \text{(III.35)}$$

On peut remarquer que la fonction de classement dépend du produit scalaire dans l'espace des caractéristiques. Ainsi, pour que le coût de calcul reste pratiquement inchangé et le nombre de paramètres libres du système n'augmente pas, il faut que la fonction $\boldsymbol{\psi}$ satisfasse la condition suivante :

$$\Psi(u)\Psi(v) = K(u.v) \qquad \text{(III.36)}$$

C'est à dire le produit scalaire dans l'espace des caractéristiques va être représentable comme un noyau de l'espace d'entrée. Le classifieur est donc construit sans utiliser explicitement la fonction $\psi$.

Suivant la théorie de Hilbert-Schmidt, une famille de fonctions qui permet cette représentation et qui sont très appropriées aux besoins des SVM peut être définie comme l'ensemble des fonctions symétriques qui satisfont la condition suivante :

**III.4.3.1 Théorème (Mercer) :**

Pour être sûr qu'une fonction symétrique $K(u; v)$ admet un développement de la forme suivante :

$$K(u; v) = \sum_{k=1}^{+\infty} \beta_k \Psi_k(u).\Psi_k(v) \qquad \text{(III.37)}$$

tel que les $\beta_k \succ 0$ (i.e. $K(u; v)$ décrit un produit interne dans l'espace des caractéristiques) il est nécessaire et suffisant que la condition suivante soit satisfaite

$$\iint K(u,v)g(u)g(v)dudv \geq 0 \qquad \text{(III.38)}$$

pour toute fonction $g \neq 0$ avec :

$$\int g^2(z)dz \geq 0$$

On appelle ces fonctions les noyaux de Hilbert-Schmidt. Plusieurs noyaux ont été utilisés par les chercheurs, en voici quelques uns :

- Le noyau linéaire :

$$K(u; v) = u.v$$

- Le noyau Polynomial :

$$K(u; v) = [(u.v) + 1]^d$$

où d est le degré du polynôme à déterminer par l'utilisateur.

-Le noyau RBF (Radial Basis Function) :

$$K(u; v) = \exp\left(-\frac{\|u-v\|^2}{2\sigma^2}\right)$$

où $\sigma$ est à déterminer.

Maintenant que nous avons défini ce qu'est un noyau, la fonction de classement (III.35) devient :

$$\text{Class}(x) = \text{sign}\left[\sum_{x_i \in VS} \alpha_i^{\,0} y_i K(x_i.x) + b_0\right] \qquad \text{(III.39)}$$

### III.4.4 Extensions des SVMs :

#### III.4.4.1 SVM multi-classes :

Les SVM multi-classe sont des modèles de l'apprentissage de conception relativement récente, dont l'étude est actuellement en plein essor. Ceci résulte en premier lieu du fait que la communauté des théoriciens, qui avait jusque dans un passé récent consacré l'essentiel de ses forces au développement de la théorie statistique du calcul des dichotomies, exprime à présent un intérêt de plus en plus marqué pour le cas multi-classes, dont elle perçoit mieux les spécificités. Cette situation nouvelle fait naître un besoin, celui de disposer d'une étude synthétique sur les SVM multi-classes, ou plus généralement l'utilisation de SVM pour la discrimination à catégories multiples.

##### III.4.4.1.1 Approches "un contre tous" :

L'approche "un contre tous" est la plus simple et la plus ancienne des méthodes de décomposition. Elle consiste à utiliser un classifieur binaire (à valeurs réelles) par catégorie. Le k-ième classifieur est destiné à distinguer la catégorie d'indice k de toutes les autres. Pour affecter un exemple, on le présente donc à Q classifieurs, et la décision s'obtient en application du principe "winner-takes-all" : l'étiquette retenue est celle associée au classifieur ayant renvoyé la valeur la plus élevée.. Les auteurs soutiennent la thèse selon laquelle cette approche, aussi simple soit-elle, lorsqu'elle est mise en œuvre avec des SVM correctement paramétrées, obtient des performances qui ne sont pas significativement inférieures à celles des autres méthodes de décomposition et des SVM multi-classes actuelles.

Il convient cependant de souligner qu' elle implique d'effectuer des apprentissages aux répartitions entre catégories très déséquilibrées, ce qui soulève souvent des difficultés pratiques.

**III.4.4.1.2 Approches "un contre un" :**

Une autre approche appelée "un contre un". Ordinairement attribuée à Knerr et ses co-auteurs , elle consiste à utiliser un classifieur par couple de catégories. Le classifieur indicé par le couple (k, l) (avec 1 < k < l < Q), est destiné à distinguer la catégorie d'indice k de celle d'indice l. Pour affecter un exemple, on le présente donc à $C_Q^2$ classifieurs, et la décision s'obtient habituellement en effectuant un vote majoritaire ("max-wins voting").
Sous l'hypothèse que la frontière séparant une catégorie d'une autre peut être moins complexe que celle séparant cette même catégorie de toutes les autres, il y voit un moyen d'obtenir des estimateurs présentant un biais plus faible qu'avec l'approche un contre tous. Naturellement, le prix à payer est un possible accroissement de la variance de ces estimateurs, compte tenu du fait que les bases d'apprentissage de chacun des classifieurs sont plus petites que l'échantillon initial.

**III.5 Conclusion :**

Dans ce chapitre, nous avons décrit quelques méthodes à noyaux parmi les plus prometteuses pour effectuer les tâches de discrimination et de classification telles que les GMM, réseau de neurones de type GRNN et SVM (binaire et multi-classes). Nous avons détaillé leurs principes afin de faciliter leur utilisation dans la RAL, objet du prochain chapitre.

# IV Mise en œuvre d’un Système de Reconnaissance Automatique de Locuteurs Basé sur les Méthodes à Noyaux.

# Chapitre IV :

# Mise en œuvre d'un Système de Reconnaissance Automatique de Locuteurs Basé sur les Méthodes à Noyaux

## IV.1 Introduction :

Dans ce chapitre nous allons présenter les résultats obtenus avec un système de reconnaissance automatique du locuteur que nous avons élaboré et qui est basé sur les méthodes à noyaux. Dans ce système, la tache de reconnaissance est dévolue aux SVMs , GRNN et au GMM.

Les protocoles de développement et d'évaluation des différentes techniques de modélisation pour l'identification du locuteur sont décrits dans cette section. Ils mettent en jeu des modules d'extraction des paramètres acoustiques à savoir MFCC et LSF. Les expériences ont été menées en mode dépendant et indépendant du texte. L'influence de l'environnement a également été évaluée par simulation de différents milieux bruités.

## IV.2 Protocole expérimental :

Le protocole expérimental retenu est résumé en fig IV.1 suivante :

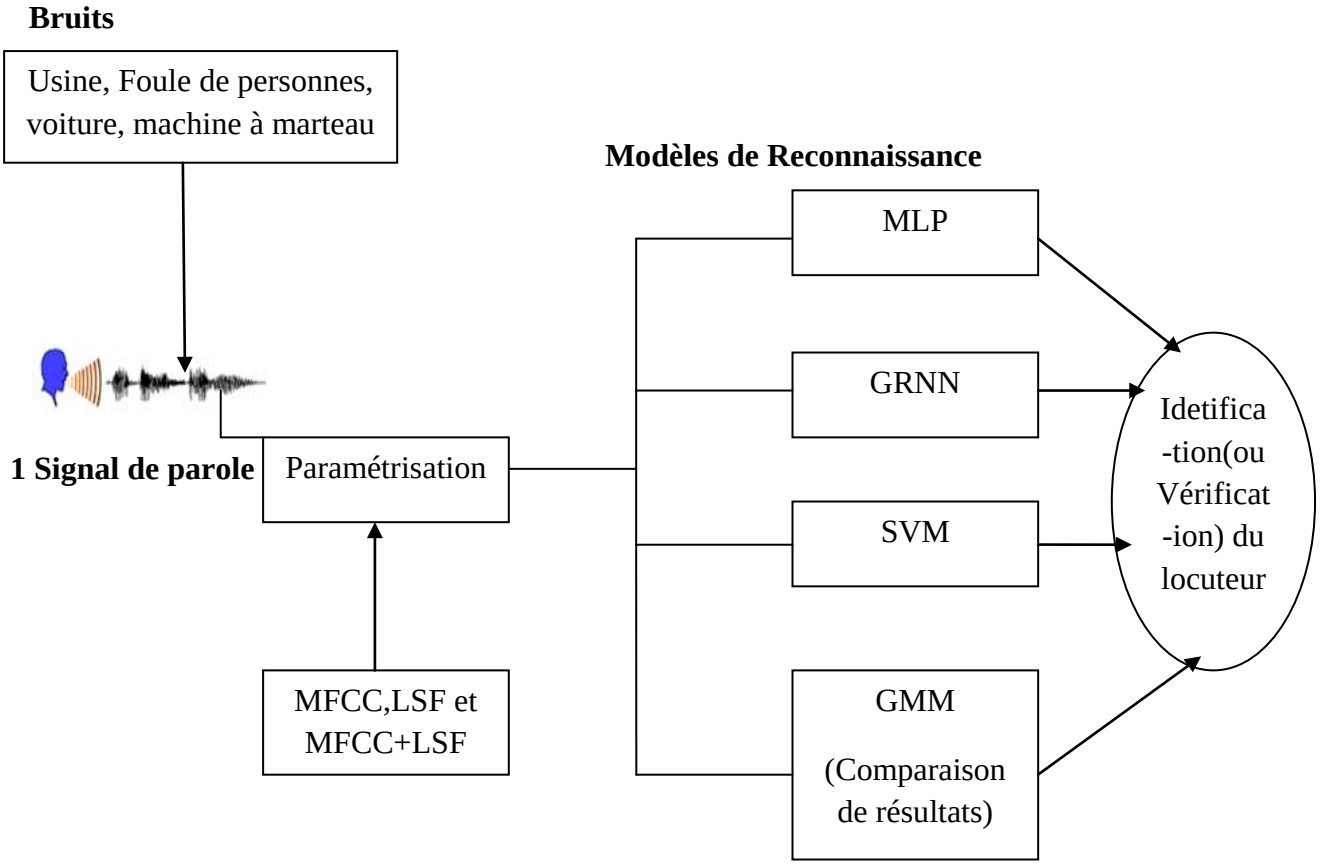


**Fig.IV.1 Schéma général de notre système RAL**

Nous testons nos systèmes dans un milieu clean sur les données de **ARADIGITS**, en utilisant le protocole de développement définit comme suit :

### IV.3 La base de données sonores :ARADIGITS

La base de données parole utilisée dans ce travail est une partie de la base de données **ARADIGITS**. Elle est constituée des 10 chiffres de la langue Arabe (zéro jusqu'à neuf) prononcés par 60 locuteurs des deux sexes avec trois répétitions pour chaque chiffre. Cette base a été enregistrée par des locuteurs algériens de différentes régions âgés entre 18 et 50 ans dans un environnement calme avec un niveau de bruit ambiant inférieur à 35 dB, sous le format WAV, avec une fréquence d'échantillonnage égale à 16 KHz.

Dans nos expériences, une analyse est appliquée toutes les 10 ms sur des fenêtres d'analyse de 20 ms (par glissement et recouvrement des fenêtres d'analyse). A chaque trame, on associe un vecteur de représentation acoustique. Douze (12) coefficients LSF et Douze (12) MFCC sont

calculés pour chaque trame à partir d'un banc de 24 filtres répartis dans l'échelle fréquentielle Mel. Des coefficients différentiels du premier ordre ($\Delta$) et du second ordre ($\Delta\Delta$) sont ensuite calculés pour former un vecteur de dimension 36 (12 MFCC + $\Delta$12 MFCC+ $\Delta\Delta$12 MFCC).

### IV. 3.1 Identification du locuteur :

Pour évaluer notre système de reconnaissance en ensemble fermé, nous avons structuré cette base selon la hiérarchie suivante :

- B.A et B.T contiennent dix (10) locuteurs chacune (5 masculins et 5 féminins).
- B.A et B.T contiennent trente (30) locuteurs chacune (15 masculins et 15 féminins).
- B.A et B.T contiennent soixante (60) locuteurs chacune (30 masculins et 30 féminins).

B.A : base d'apprentissage.

B.T : base de test.

Les bases d'apprentissage et de test sont disjointes.

### IV.3.2 Vérification du locuteur :

Pour la vérification du locuteur, en complément de la base initiale, nous avons utilisé comme base de test supplémentaire un ensemble de trente (30) locuteurs différents de ceux de la base d'apprentissage (pour simuler les Imposteurs).

### IV. 3. 3 Identification en mode dépendant du texte :

- Pour la base d'apprentissage :

Chaque locuteur prononce les chiffres (0 à 9) en arabe deux (2) fois.

- Pour la base de test :

Chaque locuteur prononce les chiffres (0 à 9) en arabe une fois.

### IV. 3. 4 Identification en mode indépendant du texte :

- Pour la base d'apprentissage :

Chaque locuteur prononce les chiffres (0 à 7) en arabe trois (3) fois.

- Pour la base de test :

Chaque locuteur prononce les chiffres (8 et 9) en arabe deux fois.

### IV. 3.5 Concaténation des paramètres :

Pour toutes les expériences une combinaison entre les coefficients MFCC et LSF a été effectuée afin de mesurer l'influence de cette fusion sur le taux de reconnaissance.

### IV.4 Résultats expérimentaux :

Nous avons appliqué ce protocole aux systèmes de RAL en utilisant les reconnaisseurs suivants :

1. le réseau de neurones**, GRNN** et le MLP.
2. les supports vecteurs machines (**SVM**) avec trois(3) différents noyaux (**Linéaire** , **polynomial** avec degré égal à 2 et le noyau **RBF**).
3. Les GMMs à titre de comparaison.

Les résultats obtenus de ces systèmes sont mis sous forme de tableaux et de graphes ci-dessous :

### IV.4.1 Vérification du locuteur :

Tab. 4.1 - Taux de fausse acceptation du locuteur selon le nombre de locuteurs et le classifieur utilisé.

| Locuteurs / Classifieur | 10 | 30 | 60 |
|---|---|---|---|
| MLP | 0% | 0% | 0% |
| GRNN | 0% | 0.89% | 0% |
| SVM(noyauLinéaire) | 3.33% | 10% | 0% |
| SVM(noyauPolyd=2) | 0% | 3.33% | 0% |
| SVM(noyau RBF) | 0% | 0% | 0% |

### IV.4.2 Identification du locuteur :

#### IV.4.2.1 En mode dépendant du texte :

Tab.4.2 - Taux d'identification du locuteur selon le nombre de locuteurs et le classifieur utilisé.

| Locuteurs / Classifieur | 10 | 30 | 60 |
|---|---|---|---|
| MLP | 90% | 89.73% | 94.26% |
| GRNN | 25% | 17.33% | 21.54% |
| SVM(noyauLinéaire) | 80 .89% | 83% | 81.30% |
| SVM(noyauPolyd=2) | 80 .89% | 80.03% | 81.11% |
| SVM(noyau RBF) $\sigma = 0.5$ | 80 .89% | 84% | 82.56% |

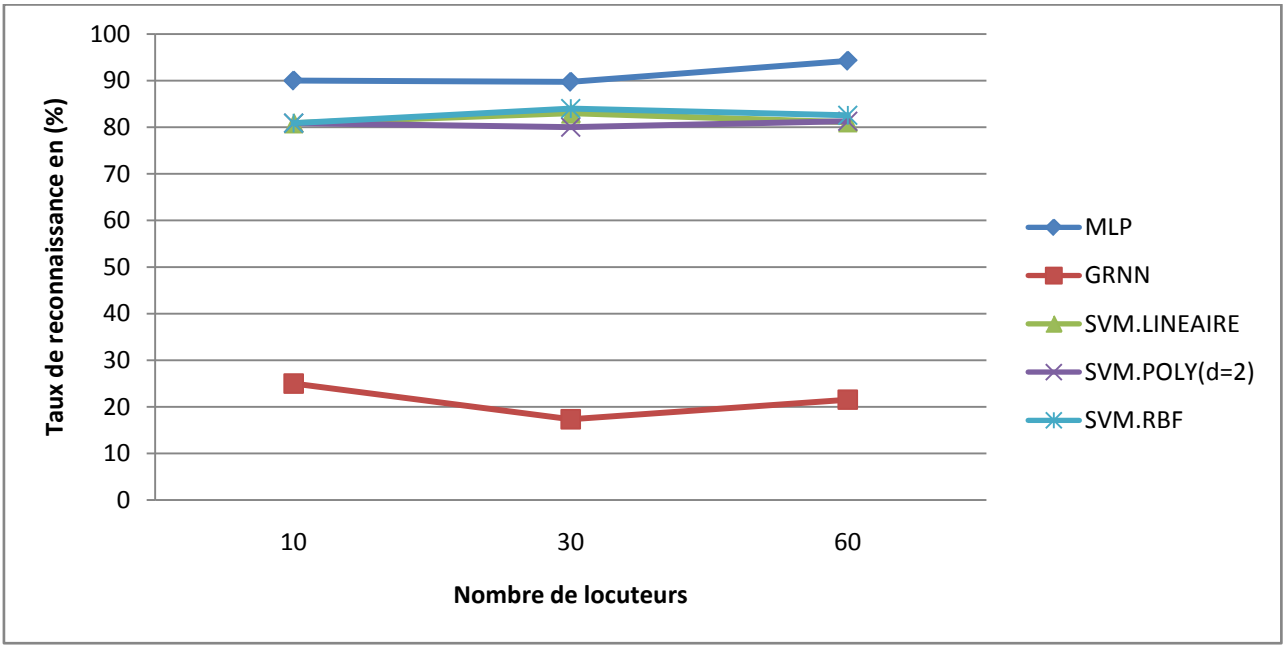


**Fig. IV.2 Influence du nombre de locuteurs sur le taux d' identification .**

### IV.4.2.2 En mode indépendant du texte :

Tab.4.3- Taux d'identification du locuteur selon le nombre de locuteurs et le classifieur utilisé

| Locuteurs / Classifieur | 10 | 30 | 60 |
|---|---|---|---|
| MLP | 20% | 03.65% | 2.22% |
| GRNN | 24% | 20% | 15.65% |
| SVM(noyauLinéaire) | 50% | 89.17% | 88.33% |
| SVM(noyauPolyd=2) | 72.5% | 87.5% | 82.78% |
| SVM(noyau RBF) | 97.5% ($\sigma = 0.001$) | 92.5% ($\sigma = 0.001$) | 86.36% ($\sigma = 0.003$) |

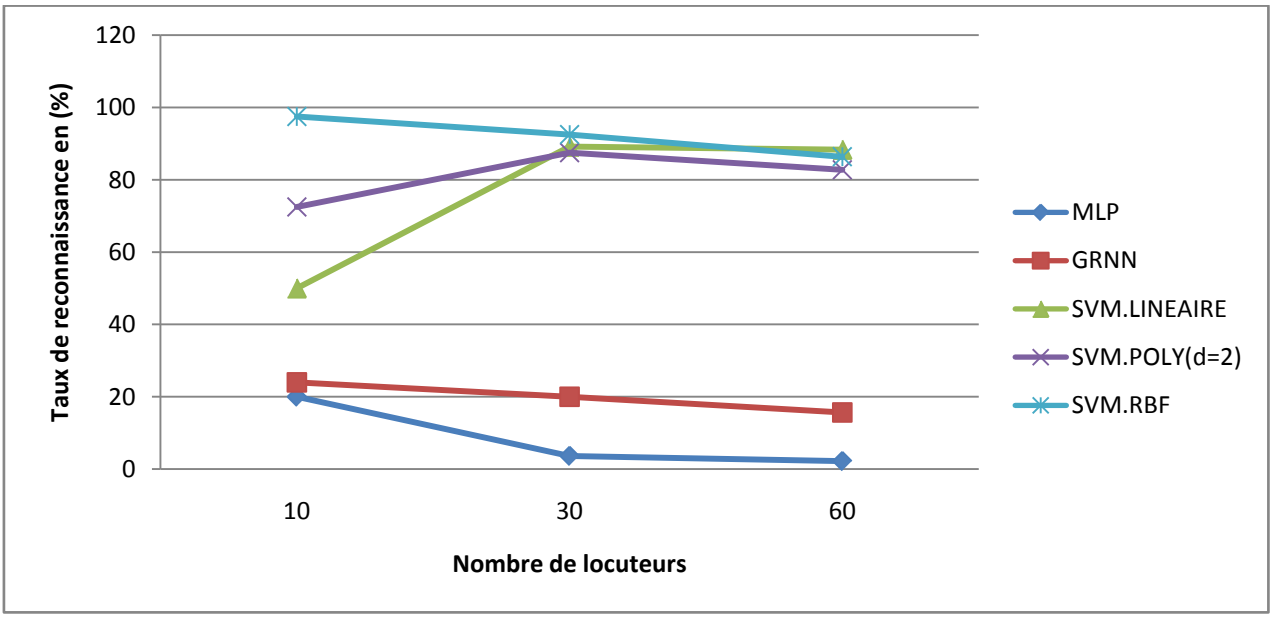


**Fig. IV.3 Influence du nombre de locuteurs sur le taux d'identification.**

### IV.4.3 Identification du locuteur dans un milieu bruité :

#### IV.4.3.1 En mode indépendant du texte :

Dans cette partie, nous bruitons notre base de test **ARADIGITS** qui contient 60 locuteurs (30 masculins et 30 féminins) dont chacun d'eux prononce les chiffres 8 et 9 en arabe par des bruits différents (voiture, speech babble, usine et machine à pistolet) issus de la base NOISEX'92 NATO (Varga).

Les graphes ci-dessous montrent les formes temporelles et spectrogrammes des signaux de parole d'une locutrice prononçant le digit [Tamania] (Chiffre 8 en arabe ) dans un milieu clean (Fig .IV.4 ) ,d'un bruit d'une usine., machine à marteau et voiture. Les figures Fig. IV.6, Fig. IV.7, Fig. IV.8 représentent la forme précédente bruitée avec un bruit d'une usine à différents SNRs.

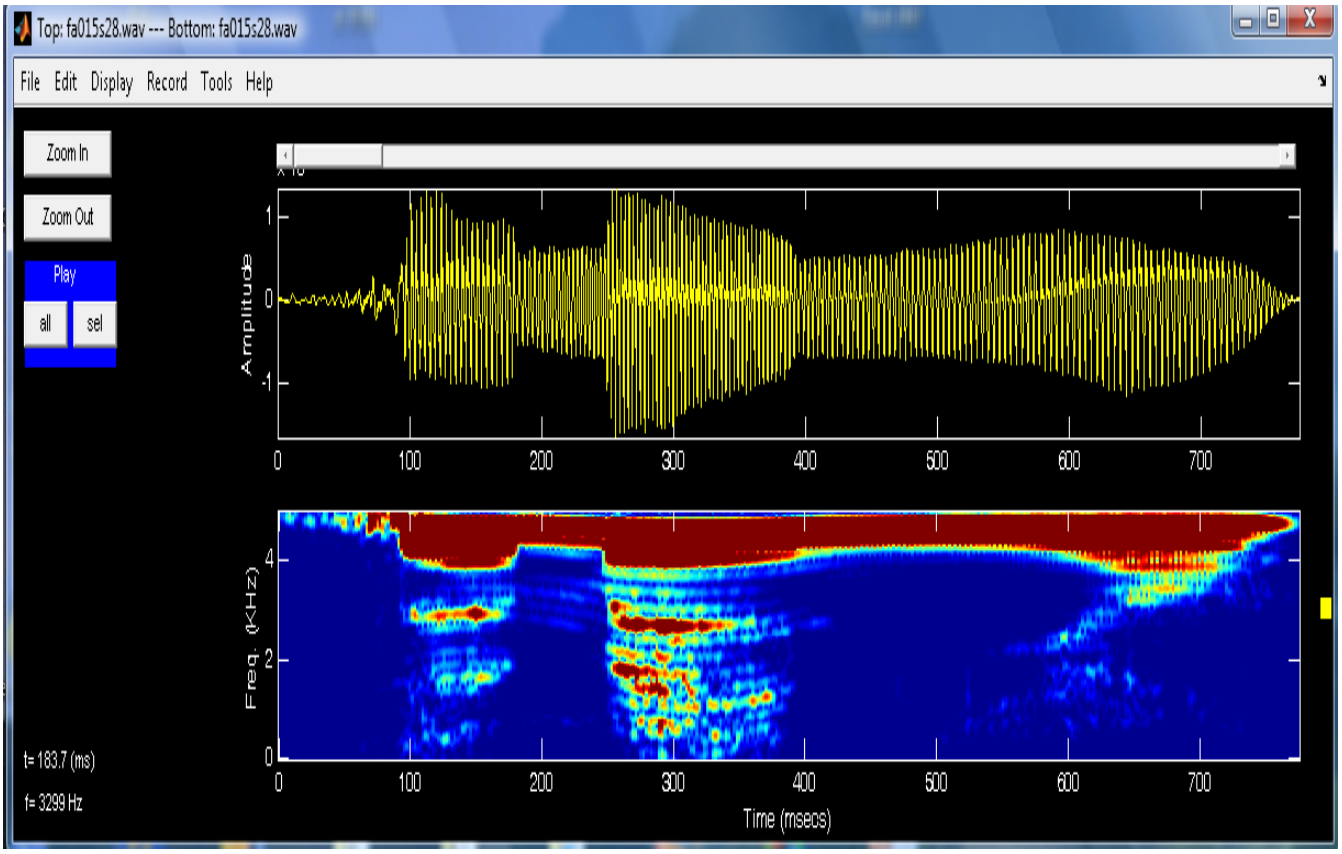


**Fig. IV.4 Forme temporelle et spectrogramme d'un signal de parole du mot [tamania] prononcé par une locutrice.**

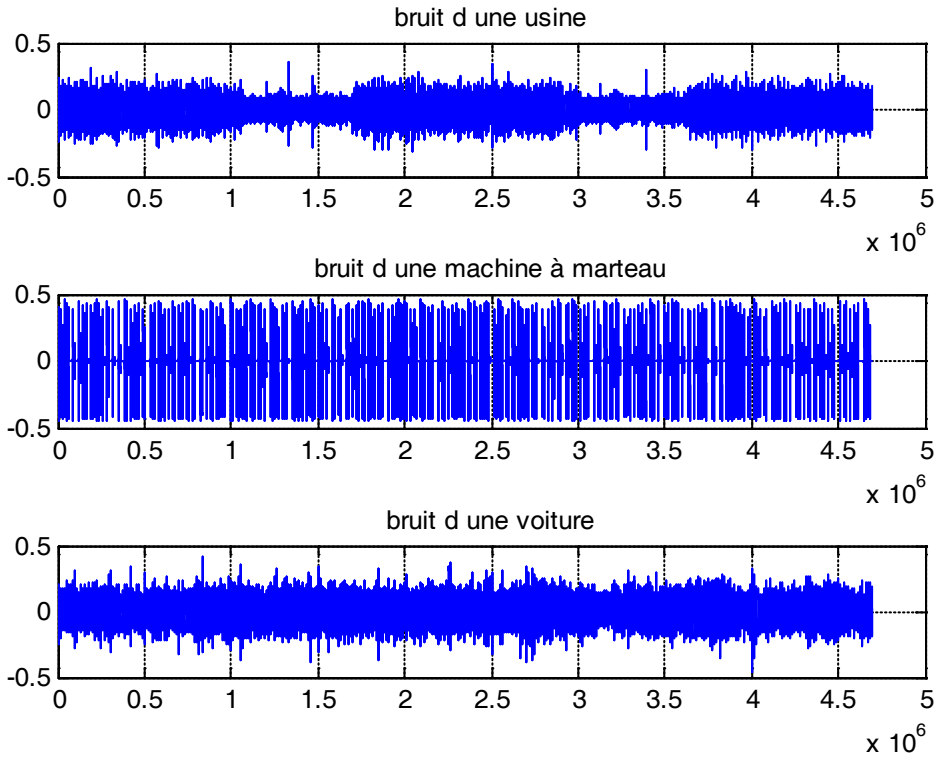


**Fig.IV.5 La forme temporelle des bruits de :Usine, Machine à marteau et Voiture.**

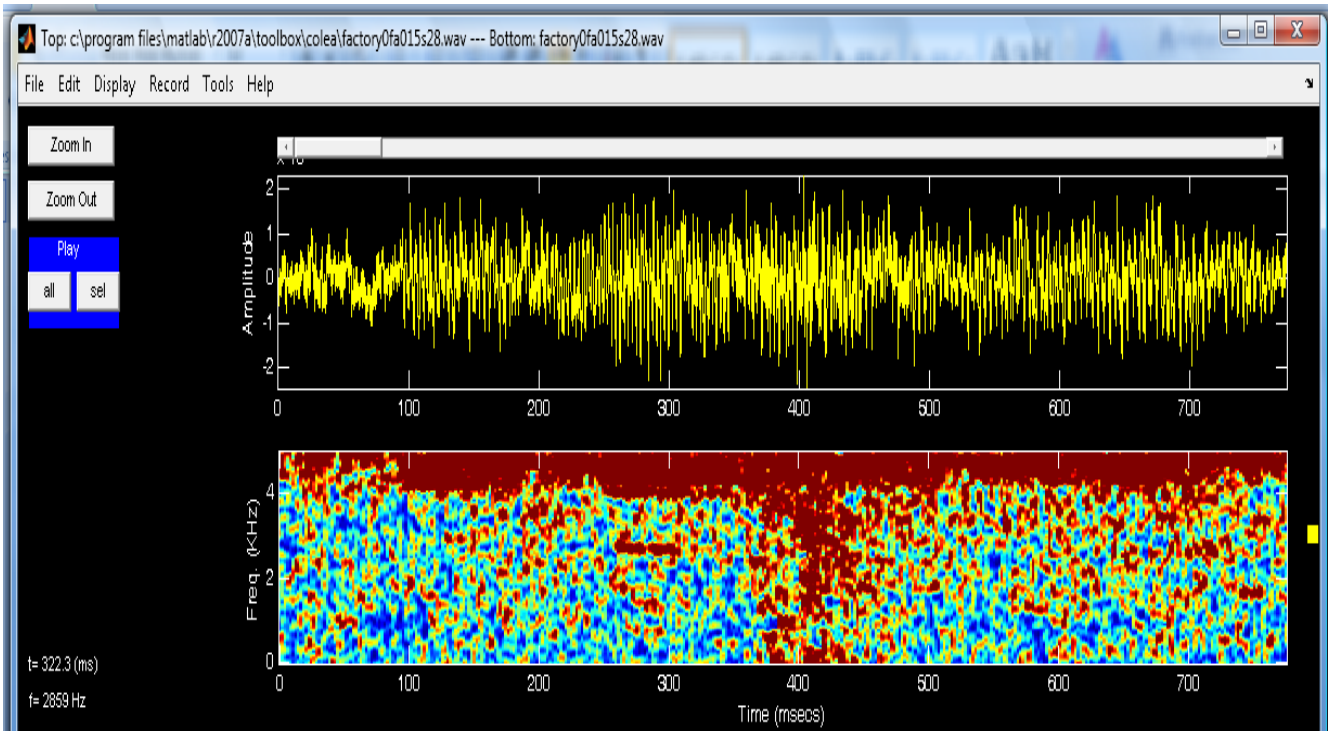


**Fig. IV.6 Forme temporelle et spectrogramme du mot [tamania ]bruité avec le bruit d'une usine à SNR égal à 0.**

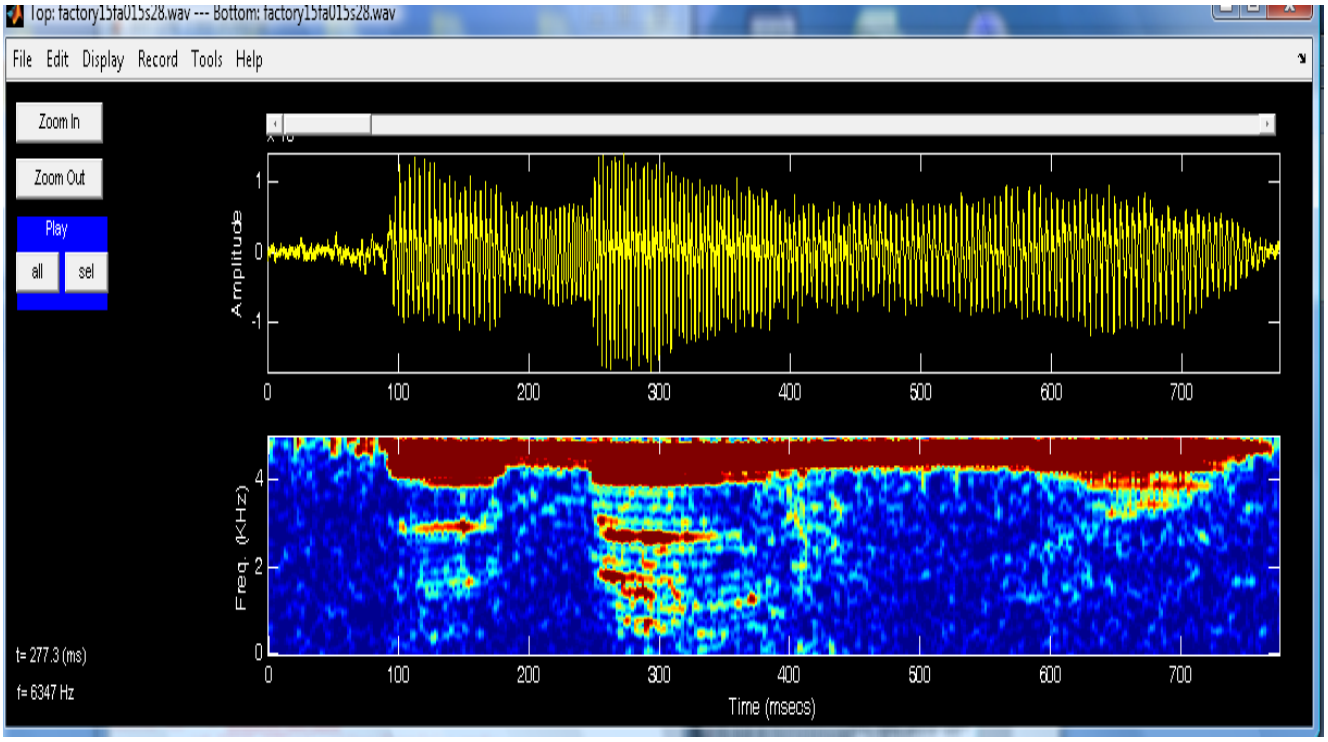


**Fig. IV.8 Forme temporelle et spectrogramme du mot [tamania ]bruité avec le bruit d'une usine à SNR égal à 15.**

Les résultats obtenus sont résumés ci-dessous sous forme de tableaux et graphes.Sachant que le SNR utilisé dans l'opération de bruitage est un SNR global.

Tab.4.4 Taux d'identification en (%) en présence d'un bruit d'une voiture (Volvo)

| SNR(dB) / Classifieur | 0 | 5 | 10 | 15 | 20 |
|---|---|---|---|---|---|
| MLP | 0.33 | 0.75 | 1.32 | 1.43 | 2.15 |
| GRNN | 2 | 3.5 | 8.11 | 9 | 11.33 |
| SVM(noyau Linéaire) | 65 | 68.33 | 71.25 | 74.58 | 77.92 |
| SVM(noyau Polyd=2) | 67.5 | 71.67 | 72.92 | 77.5 | 77.92 |
| SVM(noyau RBF) | 65.83 | 67.5 | 70.83 | 72.08 | 72.5 |

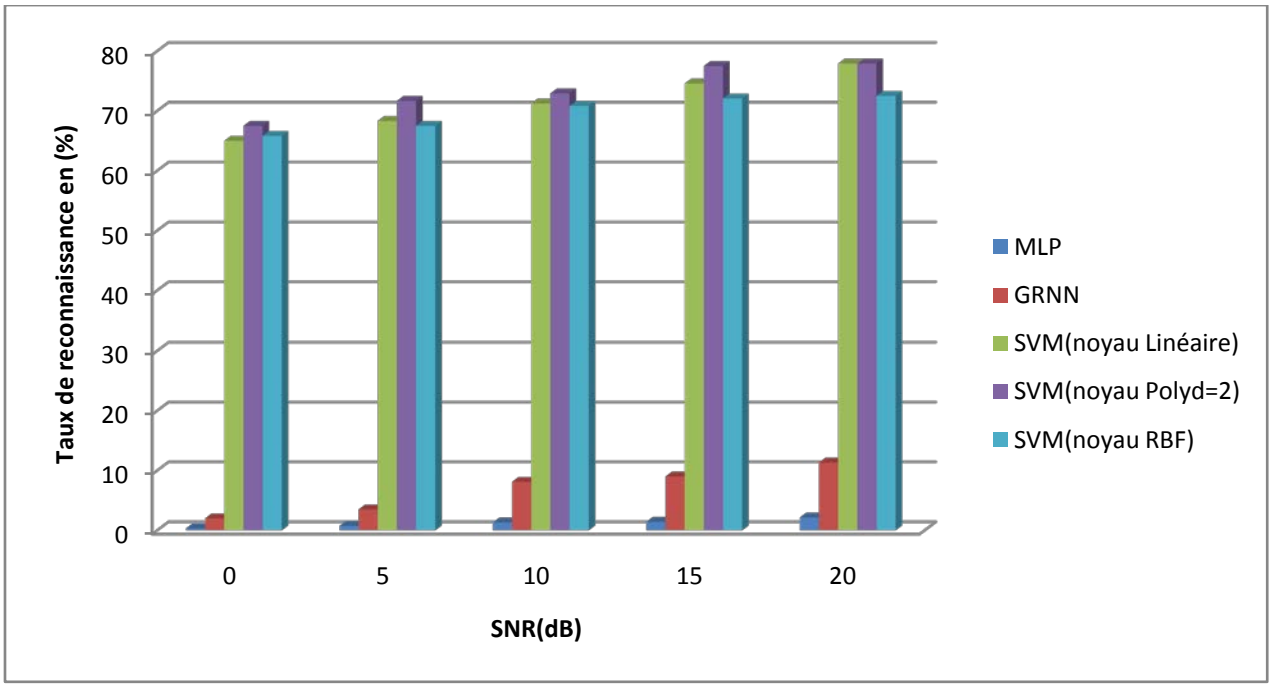


**Fig. IV.9 Histogrammes de taux d'identification en (%) en présence de bruit d'une voiture (volvo)**

Tab.4.5 Taux d'identification en (%) en présence de bruit d'une foule de personnes (babble speech).

| SNR(dB) / Classifieur | 0 | 5 | 10 | 15 | 20 |
|---|---|---|---|---|---|
| MLP | 0.08 | 0.33 | 1.3 | 1.56 | 2.05 |
| GRNN | 3.5 | 5.28 | 9.8 | 10.5 | 7.85 |
| SVM(noyauLinéaire) | 8.33 | 66.67 | 72.5 | 76.25 | 80.42 |
| SVM(noyauPolyd=2) | 7.33 | 66.26 | 75 | 78.75 | 80.42 |
| SVM(noyau RBF) | 9.33 | 62.08 | 67.92 | 69.17 | 72.08 |

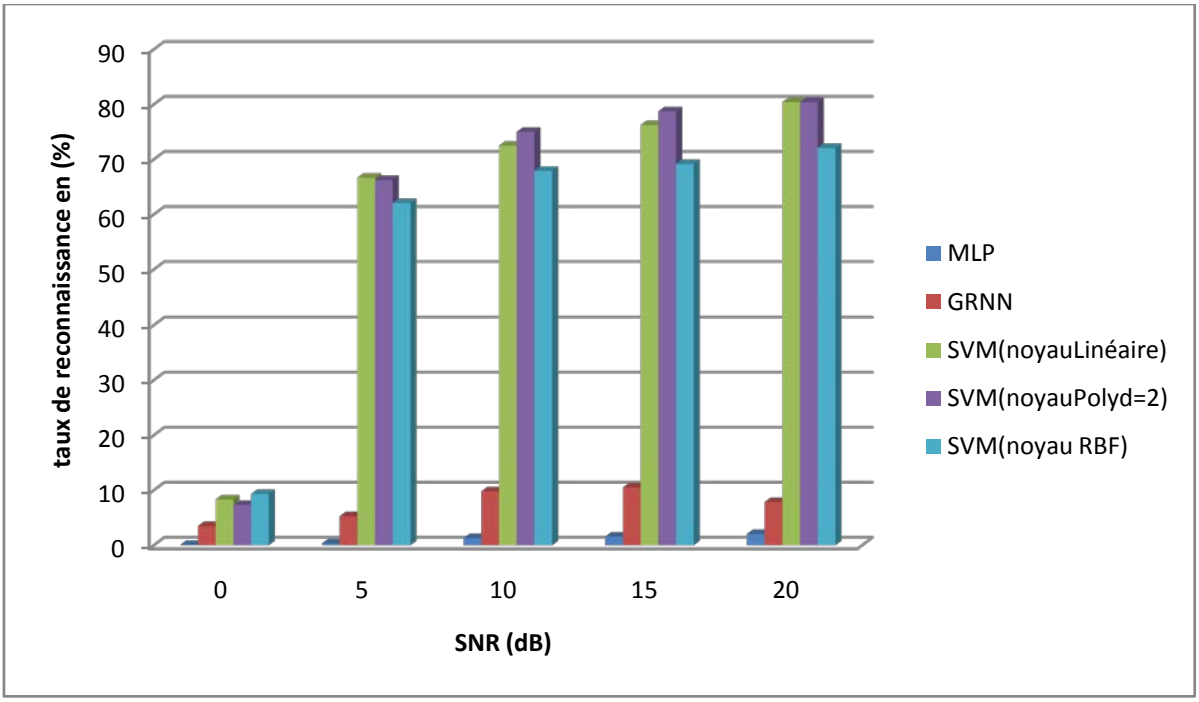


**Fig. IV.10 Histogrammes de taux d'identification en (%) en présence de bruit d'une foule de personnes (babble speech)**

Tab.4.6 Taux d'identification en (%) en présence de bruit d'une machine à marteau (machine gun)

| SNR (dB) / Classifieur | 0 | 5 | 10 | 15 | 20 |
|---|---|---|---|---|---|
| MLP | 0.08 | 0.67 | 0.83 | 1.25 | 2.18 |
| GRNN | 4 | 4.65 | 5.58 | 6.22 | 8.51 |
| SVM(noyauLinéaire) | 5.83 | 60.42 | 66.25 | 70 | 75 |
| SVM(noyauPolyd=2) | 1.67 | 3.33 | 8.33 | 66.25 | 74.17 |
| SVM(noyau RBF) | 3.75 | 4.58 | 7.5 | 10 | 64.58 |

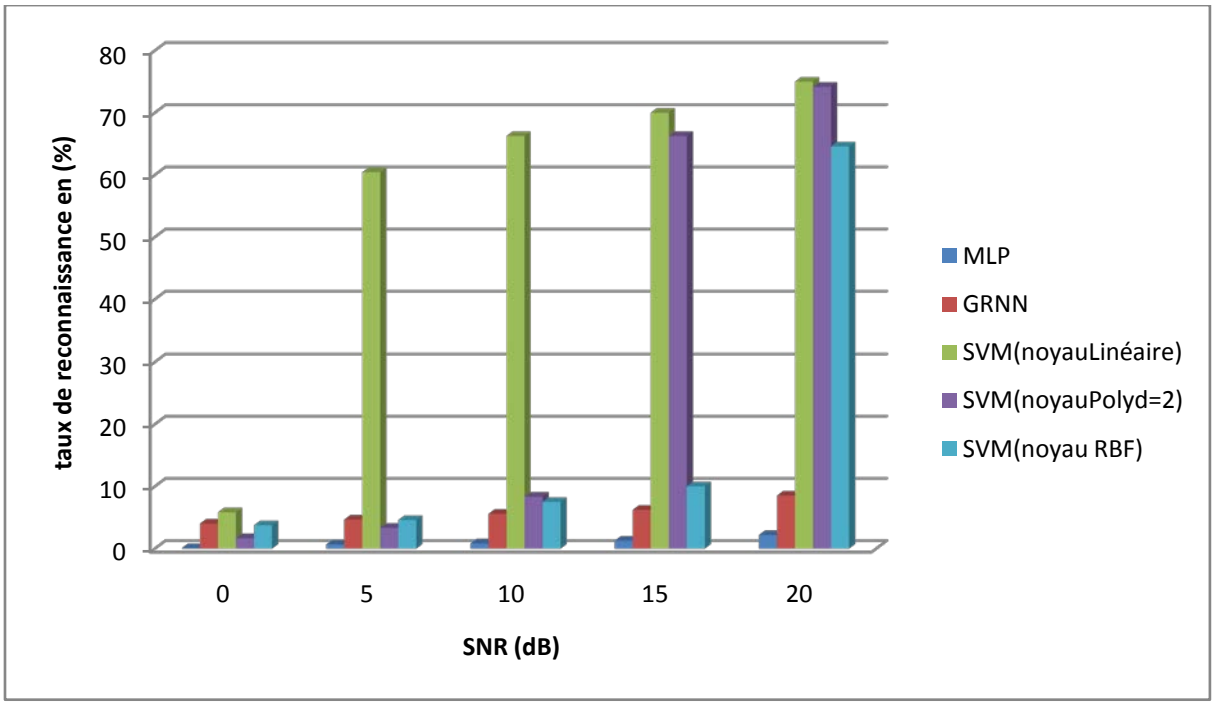


**Fig. IV.11 Histogrammes de taux d'identification en (%) en présence de bruit d'une machine à marteau (machinegun)**

Tab.4.7 Taux d'identification en (%) en présence de bruit d'une usine (factory)

| SNR (dB) / Classifieur | 0 | 5 | 10 | 15 | 20 |
|---|---|---|---|---|---|
| MLP | 0.8 | 1.15 | 1.83 | 1.95 | 2.2 |
| GRNN | 5.6 | 6.95 | 7.49 | 10.75 | 11.67 |
| SVM(noyauLinéaire) | 7.13 | 61.82 | 64.36 | 71 | 78.2 |
| SVM(noyauPolyd=2) | 3.61 | 5.85 | 57.76 | 65.15 | 73.67 |
| SVM(noyau RBF) | 4.15 | 6.83 | 38.42 | 42 | 65 |

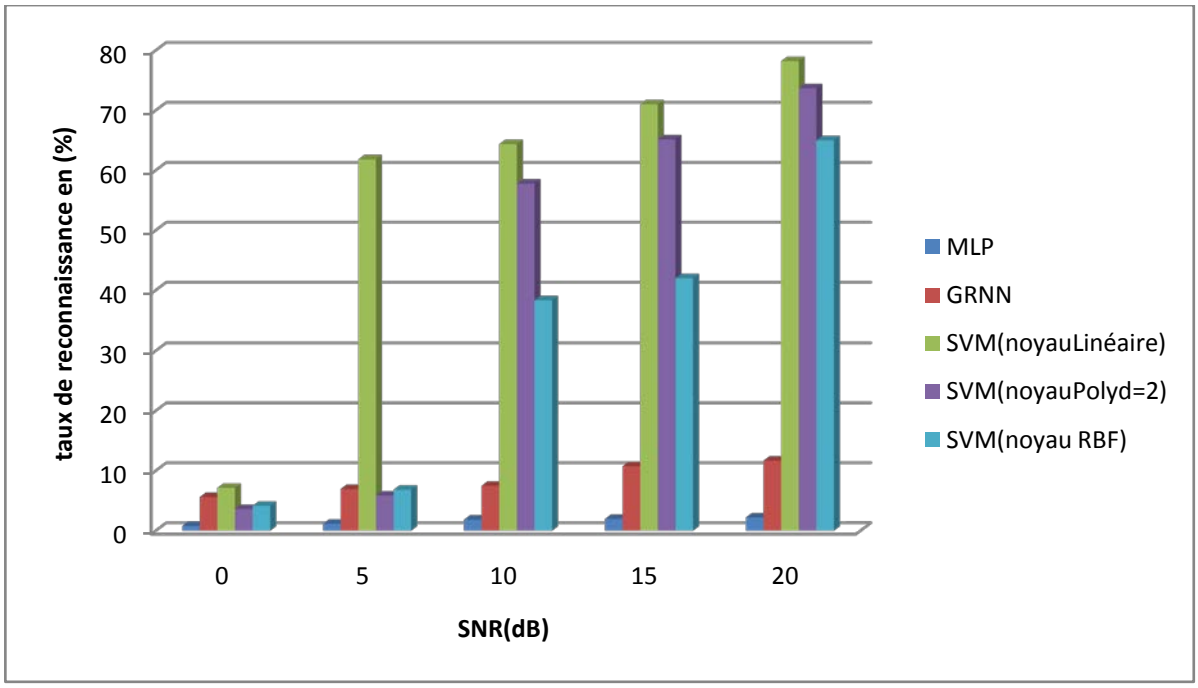


**Fig. IV.12- Histogrammes de Taux d'identification en (%) en présence de bruit d'une usine (factory)**

### IV.4.4 Concaténation des paramètres (36 Coefficients MFCC + 12 Coefficients LSF) :

### IV.4.4.1 Identification du locuteur :

#### IV.4.4.1.1 En mode dépendant du texte :

Tab.4.8 Influence du nombre de locuteurs sur le taux d'identification.

| Locuteurs / Classifieur | 10 | 30 | 60 |
|---|---|---|---|
| MLP | 92% | 90.73% | 92.26% |
| GRNN | 28% | 19.33% | 25.54% |
| SVM(noyauLinéaire) | 88 .89% | 87% | 86.11% |
| SVM(noyauPolyd=2) | 88 .89% | 84.03% | 82.30% |
| SVM(noyau RBF) $\sigma = 0.5$ | 88 .89% | 87.52% | 84.56% |

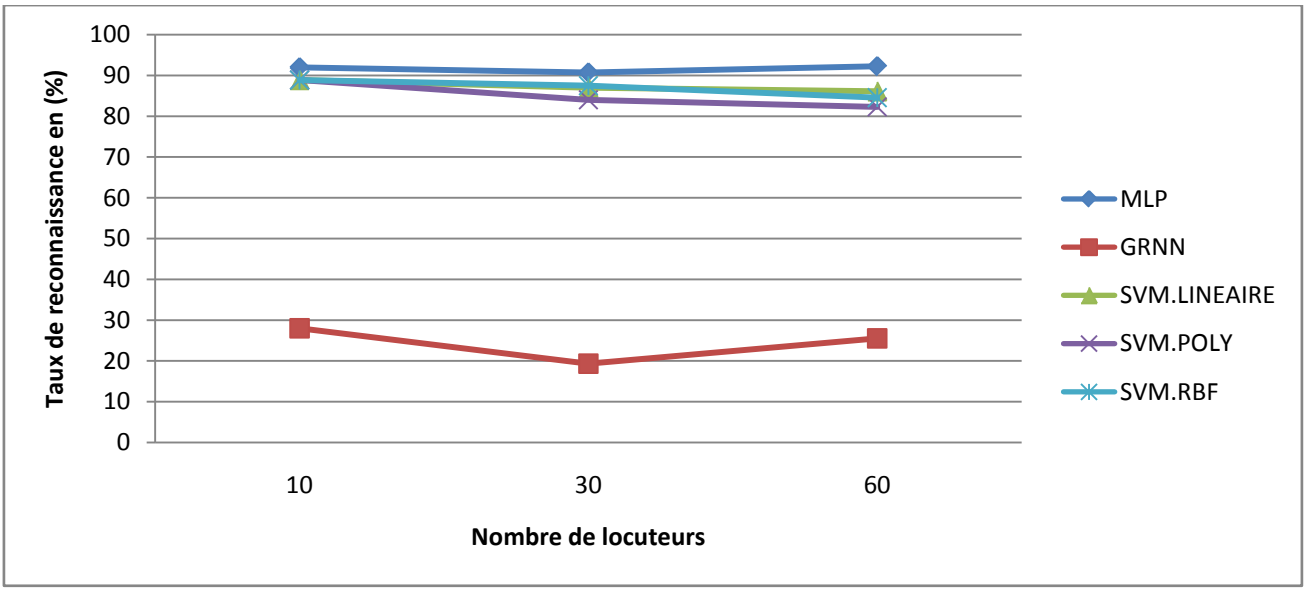


**Fig. IV.13 Influence du nombre de locuteurs sur le taux d'identification.**

### IV.4.4.1.2 En mode indépendant du texte :

Tab.4.9 Taux d'identification du locuteur selon le nombre de locuteurs et le classifieur utilisé.

| Locuteurs / Classifieur | 10 | 30 | 60 |
|---|---|---|---|
| MLP | 26% | 13.45% | 14.32% |
| GRNN | 32.33% | 25.78% | 18.56% |
| SVM(noyauLinéaire) | 89.41% | 91.12% | 89.73% |
| SVM(noyauPolyd=2) | 76.25% | 89.57% | 83.78% |
| SVM(noyau RBF) $\sigma = 0.5$ | 72.1% | 77.22% | 85.84% |

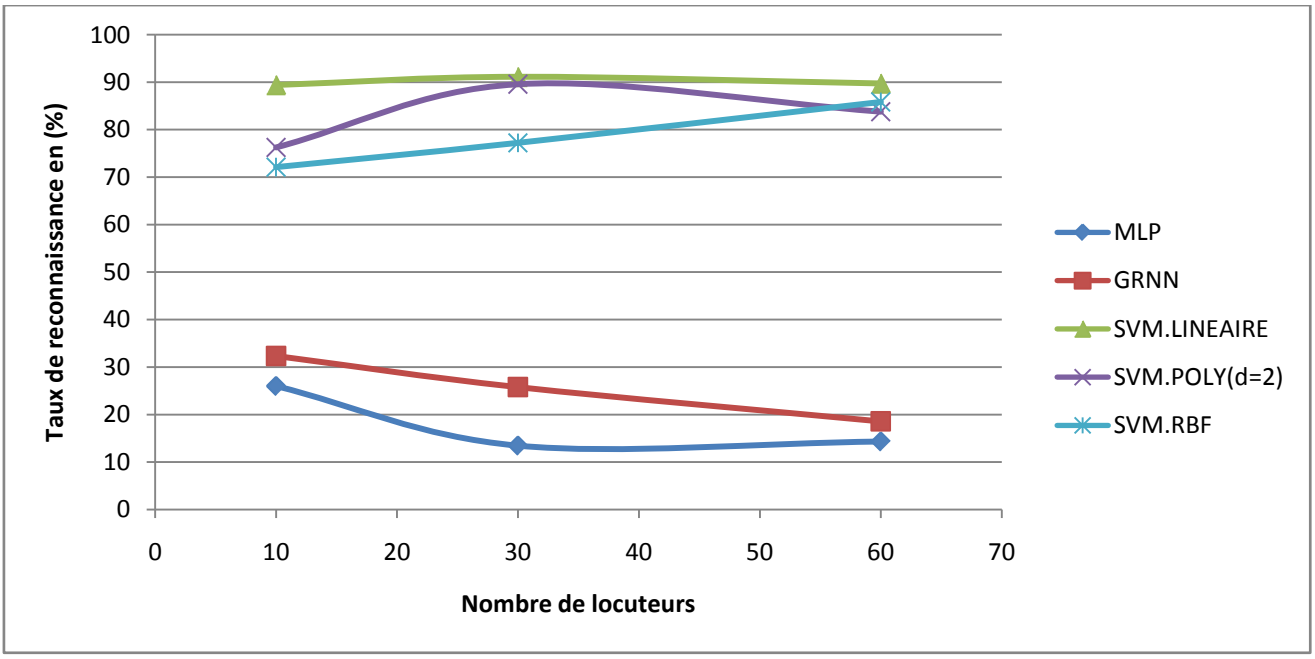


**Fig. IV.14 Influence du nombre de locuteurs sur le taux d'identification .**

### IV.4.5 Milieu bruité :

#### IV.4.5.1 En mode indépendant du texte :

Tab.4.10 Taux d'identification en (%) en présence de bruit d'une foule de personnes (babble speech)

| SNR (dB) / Classifieur | 0 | 5 | 10 | 15 | 20 |
|---|---|---|---|---|---|
| MLP | 3.08 | 6.33 | 9.22 | 11.25 | 13.75 |
| GRNN | 6.15 | 9.29 | 11.83 | 14.53 | 17.85 |
| SVM(noyauLinéaire) | 9.13 | 68.87 | 74.15 | 79.65 | 85.42 |
| SVM(noyauPolyd=2) | 10.33 | 68.26 | 77.17 | 79.75 | 80.22 |
| SVM(noyau RBF) | 9.31 | 65.05 | 68.97 | 72.27 | 82.88 |

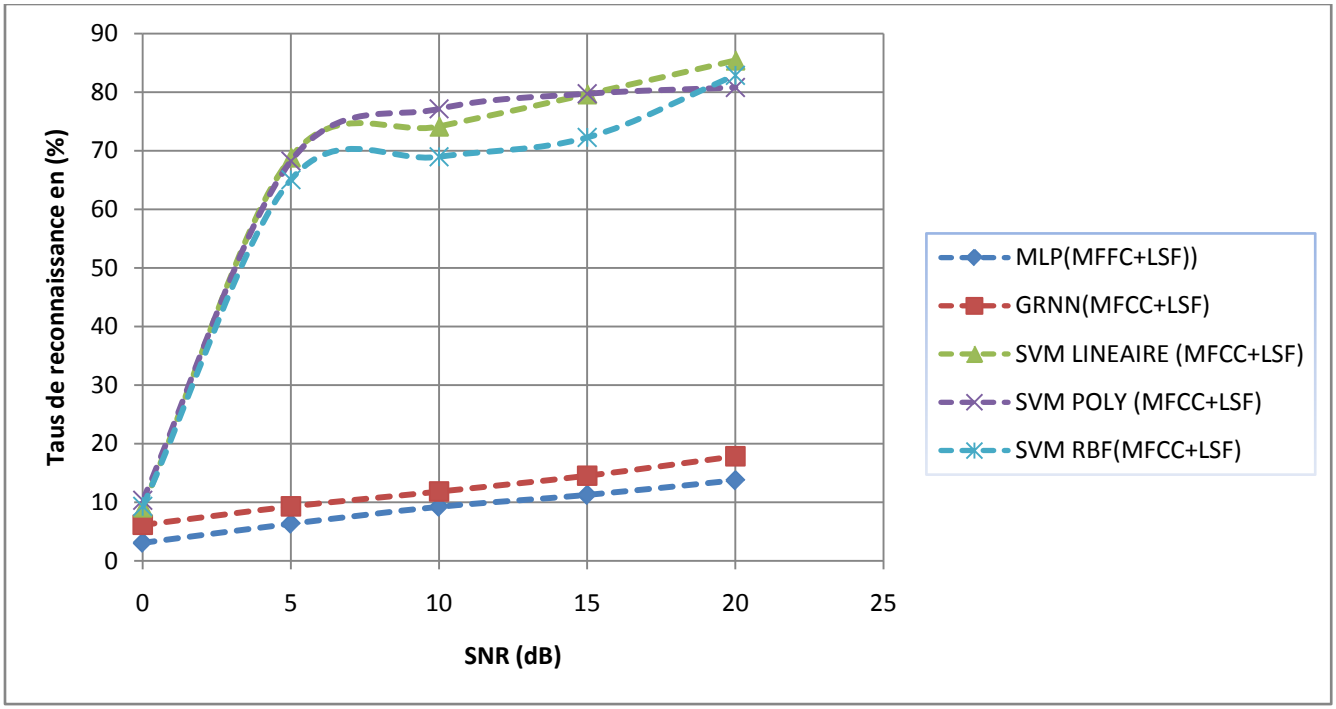


**Fig. IV.15 Evolution de Taux d'identification en (%) selon le classifieur utilisé et la valeur du SNR**

### IV.4.6 Comparaison entre MLP , GRNN , SVM et GMM :

Comparons nos résultats avec ceux obtenus dans les travaux de **KROBBA.A** [**30**], sur la reconnaissance automatique du locuteur par GMM en utilisant même base de données utilisée dans ce travail avec mêmes conditions et même hiérarchie (base d'apprentissage et test comporte 60 locuteurs chacune), en mode indépendant du texte, nous obtenons les tableaux et le graphes suivants :

### IV.4.6.1 Milieu clean :

Tab.4.11 Comparaison des taux d'identification des différents modèles utilisés avec GMM dans un milieu clean.

| | MLP | GRNN | SVM.LINAIRE | SVM.POLYd=2 | SVM.RBF | GMM(32) |
|---|---|---|---|---|---|---|
| Taux d'identification en (%) | 2.22 % | 15.65 % | 88.33% | 82.78% | 86.36% | 89.34% |

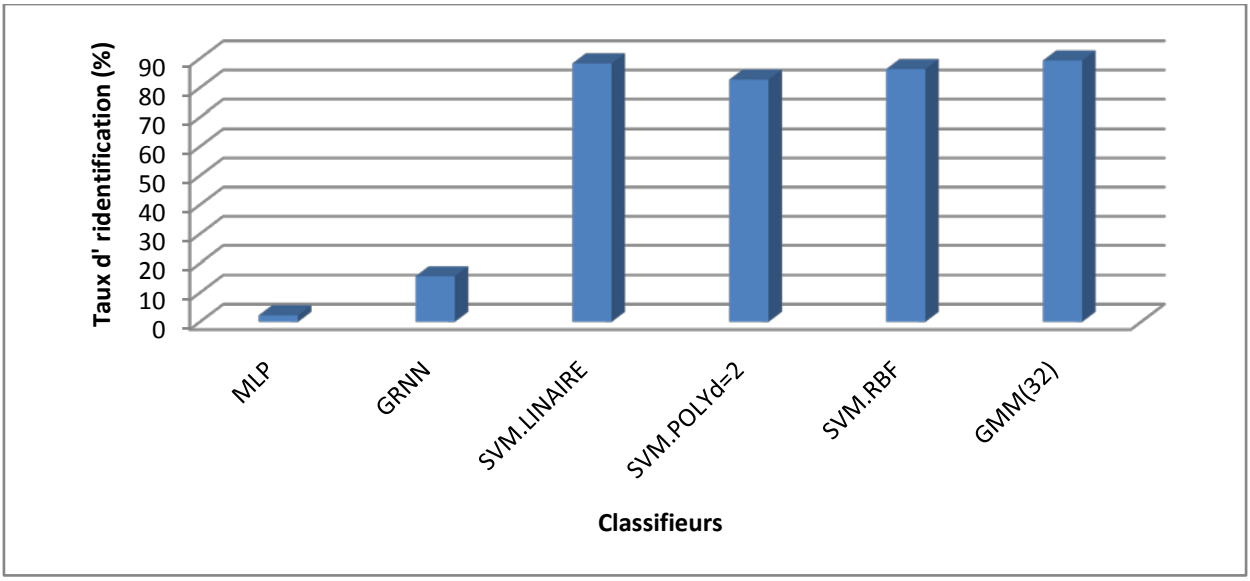


**Fig. IV.16 Histogrammes d'une comparaison des taux d'identification des différents classifieurs utilisés avec GMM dans un milieu clean.**

### IV.4.6.2 Milieu bruité par un bruit Blanc Gaussien(AW):

Tab.4.12 Comparaison des taux d'identification des différents modèles utilisés avec GMM en un milieu bruité.

| | MLP | GRNN | SVM.LINEAIR | SVM.POLY d=2 | SVM.RBF | GMM(32) |
|---|---|---|---|---|---|---|
| SNR=20 dB | 1.75% | 11.85% | 79.42% | 73.42% | 75.18% | 68.73% |

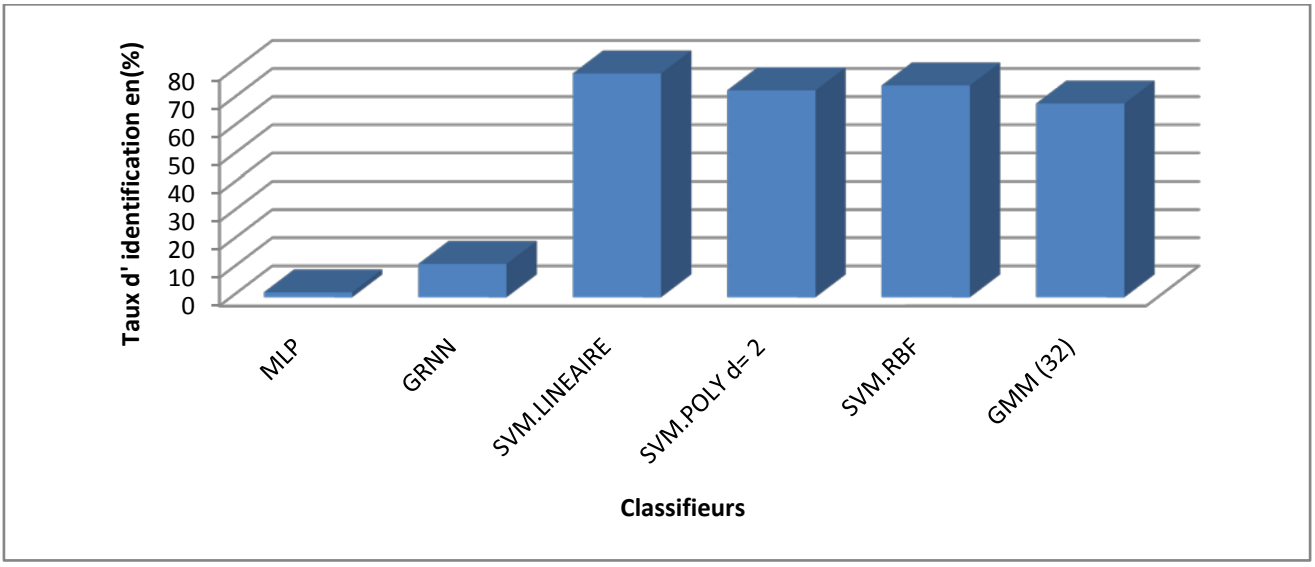


**Fig. IV.17 Histogrammes des taux d'identification des différents classifieurs utilisés avec GMM en un milieu bruité.**

### IV.5 Lecture et interprétations des résultats :

Comparons les résultats des tableaux représentant le taux de reconnaissance du locuteur obtenus suite à l'utilisation de nos trois modèles de classification à savoir MLP,GRNN et SVM avec ces trois noyaux (Linéaire, RBF et Polynomial) appliqués aux Trois différents ensembles de locuteurs (10,30 et 60) on constate ceci :

#### IV.5.1 Milieu clean :

#### IV.5.1.1 Vérification du locuteur :

On voit bien que les trois(3) modèles cités auparavant donnent de bons résultats concernant la tache de vérification. Tel que le taux de fausse acceptation ( vérification) du locuteur est autour de zéro 0%

#### IV.5.1.2 Identification du locuteur :

#### IV.5.1.2.1 En mode dépendant du texte :

Les performances de nos systèmes mises en œuvre en identification varient d'un système à l'autre et aussi d'un ensemble à l'autre. Tel que :

Le taux de reconnaissance (identification) pour le MLP est de 94,26% et le GRNN il est autour de 21%, ce qui explique que les GRNN ne sont pas adaptés pour la reconnaissance du locuteur dans notre travail.

Par contre, en SVM on constate qu'au fur et à mesure que le nombre de classes (locuteurs) augmente le taux de reconnaissance baisse par quelques pourcent, parce qu'à chaque fois qu'on ajoute une classe (locuteur) en plus, y'a un ajout des données en plus. Et çà, a une grande influence sur les paramètres de noyau utilisés, à savoir le paramètre $\boldsymbol{\sigma}$ (Pour le noyau RBF).

Autrement dit, comme le critère de décision utilisé dans notre cas (SVM multi-classes approche « **un contre tous** »), est celui de "winner-takes-all» c'est à dire Le maximum de score délivré par la fonction class(x) décrite dans **IV.5.2,** alors il s'avère que plusieurs classes ont même score et qui est en même temps le score le plus élevé, d'ou alors, la confusion dans le classement des données d'apprentissage qui engendre à son tour un mauvais classement et puis un mauvais apprentissage.

Maintenant si on fait une comparaison entre les différents noyaux, on peut dire que les deux noyaux adaptés à la reconnaissance sont le linaire et le RBF avec un taux de reconnaissance proche de 82%.

**IV.5.1.2.2 En mode indépendant du texte :**

Même interprétation qu'en mode dépendant du texte, sauf que les réseaux de neurones ne sont pas adaptés dans ce cas. Parce qu'ils ne donnent pas de bons résultats, le taux de reconnaissance égal presque à 3%(MLP) et 16%(GRNN).A propos les SVMs les noyaux les plus adaptés sont le RBF avec un taux de reconnaissance égal à 86,36% et le linéaire avec 88,33% de reconnaissance.

**IV.5.2 Milieu Bruité :**

D'après les résultats présentés dans les tableaux **Tab.4.4** ,**Tab.4.5** ,**Tab.4.6**, et **Tab.4.7**, on constate que les SVMs sont plus robustes au bruit (Exp : un taux d'identification égale à 67,5%(SVM .Poly d=2) dans un milieu bruité avec bruit d'une voiture (Volvo) à SNR= 0dB) que d'autres systèmes (MLP et GRNN) utilisés dans ce travail. Ce qui implique que les SVMs sont plus adaptés à la reconnaissance du locuteur dans un milieu bruité que les réseaux de neurones. Et aussi on remarque que le taux de'identification varie d'un bruit à l'autre et surtout d'un SNR à un autre .Tel qu'au fur et à mesure que le SNR augmente (moins de bruit), l'identification est meilleure.

### IV.5.3 Concaténation des paramètres :

Lorsqu'on a concaténé les paramètres MFCCs et LSFs (36 coefficients MFCCs + 12 LSFS), on a constaté que le taux d'identification en mode indépendant du texte est augmenté pour les SVMs que se soit dans le milieu clean (avec un taux d'identification égal à 89,73%, avec le noyau linéaire) ou dans le milieu bruité (avec un taux de reconnaissance égal à 68,87%, avec le noyau linéaire en présence de bruit babble speech avec SNR=5 dB).Ce qui prouve que la meilleure reconnaissance du locuteur est obtenue avec la concaténation des coefficients MFCCs et LSFs.

## IV.6 Conclusion :

Dans ce chapitre, nous avons décrit des expériences de reconnaissance automatique de locuteurs sans et avec concaténation des paramètres dans deux milieux différents (clean et bruité), en utilisant les techniques SVM multi-classes et les réseaux de neurones (MLP et GRNN). Nous avons montré que les SVM peuvent être très efficaces pour des tâches d'identifications de locuteurs surtout avec concaténation des paramètres.

Les résultats que nous avons obtenus et que nous avons mis sous forme de tableaux (**Tab.4.11** et **Tab.4.12** ) conforment tous que les SVM sont des techniques très intéressantes et surtout prometteuses pour des tâches de reconnaissance dans un milieu bruité mieux que les GMMs ou d'autres approches .

# Conclusion générale

La reconnaissance automatique de locuteurs (RAL) qui pourrait constitue un élément important de la biométrie humaine à travers la signature vocale est une discipline prometteuse en regard des ses nombreuse applications et perspectives. En effet, la signature vocale, au même titre que les empreintes digitales ou l'iris, peut contribuer à l'identification ou la vérification authentification d'un individu à travers la voix.

Cette technique émergeante est appelée à jouer un rôle important dans les sciences criminalistiques, en complément des autres modalités reposant sur l'image telle que les empreintes digitales ou l'iris, pour apporter la matérialisation des faits pouvant aider la justice dans sa quête de la manifestation de la vérité. Par ailleurs, l'émergence de la reconnaissance vocale dans les réseaux de communication est née du besoin d'éviter les impostures dans certains domaines sensibles. Les personnalités occupant des responsabilités stratégiques, les transactions boursières via le téléphone, l'accès sécurisé restrictif, etc., nécessitent l'authentification ou la vérification du donneur d'ordre distant.

L'objectif de notre travail était d'évaluer l'apport des méthodes à noyaux dans l'amélioration des performances des systèmes de reconnaissance automatique de locuteurs (RAL) en milieu réel, représenté souvent par un environnement acoustique fortement dégradé. En effet, la détermination des caractéristiques physiques discriminant un locuteur d'un autre est une tache très difficile, notamment en environnement adverse.

Pour cela, nous avons élaboré un système de reconnaissance automatique de locuteur, en mode dépendant et indépendant du texte, dont la partie reconnaissance repose sur des classificateur utilisant des fonctions noyaux, et qui sont alternativement : les Support Vector Machines ou SVM (avec noyau linéaire, polynomial, radial), en particulier l'approche (**SVM$^{multiclass}$** ), les réseau de neurones de type GRNN (avec noyaux de Parzen) et les mélanges de gaussiennes ou GMM..

L'application a porté sur l'utilisation de la base de données sonore arabes ARADIGITS, d'où ont été extraits les vecteurs acoustiques MFCC et LSF. Nous avons également étudié l'apport de fusion des paramètres d'entrées sur la reconnaissance de locuteurs, notamment en environnement acoustique dégradé. Ce dernier a été simulé par quatre situations : bruit de chahut dans un restaurant (speech babble), bruit de voiture (car noise Volvo), bruit d'usine (factory noise), issus de la base de données bruitée NOISEX'92, avec des niveaux SNR échelonnés de 0 à 20dB.

Une difficulté majeure pour la mise en application des SVMs à la RAL est liée au volume des bases de données nécessaires pour que les machines puissent apprendre à réaliser des tâches automatique sur le signal de parole de manière suffisamment robuste. La complexité calculatoire des algorithmes d'apprentissage et de prise de décision des SVMs peut être rédhibitoire dans le cas d'un corpus d'apprentissage trop volumineux. Ce mémoire montre deux solutions pour remédier à ce problème de façon élégante. Premièrement, adopter un bon noyau permet la synthèse de l'information de manière judicieuse. Deuxièmement, utiliser le LIBSVM et SVMligth qui incluent l'algorithme de SMO (**S**equential **M**inimal **O**ptimization) de telle façon d'accélérer la vitesse d'apprentissage autrement dit réduire le temps d'apprentissage et faire une bonne décision.

En environnement calme, les performances des différents systèmes sont presque similaires, avec un léger avantage pour les SVMs. Cependant, nous avons montré aussi à travers nos expériences que les SVMs sont plus robustes au bruit que les GMM et les réseau de neurones (GRNN et MLP). C'est ce qui explique l'intérêt grandissant pour les SVMs dans les systèmes de reconnaissances de locuteurs.

En perspective de ce travail, et pour améliorer les performances des systèmes de RAL, nous envisageons de mener des actions dans trois directions :

D'abord agir sur le choix des paramètres d'entrées pour les rendre moins sensibles aux bruits, ou mettre en place un module de rehaussement de la parole à l'entrée du système de reconnaissance.

Utiliser une approche hybride de type SVM/GMM pour allier le pouvoir discriminant des SVM au pouvoir de modélisation des GMMs.

Envisager une application dans le domaine des communications qu'elle soit filaires ou mobiles.

# Bibliographie

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, pp. 19-41, 2000 .

[2] J. Oglesby, J. Mason .Optimization of neural models for speaker identification. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp 261-264, 1990.

[3] J. Oglesby, J. S. Mason Radial basis function networks for speaker recognition. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp 393-396, Toronto (Canada), 1991.

[4] W. M. Campbell, "A covariance kernel for SVM language recognition," in *Proc. Int.Conf. Acoust. Speech Signal Process.*, 4141-4144, pp, 2008.

[5] Shakhnarovish, Darrell, et Indyk. Nearest-Neighbor Methods in Learning and Vision. MIT Press, editors 2005.

[6] M. J. F. Gales. “Maximum likelihood linear transformations for HMM-based speech recognition,” Computer Speech and Language, vol.12, no. 2, pp. 75–98, 1998.

[7] L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1559–1563, 1966.

[8] N. Minh, Do. “Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models,”IEEE Signal Processing Letters,vol. 10, no. 4, pp. 115–118, 2003.

[9] P. Kenny and P. Dumouchel. “Experiments in speaker verification using factor analysis likelihood ratios,” in Proc. Odyssey04, pp. 219–226, 2004.

[10] A. Sankar and C. Lee. “A maximum-likelihood approach to stochastic matching for robust speech recognition,” IEEE Transactions on Speech and Audio Processing, vol. 4, pp. 190–202, 1996.

[11] S. Furui .Cepstral analysis technique for automatic speaker verification. IEEE Transactions Acoustics, Speech, and Signal Processing (ASSP), volume 29(2), pp 254-272, 1981.

[12] I. Booth, M. Barlow, B. Watson .Enhancements to DTW and VQ decision algorithms for speaker recognition. Speech Communication, volume 13(3-4), pp 427-433, 1993.

[13] K. Yu, J. S. Mason, J. Oglesby .Speaker recognition using hidden Markov models, dynamic time warping and vector quantization. IEE vision, image and signal processing, Berlin, 1995.

[14] V. Vapnik. Statistical Learning Theory, Johm Wiley and Sons, New York, 1998.

[15] O. Richard E. Peter, Duda . Hart, Pattern classification and Scene Analysis. Wiley, 1973.

[16] J. Mercer. *Functions of positive and negative type and their connection with the theory of integral equations*. Philos. Trans. Roy. Soc. London, A 209:415--446, 1909.

[17] M. Aizerman, E. Braverman, and L. Rozonoer. *Theoretical foundations of the potential function method in pattern recognition learning*, Automation and Remote Control 25:821--837, 1964.

[18] E. Bernhard , M. Isabelle , Boser, Guyon, Vladimir N. Vapnik. *A Training Algorithm for Optimal Margin Classifiers [archive]* In Fifth Annual Workshop on Computational Learning Theory, pp 144--152, Pittsburgh, ACM. 1992 .

[19] V. Vapnik, "The nature of statistical learning theory", N-Y: Springer-Verlag, 1995.

[20] Najim Dehak, Reda Dehak, Patrick Kenny, and Pierre Dumouchel. “The Comparison Between Factor Analysis and GMM Support Vector Machines for Speaker Verification,” in Submited to Odyssey,2008.

[21] V. Wan and S. Renals. “SVM: Support vector machine speaker verification methodology,” in Proceedings of the International Conference on Acoustics Speech and Signal Processing, vol. 2, pp. 221–224, 2003.

[22] R.Dehak, N. Dehak, P.Kenny,and P.Dumouchel. “Linear and Non Linear Kernel GMM Super Vector Machines for Speaker Verification,” in Inter speech, Antwerp, Belgium, 2007.

[23] P.J. Moreno, P.P. Ho, and N. Vasconcelos. “A Generative Model Based Kernel for SVM Classification in Multimedia Applications,” in NIPS,2003

[24] M. Unser, A. Adroubi et Eden { « B-spline signal processing », *IEEE Transactions on Speech and Audio Processing* 41(2) , pp. 821,1993.

[25] J. Kharroubi, G. Chollet. " Utilisation de mots de passe personnalisés pour la vérification du locuteur " Jep2000, pp 361-364, 2000.

[26] J. B. Pierrot " Elaboration et validation d'approches en vérification du locuteur " Thèse de l'Ecole Nationale Supérieure des Télécommunications,1998.

[27] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America, 1990

[28] T. Kinnunen and P. Franti .Speaker Discriminative Weighting Method for VQ-based Speaker identification, 2001.

[29] A. Amrouche .“Reconnaissance automatique de la parole par les modèles connexionnistes“. Thèse de doctorat, faculté d’électronique et d’informatique, USTHB. 2007.

[30] Ahmed.Krobba, Mohmed. Deybeche .‘’ Effet de la Parole Transcodée GSM sur la Performance d’un Système de Reconnaissance Automatique du Locuteur’’, 1st International Conference on Image and Signal Processing and their Applications, Mostaganem, Algeria, octobre, 2009.

[31] J. Platt. Sequential Minimal Optimization : A fast algorithm for training support vector machines. Technecal Report MSR-TR-98-14,Microsoft Research,1998.

[32] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001.

[33] L. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3:1–8, 1972.

[34] W. M. Campbell .”A SVM/HMM system for speaker recognition,” in Proceedings of the International Conference on Acoustics Speech and Signal Processing, vol.2, pp. 209–212, 2003.

[35] L. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[36] T. Dutoit . **.**Introduction au Traitement Automatique de la Parole. Facultés poly de Mons.2000.

[37] Caliope. La parole et son traitement automatique Masson.Paris.1989.

[38] Ronan Collobert, Samy Bengio and Yoshua Bengio. “A Parallel Mixture of SVMs for Very Large Scale Problems,” Advances in Neural Information Processing Systems, Neural Computation, 2002.

[39] L.Rabiner,B.H.J. *Fundamentals of speech recognitions*. Prentice Hall Signal Processing Series 1993.

[40] L. Feng,Speaker. Recognition, Master's thesis, Technical University of Denmark, Informatics and Mathematical Modelling, 2004.

[41] S. S. Kajarekar. “Four Weightings and a Fusion: A Cepstral-SVM System for Speaker Recognition,” in Proc. IEEE Speech Recognition and Understanding Workshop, San Juan, PuertoRico, pp.17–22, 2005

[42] E. Karpov. Real-Time Speaker Identification, Master's thesis, University of Joensuu Department of Computer Science, 2003.

[43] Bahler, , Porter et Higgins. Improved voice identification using a nearest-neighbor distance measure. Dans Proc. (ICASSP) ,1994.

[44] C. M. Bishop, Neural Networks for Pattern Recognition. 1995.

[45] A. Amrouche, et al. An efficient speech recognition system in adverse condition using the nonparametric regression. Engineering Applications of Artificial Intelligence, doi:10.1016/j. pp.6-9, 2009.

[46] S. Knerr, L. Personnaz, and G. Dreyfus. Single-layer learning revisited : A stepwise procedure for building and training a neural network. In F. Fogelman-Soulié and J. Hérault, editors, Neurocomputing : Algorithms, Architectures and Applications, volume F68 of NATO ASI Series, pp 41_50. Springer-Verlag, 1990.

[47] J. Friedman. Another approach to polychotomous classification. Technical report, Department of Statistics, Stanford University, 1996.

[48] John Shawe-Taylor and Nello Cristianini, Kernel Methods for Pattern Analysis, Cambrige,2004.

[49] M. J. Carey, E. S. Parris " Speaker verification using connected words " Proceedings of Institute of Acoustics, Vol. 14, pp 95-100, 1998.

[50] Hyunchul Kim, etc. Constructing support vector machine ensemble, Pattern Recognition, vol 36, pp.2757-2767, 2003

[51] P. Ciarlet " Introduction à l'analyse numérique matricielle et à l'optimisation " Masson, nouvelle édition, 1994.

[52] ELISA Consortium ." The ELISA systems for NIST'99 evaluation in speaker detection and tracking " Digital Signal Processing Journal, Vol. 10 N 1-3, pp 143-153, 2000.

[53] Drish, Joseph, “Obtaining calibrated probability estimates from support vector machines”.

[54] Hearst, Marti A., “Trends controversies : Support vector machines”, IEEE Intelligent System, vol. 13, no. 4, pp. 18–28, 1998.

[55] A. Bellili, M. Gillouxand, P.Gallinari.“Anhybrid mlp-svm hand written digit recognizer”. In ICDAR’01 ,2001.

[56] Karacali, Bilge, Rajeev Ramanath and W.E. Snyder. “A nearest neighbor classifier based on structural risk minimization”, 2004.

.