UNIVERSITE SIDI MOHAMED BEN ABDELLAH
FACULTÉ DES SCIENCES ET TECHNIQUES FÈS
DÉPARTEMENT D'INFORMATIQUE

FST FES

# PROJET DE FIN D'ETUDES

MASTER SCIENCES ET TECHNIQUES
SYSTÈMES INTELLIGENTS & RÉSEAUX

# L'APPROCHE HYBRIDE GMM-SVM POUR L'IDENTIFICATION DES LOCUTEURS EN MODE INDÉPENDANT DU TEXTE EN MILIEU FERMÉ


**Lieu de stage : laboratoire système intelligents et réseau(LSIA)**

**Réalisé par : Ben mahria bilal**

**Soutenu le : 25/06/2015**

**Encadré par :**

Pr. Kharroubi jamal

**Devant le jury composé de :**

Pr. Kharroubi jamal
Pr. zenkouar khalid
Pr. lamrini loubna
Pr. Abounaima med chaouki

**Année Universitaire 2014-2015**

# *Dédicace*

Je dédie ce modeste travail à :

## A

**Mes très chers parents**

En témoignage de ma reconnaissance envers le soutien, les sacrifices et tous les efforts qu'ils ont fait pour mon éducation ainsi que ma formation

**Mes chers frères, et ma chère sœur**

Je vous souhaite la réussite dans votre vie, avec tout le bonheur qu'il faut pour vous combler.

**A tous les membres de ma famille, petits et grands**

**A tous mes ami(e)s**

**A tous mes enseignants au long de mes études**

**Benmahria bilal**

# *Remerciements*

Je tiens tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui m'a donné la santé et la patience d'accomplir ce modeste travail.

**A mon Encadrant le Professeur J.KHARROUBI**

**Responsable MST SIR**

Durant toute la période de mes études, j'ai été impressionné par votre droiture, votre sérieux, votre encouragement et vos conseils. Vous me faites un grand honneur en acceptant de me confier et diriger ce travail. Veuillez trouver ici le témoignage de ma gratitude avec mes remerciements pour votre bienveillance et votre disponibilité.

**A Mr Ayoub BOUZIANE**

**Doctorant au laboratoire SIA**

Pour l'aide précieuse, la disponibilité et le soutien. Je vous prie de trouver dans ce travail toute la reconnaissance que je vous témoigne.

**A tous les membres de jury**

Vous me faites le grand honneur en acceptant de juger ce modeste travail, veuillez trouver ici l'expression de mes sincères gratitudes et mon grand respect.

A tous mes enseignants, j'ai su apprécier la qualité de l'enseignement que vous m'avez transmis, vos compétences et vos rigueur scientifique sont pour moi une référence. Je vous prie de trouver ici l'expression de mes vifs remerciements.

A tous ceux qui ont participé de près ou de loin pour la réalisation de ce travail.

A ce qui croit le changement.

# *Résumé*

La reconnaissance automatique du locuteur est une technique consiste à reconnaitre l'identité d'une personne à l'aide de sa voix. Les différents Tâches en RAL sont : l'identification et la vérification. La première tâche envisagée en reconnaissance du locuteur est l'identification Automatique du locuteur (IAL). Il s'agit de retrouver l'identité d'un locuteur parmi un panel de personnes. Dans le cadre d'une évaluation perceptive, il est possible que l'auditeur connaisse déjà le locuteur à identifier (personne connue) ou au contraire que l'auditeur ait dans un premier temps à apprendre la voix du locuteur (personne nouvelle).L'identification peut être ouverte, dans le sens où il est possible que le locuteur ne soit pas présent dans le panel, ou bien fermée, lorsque le locuteur est obligatoirement dans le panel. La vérification automatique du locuteur (VAL) consiste à vérifier l'identité proclamée par un individu par la comparaison d'un signal vocal et d'un modèle de référence du locuteur présumé, préalablement appris par le système.

Une des questions proposées dans la reconnaissance automatique du locuteur est comment représenter un énoncé, qui a un nombre variable de vecteurs caractéristiques en un seul vecteur .Dans ce rapport, nous présentons un système hybride GMM-SVM, qui combine entre l'utilisation de GMM pour la modélisation et pour générer des vecteurs de taille fixe (supervecteur) qui sont utilisés comme des entrés pour le SVM lors de la phase de la décision, il existe une autre méthode pour générer des supervecteur : l'utilisation de noyau de séquence. Dans la section des résultats, il est évident l'influence de nombre de gaussien G sur le taux de reconnaissance. Le meilleur résultat est obtenu par le noyau RBF pour G=1024 avec G est le nombre de gaussiens.

**Mot-clé :** RAL, VAL, IAL, SVM, GMM, Noyau de séquence, supervecteur.

# *Abstract*

Automatic speaker recognition is a technique consists in recognizing the identity of a person using his voice. The different tasks in RAL: identification and verification. The first task envisaged in speaker recognition is the identification Automatic speaker (IAL). This is to find the identity of a speaker from a panel of people. As part of a perceptual evaluation, it is possible that the listener already knows the speaker to be identified (known person) or on the contrary that the listener has at first to learn the speaker's voice (new person) .Identification may be open, in the sense that it is possible that the speaker is not present in the panel, or closed, when the speaker is mandatory in the panel. Automatic speaker verification (VAL) is to verify the identity proclaimed by an individual by comparing a voice signal and a reference model of the alleged speaker, previously learned by the system.

One of the questions proposed in the speaker recognition is how to represent a statement that has a variable number of feature vectors into a single vector .In this report we present a hybrid system GMM-SVM, which combines between the use of GMM for modeling and generating fixed size vectors (supervector) which are used as input for the SVM in phase of the decision, there is another method to generate supervector: using sequence core. In the results section, it is clear the number of Gaussian G influence on the recognition rate. The best result is obtained by the RBF kernel with G = 1024 G is the number of Gaussian.

**Keywords** : RAL, VAL, IAL, SVM, GMM, séquence kernel, supervector.

# Liste des abréviations

| | |
|---|---|
| DTW | Dynamic Time Warping |
| GMM | Gaussian Mixture Models |
| HMM | Hidden Markov Models |
| IAL | Identification Automatique du Locuteur |
| LLR | Log Likelihood Ratio |
| MAP | Maximum A Posteriori |
| MFCC | Mel Frequency Cepstral Coefficients. |
| MLLR | rsMaximum Likelihood Linear Regression |
| RAL | Reconnaissance Automatique du Locuteur |
| RBF | Radial Basis function |
| SVM | Support Vector Machine |
| TEE | Taux d'Égale Erreur |
| TFA | Taux de Fausses Acceptations |
| TFR | Taux de Faux Rejets |
| TIC | Taux d'Identification Correcte |
| UBM | Universal Background Model |
| VAL | Vérification Automatique du Locuteur |
| VQ | Vector Quantization |

# Table des matières

# Liste de Figures

# Liste des tableaux

# Introduction Générale

La Reconnaissance Automatique du Locuteur (RAL) est une des disciplines du traitement du signal de la parole. Elle consiste à reconnaître l'identité d'un individu à partir de sa voix. L'identification automatique du locuteur(IAL) et la vérification automatique du locuteur(VAL) sont les deux tâches les plus répandues dans le domaine de RAL.

Une des questions proposées dans la reconnaissance automatique du locuteur est comment représenter un énoncé, qui a un nombre variable de vecteurs caractéristiques en un seul vecteur. En générale le modèle le plus utilisé pour modéliser un locuteur est basé sur l'apprentissage statistique.

L'apprentissage statistique est un domaine à la frontière de l'informatique et de statistique. Il consiste à développer des algorithmes qui permettent aux ordinateurs d'apprendre grâce à l'expérience. Pour cela, les algorithmes ont besoin des exemples d'apprentissage. Le but est alors de trouver la meilleure fonction parmi un ensemble préétablie de fonctions, en minimisant une fonction de coût sur les exemples d'apprentissage. L'ensemble de fonction choisi au préalable doit être suffisamment riche pour contenir une bonne solution, mais suffisamment simple pour que la solution choisie puisse être généralisée à des exemples jamais vus par le système. La solution trouvée par un algorithme d'apprentissage statistique est appelée modèle. Mais Un problème habituelle en apprentissage statistique est de classer des exemples en deux catégories, c'est ce qu'on appelle le problème de classification supervisée à deux classes.

Les modèles utilisés pour résoudre ces problèmes sont soit discriminants (il cherche un hyperplan qui sépare le mieux les deux classes) soit génératifs (il estime indépendamment la distribution de chacun de deux classes et utilise la règle de Bayes pour prendre une décision). L'approche statistique consiste à représenter chaque locuteur par une densité de probabilité dans l'espace des paramètres acoustique, comme par exemple GMM et HMM.

Le modèle le plus utilisé en identification du locuteur est basé sur des mélanges des distributions gaussiennes GMM, cette méthode est proposée par Rynolds et Rose en 1995.l'idée de cette méthode consiste à concevoir un modèle GMM universel appelé modèle du monde ou encore Universel Background Model(UBM). Ce modèle est entrainé sur des enregistrements d'un grand nombre de locuteurs. Plus la diversité entre locuteurs est grande, meilleur sera le modèle. Par la suite, le modèle de chaque client est construit par adaptation du modèle du monde(UBM) en utilisant que les enregistrements de clients. Contrairement au modèle du monde, le nombre d'exemples d'entrainement disponible pour estimer ce modèle est restreint : le client prononce généralement entre une ou trois phrases avant d'utiliser le système.

Donc comme peu de données sont disponibles, plutôt que d'apprendre un nouveau GMM avec les données du client, les paramètres du monde sont adaptés avec ces données. Cette méthode est appelé Maximum A Posteriori(MAP) comporte un paramètre à ajuster qui permet de contraindre le modèle client à rester plus ou moins proche du modèle du monde.

Habituellement seules les moyennes de gaussiennes sont modifiées. Finalement lors de la prise de décision, chaque hypothèse est testée en calculant un score, appelé vraisemblance, pour chacun des modèles. Le ratio de cette vraisemblance est comparé avec un seuil de décision appris au préalable sur un autre en ensemble de client.

La plupart des études publiées dans le domaine de la RAL utilisent le log du rapport de vraisemblance comme fonction de décision. Mais dans ce projet, nous avons voulu étudier l'utilisation de la méthode discriminante (SVM : Support Vector Machine). Mais la longueur variables des enregistrements vocaux utilisés dans la RAL rendent l'utilisation des modèles discriminants plus difficile.

La communauté de reconnaissance du locuteur a proposé un moyen robuste pour présenter des énoncés de tailles variables par un seul vecteur, appelé super vecteur. Ces supers vecteurs peuvent être utilisés comme entrées pour les méthodes discriminantes de décision entre autre les vecteurs supports machines de vecteur (SVM). Pour créer des super-vecteurs, il existe deux méthodes :

1- Soit en utilisant le noyau de séquence (chapitre 4, section 4).
2- Soit en utilisant le modèle du mélange de gaussienne (GMM : chapitre 2)

L'utilisation de SVM avec un noyau de séquence permet de remplacer un énoncé représenté par un nombre variable de vecteur caractéristique par un autre vecteur de taille fixe dans un autre espace de caractéristique. Parmi ces noyaux, nous trouvons des noyaux classique comme : linéaire, polynomiale et Gaussien et des autres plus récents comme : Fisher et GLDS.

Dans ce projet, nous avons utilisé le modèle GMM pour générer ces supers vecteurs qui sont utilisées comme des entrées pour le classifieur SVM. Notre système hybride GMM-SVM permet de faire une combinaison entre le GMM qui est utilisé pour la modélisation et la génération des supers vecteurs et le classifieur SVM qui est utilisé pour la décision.

Ce rapport est organisé comme suit :

Le premier chapitre est consacré à introduire le système Automatique du Locuteur.

Le Deuxième chapitre consiste à présenter le système GMM-UBM qui représente le modèle classique pour la modélisation du locuteur.

Dans le troisième chapitre, je vais donner une vision globale sur le principe des SVMs,ainsi que les différents types de noyau.

Le quatrième chapitre est consacré à représenter un historique sur l'utilisation des SVMs pour l'identification du locuteur, notre approche utilisée pour l'identification du locuteur et ensuite les résultats obtenus et les évaluations.

# Chapitre I

Système de reconnaissance
Automatique de locuteur

# I. Introduction aux systèmes de reconnaissance Automatique de locuteur

L'objectif de la reconnaissance du locuteur est de reconnaître l'identité d'une personne à l'aide de sa voix. Les applications de la RAL sont principalement liées aux problèmes d'authentification ou de confidentialité [1].

La reconnaissance du locuteur est un terme générique qui répond à plusieurs définitions selon le scénario applicatif envisagé. Les scénarios applicatifs sont regroupés en trois catégories principales[1], la figure1 montre les différents tâches en RAL :

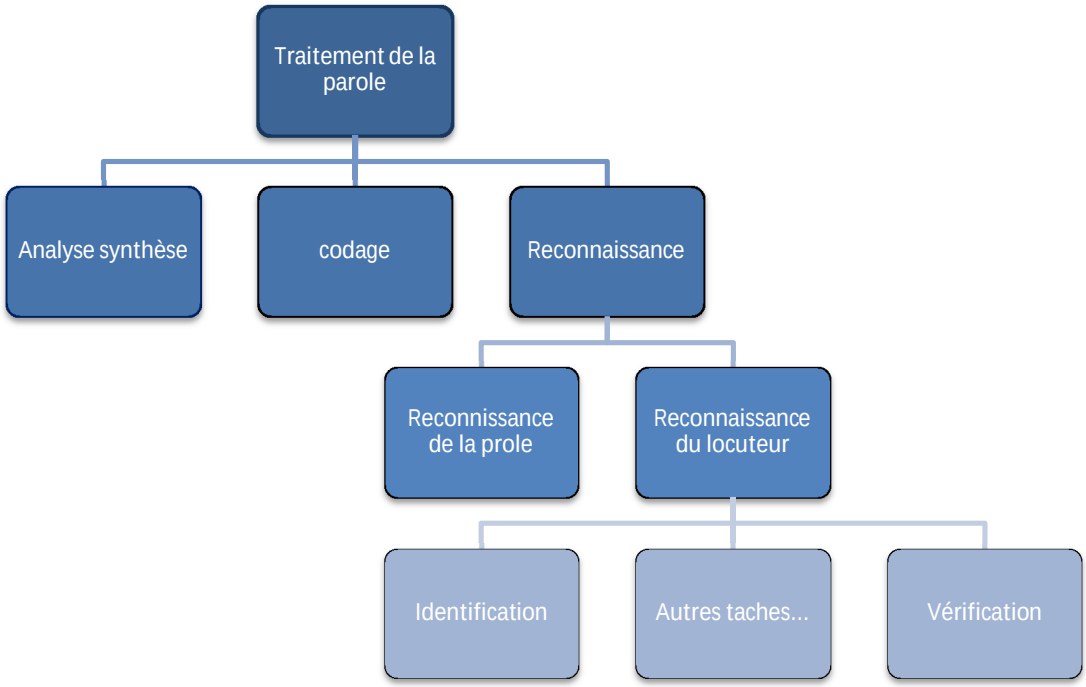



**Figure 1: Les différents tâches en RAL**

- L'identification du locuteur.
- La vérification du locuteur.
- L'indexation du locuteur ou le suivi du locuteur.

Chacune de ces catégories propose son protocole de reconnaissance selon que l'identité du locuteur à reconnaître soit proclamée, ou que les locuteurs à reconnaître soient connus ou non du système de RAL. Le système de RAL peut valider une identité pour la vérification du locuteur, proposer une identité à partir d'un ensemble de locuteurs, déterminer les durées de parole d'un locuteur, compter le nombre de locuteurs présents dans un signal[1].

Une seconde classification à l'intérieur de ces catégories repose sur le niveau de dépendance au texte. La reconnaissance peut être indépendante du texte ou dépendante du texte. En mode dépendant du texte la reconnaissance bénéficie de la connaissance du contenu linguistique prononcé (fixe ou prompté). L'estimation des paramètres caractéristiques du locuteur est alors plus robuste. En mode indépendant du texte, le système de reconnaissance n'a aucune connaissance sur le message linguistique prononcé par la personne[1].

# 1. Les différentes tâches en RAL

## 1.1. Identification Automatique du Locuteur

L'Identification Automatique du Locuteur (IAL) consiste à déterminer, à partir d'un ensemble de locuteurs référencés dans le système, l'identité du locuteur présent dans un signal vocal (signal de test). Pour cela, le système calcule des mesures de similarités entre ce signal et tous les modèles des locuteurs de la base. Deux conditions d'identification sont connues : milieu fermé et milieu ouvert[2].

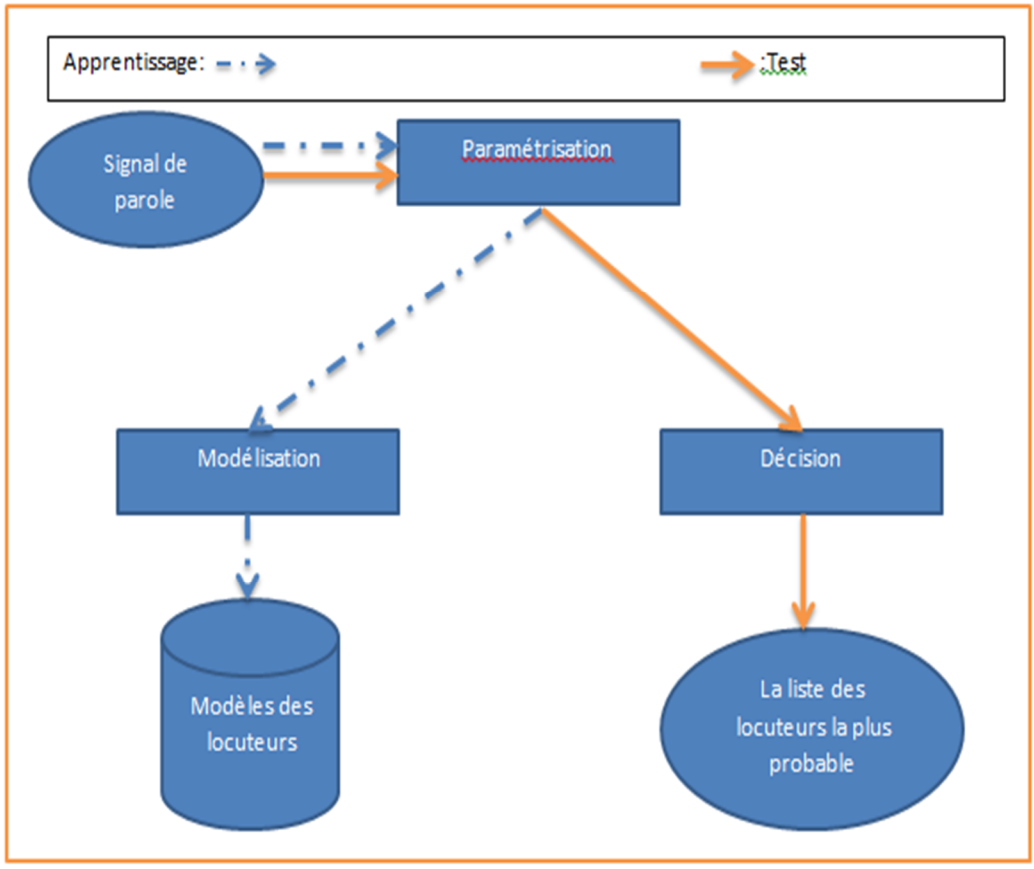



**Figure 2: Structure d'un système d'identification du locuteur**

Dans le cas où le système doit fournir un ensemble d'au moins un locuteur, on parle d'une identification en milieu fermé. Mais dans certaines applications, le système peut être amené à fournir un ensemble vide : c'est l'identification en milieu ouvert [figure 1]. En milieu fermé [figure 2], chaque accès de test est comparé à tous les modèles des locuteurs référencés dans le système. L'identité du locuteur possédant la référence la plus proche est émise en sortie du système[2].

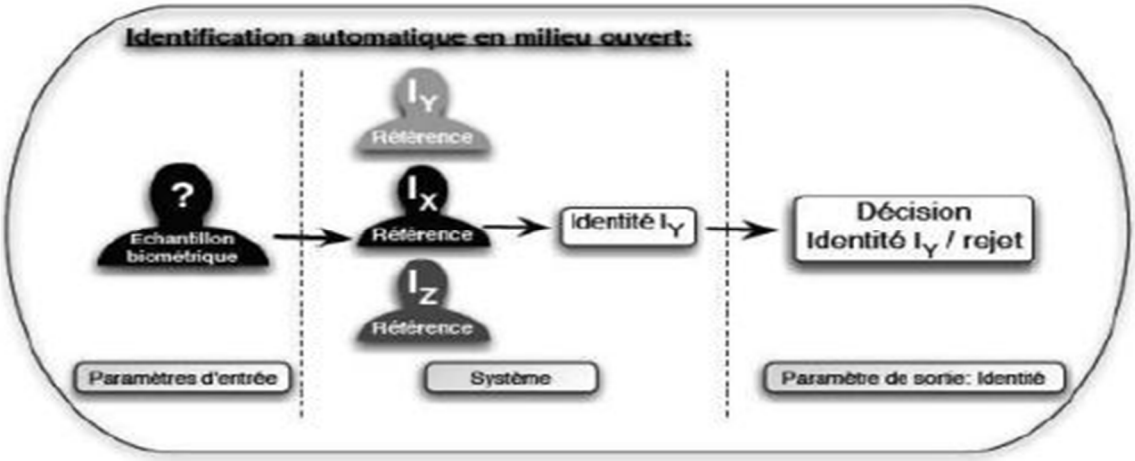



**Figure 3 : Identification automatique de locuteur en un groupe ouvert**

En identification en milieu ouvert, le système répond à deux interrogations: « *Quelle est l'identité la plus probable ?*» et « *Les données biométriques analysées correspondent-elles à cette identité ?*»Alors qu'en milieu fermé il ne répond qu'à la première question.

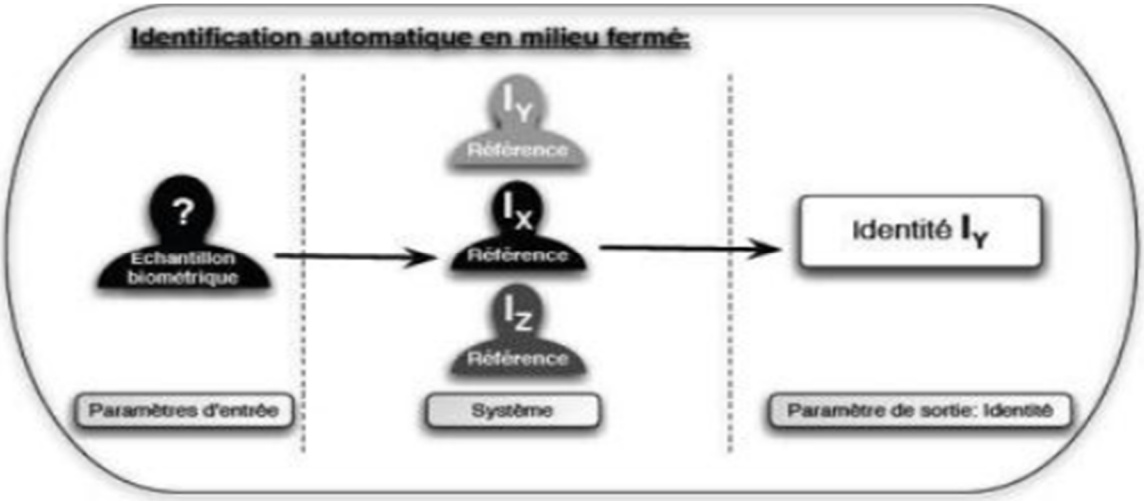



**Figure 4 : Identification automatique de locuteur en un groupe fermé**

## 1.2. Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (VAL) consiste à vérifier l'identité proclamée par un individu par la comparaison d'un signal vocal et d'un modèle de référence du locuteur présumé, préalablement appris par le système. Un système de VAL a donc deux entrées : une identité et un accès de test. Le résultat de cette comparaison est considéré comme une mesure de similarité avant d'être comparé à un seuil d'acceptation. Lorsque la mesure de similarité est supérieure à ce seuil, l'individu est accepté, sinon il sera rejeté (figure3) [2].

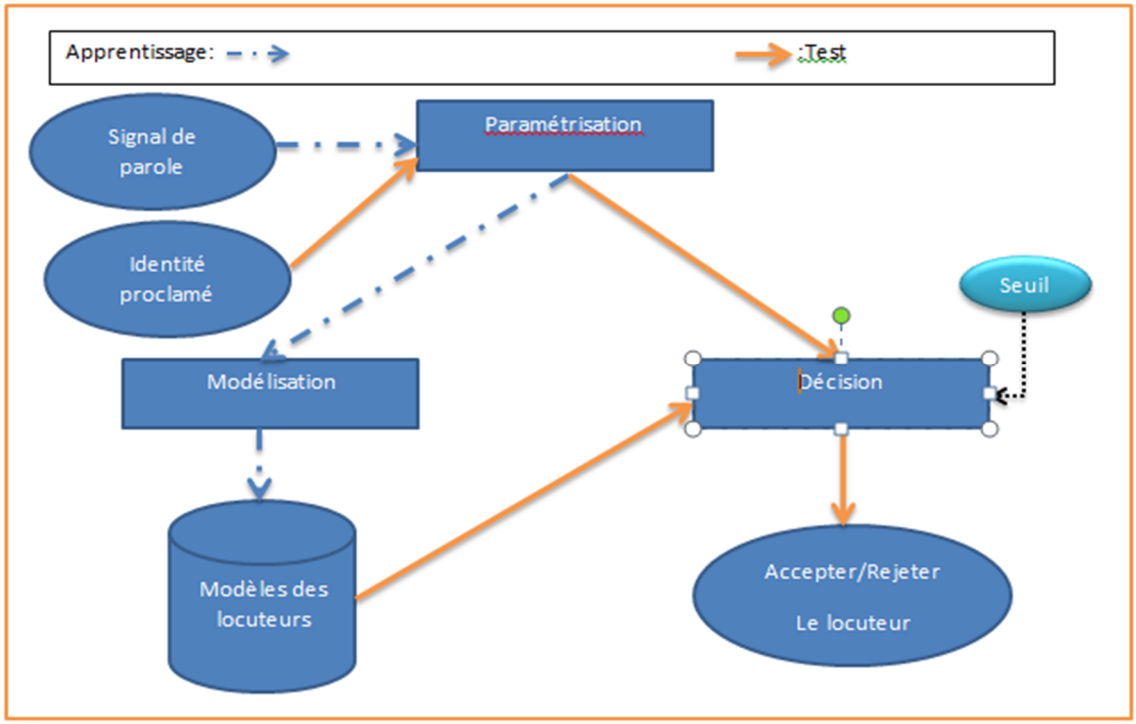



Figure 5: Structure d'un système de vérification du locuteur

### 1.3. Indexation automatique en locuteur

L'indexation en locuteur permet de déterminer les temps de parole des individus dans un signal audio.

La spécificité de cette tâche réside dans le fait que le système ne détienne pas de référence pour les locuteurs présents dans le signal audio. Un mécanisme d'apprentissage aveugle et adaptatif est alors mis en place. Il est possible de segmenter un signal audio par prise de parole des intervenants, étiqueter des données audio pour permettre des recherches de documents audio par locuteurs ou, enfin, identifier le nombre de locuteurs présents dans le signal.

Le suivi de locuteur est similaire à l'indexation en locuteur, à ceci près que les locuteurs présents dans le signal sont connus par le système de RAL. Il s'agit donc d'une simplification de la tâche d'indexation en locuteur mais qui reste néanmoins, une tâche très complexe.

## 2. Dépendance et Indépendance du Texte

En mode dépendant du texte, le texte prononcé par le locuteur est le même que celui qu'il a prononcé lors de l'apprentissage de sa voix. Les niveaux de dépendance au texte sont classés suivant les applications : systèmes à texte libre (*free-texte*), systèmes à texte suggérée (*text-prompted*), systèmes dépendants du vocabulaire (*vocabulary-dependant*) ou système personnalisés dépendants du texte (user specific text dependent). D'évidence, la connaissance a priori du message vocal rend la tâche des systèmes de RAL plus facile et les performances plus meilleures[2].

En mode indépendant du texte, le locuteur peut prononcer n'importe quelle phrase pour être reconnu. Dans ce mode, il n'existe aucune contrainte sur le message que le locuteur doit prononcer ni sur la langue qu'il peut utiliser[2].

## 3. La variabilité du signal de la parole

La variabilité **intra-locuteur** concerne les changements de la voix du même locuteur et qui sont dus, en général, à la fatigue, le stress, le sommeil, l'horaire de la journée (matin ou soir), le débit de l'élocution, l'état émotionnel,...

Le signal de parole varie selon le locuteur, on parle de la variabilité **interlocuteur** lorsque les caractéristiques qui sont propres à chaque locuteur ne sont pas les mêmes chez d'autre locuteur.

La variabilité **intersession** (entre sessions d'enregistrements) fait apparaître l'influence de facteurs extérieurs sur le signal de parole. A la sortie du conduit vocal humain, l'onde de parole est considérée comme idéale, car aucune déformation/distorsion de l'environnement extérieur ne l'a modifiée.

L'environnement sonore lors de l'enregistrement, le matériel d'acquisition ou le canal de transmission utilisé vont ensuite déformer l'onde sonore originelle. Le canal de transmission, par exemple, agit comme un filtre en fréquence sur l'onde sonore.

Ces facteurs rendent complexe la comparaison entre plusieurs échantillons d'un même individu. De nombreux travaux expérimentaux ont montré que des variations de matériel entre les phases d'apprentissage et de test sont à l'origine de graves dégradations des performances[3]. Par exemple, l'acquisition d'un signal de parole sur le réseau GSM introduit les dégradations suivantes sur le signal de parole[1]:

- l'ajout du bruit de l'environnement,
- le sous-échantillonnage à 8kHz du signal.
- le filtrage sur la bande de fréquence [300-3400] Hz.
- le codage à bas débit de la parole.
- l'ajout du bruit de quantification des paramètres émis.
- la transmission sur un lien sans-fil avec pertes.

# II. Système de reconnaissance automatique du locuteur

La tâche de reconnaissance automatique du locuteur est composée de trois principales phases. Nous pouvons distinguer la phase de paramétrisation (analyse acoustique), la phase de modélisation et la phase de décision. En plus, un système de RAL possède deux modes[4] :

- Apprentissage, où un modèle est estimé pour chaque locuteur « client » du système puis servira de référence pour les tâches de reconnaissance à venir.

- Test, où l'étape de reconnaissance (vérification, identification…) est effectuée. En sortie de cette phase, le système émet une réponse : une identité pour la tâche d'identification ou une décision accès/rejet pour la vérification.

# 1. Paramétrisation du signal de parole

Divers paramètres du signal de parole ont été proposé en reconnaissance automatique du locuteur. Idéalement, ces paramètres doivent avoir une forte variabilité inter-locuteur et une faible variabilité intra-locuteur, permettant ainsi de discriminer plus facilement différents individus. De plus, ces paramètres doivent être robustes aux différents bruits et variation inter-session, et difficile à reproduire par un imposteur[5].

La littérature aborde de nombreux types de paramétrisation du signal de la parole[6], [7][8], dont les paramètres doivent être fréquents, facilement mesurables, pas trop sensibles à la variabilité intra-locuteur, robustes face aux imitateurs, etc. Il ressort que les seuls types de paramètres vraiment pertinents et utilisables efficacement sont les paramètres de l'analyse spectrale, les paramètres dynamiques et éventuellement les paramètres prosodiques. Combiner ces différentes sources d'information peut conduire à des performances supérieures à celles que l'on obtient en les utilisant séparément [1, 9].

## 1.1. Paramètres de l'analyse spectrale

La représentation du signal par les coefficients Mel Cepstre (MFCC) est très utilisée dans les tâches de reconnaissance du locuteur. Ces coefficients caractérisent bien la forme du spectre et permettent de séparer l'influence de la source du signal vocal de celle du conduit vocal. Cette séparation est rendue possible grâce à un filtre déconvolutif.

Le codage MFCC (Mel Frequency Cepstral Coding) est une technique très utilisée en traitement de la parole.

Il est basé sur la variation des bandes critiques de l'oreille humaine avec la fréquence, les filtres espacés linéairement aux basses fréquences et logarithmiquement à hautes fréquences [5]. Ces filtres sont modélisés par une échelle non-linéaire issue de connaissances sur la perception humaine : l'échelle Mel[10].

Pour les MFCCs, on utilise la fenêtre de Hamming durant la transformation du domaine temporel au domaine fréquentiel. Cette transformation est faite en utilisant la transformée de Fourier.

Un filtrage, est appliqué ensuite, par banc de filtres triangulaires espacés selon l'échelle de Mel. Cette échelle reproduit la sélectivité de l'oreille qui diminue avec l'accroissement de la fréquence.

Après le calcul de log, une transformée en cosinus discrète est appliquée pour assurer un retour au domaine temporel. Figure suivante illustre l'algorithme de calcul des coefficients MFCC[10].

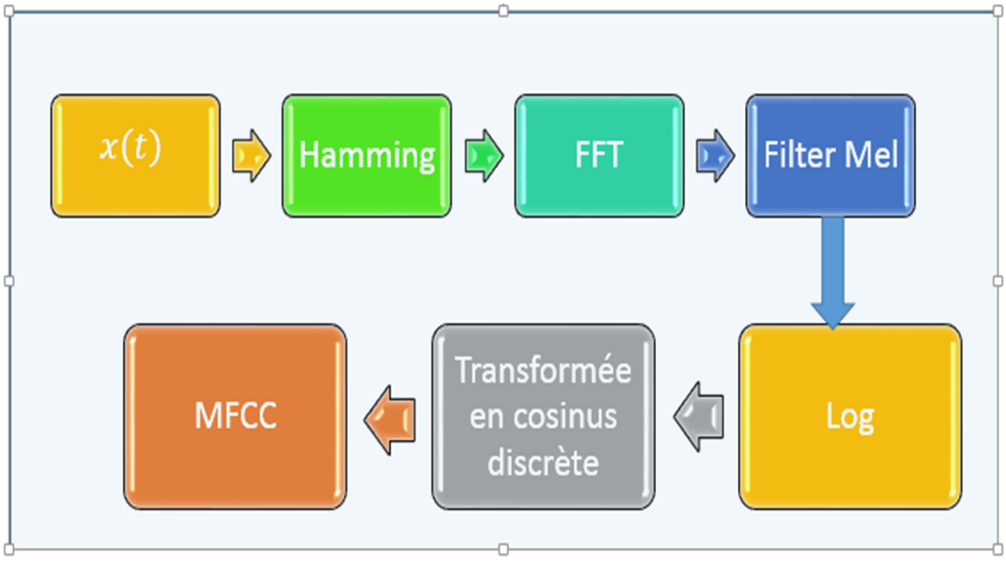



Figure 6: Algorithme de calcul des coefficients MFCC

### 1.2. Paramètres dynamique

La prise en compte d'une information de type dynamique est un facteur d'amélioration des performances d'identification du locuteur. Les dérivées du premier et du second ordre, appelées aussi coefficients de delta et delta-delta, permettent d'introduire une information concernant le contexte temporel d'une trame courante[11]. D'après Harb[12] l'utilisation de ces paramètres reste l'approche la plus populaire actuellement en raison de la simplicité pour sa mise en œuvre et l'amélioration des performances que nous pouvons observer[4].

### 1.3. Paramètres prosodiques

Ces paramètres caractérisent en grande partie le style d'élocution d'un locuteur. L'énergie contient l'information liée au niveau acoustique moyen du signal. Ces paramètres s'avèrent fragiles en pratique et ne permettent pas, à eux seuls, de discriminer de manière fiable les locuteurs.

En conséquent, ils sont souvent associés aux paramètres de l'analyse spectrale. Le terme "paramètres prosodiques" réunit l'énergie, la durée et la fréquence fondamentale (ou pitch)[4, 13].

## 2. Modélisation du locuteur

Ce paragraphe parcourt brièvement les techniques les plus couramment utilisées en reconnaissance du locuteur. Comme dans le cas de la reconnaissance de la parole, le problème de la reconnaissance de locuteur peut se reformuler selon un problème de classification[14].

L'étape de modélisation exploite les données fournis dans l'étape de la paramétrisation afin de créer la représentation d'un individu qui servira, par la suite, à l'authentifier. Le modèle utilisé est généralement une représentation statistique des données acquises.

Différentes approches ont été développées, néanmoins on peut les classer en quatre grande familles :

## 2.1. Approche Vectorielle

Le locuteur est représenté par un ensemble de vecteurs de paramètres dans l'espace acoustique .Ses principales techniques sont la reconnaissance à base de DTW et par Quantification vectorielle.

### *2.1.1. Reconnaissance du locuteur à base de DTW*

La reconnaissance par DTW(Dynamique Time Warping) repose sur le principe que chaque mot est représenté par une prononciation du référence(template).Compte tenu des décalages temporels entre les différentes prononciation d'un même mot, l'algorithme met en correspondance des séquences de paramètres par distorsion temporelle(Time Wraping).

La programmation dynamique permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage. Dans ce cas, le modèle de locuteur est tout simplement l'ensemble des vecteurs de paramètres obtenus après paramétrisation des signaux d'apprentissage. Une distance est calculée entre vecteurs d'apprentissage et de test et moyennée sur l'ensemble de la séquence.la programmation dynamique est une technique exclusivement utilisée en mode dépendant du texte. Cette approche est facile et mettre en œuvre est donnes des performances bonne .[6][14, 15].

### *2.1.2. Quantification Vectorielle(QV)*

Le principe consiste à définir un nombre limite (de quelques dizaines à quelques centaines) de vecteurs {prototype} formant un dictionnaire (codebook) obtenu dans une phase préalable, par exemple à l'aide de méthodes issues de classification automatique de données.

Chaque vecteur de paramètres est alors simplement représenté par le code du vecteur prototype qui lui est le plus proche[16].La quantification vectorielle (Vector Quantization : VQ) repose sur un partitionnement de l'espace acoustique en sous-espaces. Chaque sous-espace est associé à leur vecteur centroïde (i.e. à un vecteur de paramètres représentant l'ensemble des vecteurs composant le sous-espace). Dans ces conditions, un modèle de locuteur est composé d'un ensemble de vecteurs centroïdes, appelé dictionnaire de quantification (codebook).

Lors de la phase de reconnaissance, une distance est calculée entre un vecteur de test et chaque vecteur centroïde du dictionnaire. La distance minimale est assignée au vecteur de test. La distance d'une séquence de vecteurs de test est obtenue par moyenne des distances minimales attribuées à chacun des vecteurs de test.

La quantification vectorielle s'applique en mode dépendant ou indépendant du texte. La rapidité et les performances de cette technique dépendent fortement de la taille du dictionnaire: plus la taille du dictionnaire augmente, meilleures sont les performances sinon, le processus devient plus lent.

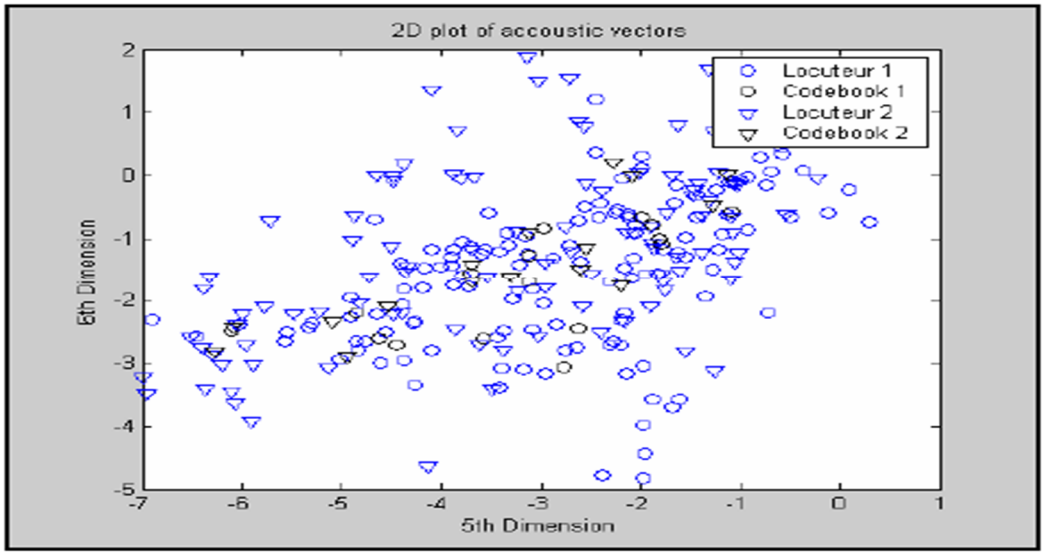



**Figure 7: Exemple de diagramme qui illustre la formation du codebook par QV**

La figure 7 illustre le processus de la reconnaissance. Les cercles font référence aux vecteurs acoustiques de locuteur 1 alors que les triangles sont du locuteur 2. Le codewords du résultat (centroids) est montré respectivement dans la figure 5 par des cercles noirs et des triangles noirs pour les locuteurs 1 et 2. La distance d'un vecteur au codeword le plus proche formé le codebook est appelée une « VQ-distorsion ».

De nombreuse articles proposent l'utilisation de la quantification vectorielle en la reconnaissance de locuteur, nous citons : [17, 18], et on peut trouver une bonne description de cette méthode dans [19].

## 2.2. L'approche statistique

Consiste à représenter chaque locuteur par une densité de probabilité dans l'espace de paramètres acoustique. Elle couvre les techniques de modélisation par les modèles de Markov caché, les mélanges de gaussiennes [14].

### *2.2.1. Modèle de Markov caché*

Les modèles de Markov (ou HMM pour Hoden Markov Models) ont été initialement introduits en la reconnaissance de la parole. Puis leur utilisation s'est étendue peu à peu au domaine de la reconnaissance de locuteur.

Dans cette approche, il ne s'agit plus de d'une mesure de distance d'une forme acoustique à une référence, mais la probabilité que la forme acoustique ait été engendrés par le modèle de référence de locuteur. Le modèle d'un locuteur est constitué de l'association d'une chaine de Markov, une succession d'états avec des probabilités de transitions d'un état à l'autre, et de lois de probabilité (probabilité d'observations d'un vecteur acoustique dans un état). Quant à l'utilisation des modèles de Markov cachés en reconnaissance du locuteur, on peut se référer [14, 20-22].

### *2.2.2. Les Mélanges de Gaussiennes*

La reconnaissance du locuteur par mélanges de gaussiennes(ou GMM pour Gaussian Mixture Models) consiste à modéliser un locuteur par une somme pondérée de composante gaussiennes[7]. Ainsi une large gamme de distributions peut être parfaitement représentée Chaque composante des gaussiennes est supposé modéliser un ensemble de classe acoustique.il semble bien modéliser les caractéristique spectrales des voix de locuteurs, et il est relativement simple à mettre en œuvre. Les mélanges de gaussiennes est considérés comme un cas particulier des HMM et une extension de quantification vectorielle. Nous allons aborder cette méthode avec plus de détails dans le chapitre 2 [14]

## 2.3. L'approche connexionniste

Cette approche consiste à modéliser les locuteurs par des réseaux de neurones .Les réseaux de neurones ont été largement utilisés en reconnaissance du locuteur. Ils offrent en effet une bonne alternative au problème de la discrimination entre les locuteurs .ces outils de classification permettent de séparer des classes, dans un espace de représentation donnée, de façon non linéaire.

Pour une bonne description de cette technique, on peut lire [23-25].L'inconvénients important de l'application de cette technique en identification du locuteur est le coût important lié à l'ajout d'un nouveau locuteur dans la base de référence (ce n'est pas le cas de la vérification du locuteur).On peut aussi utiliser les réseaux de neurones en les couplant à d'autres technique, comme par exemple le modèle de markov caché . On parle alors de méthode hybrides.

## 2.4. L'approche relative

Cette nouvelle technique consiste à modéliser un locuteur non plus de façon absolue mais relativement à un ensemble de locuteur bien appris. Cette nouvelle approche a donné la naissance à la notion d'espace locuteur où modèle de locuteur est représenté généralement par une combinaison linéaire des modèles de référence.

Cette nouvelle technique consiste à modéliser un locuteur non plus de façon absolue mais relativement à un ensemble de locuteur bien appris. Les techniques d'adaptation globale, telle que MAP(Maximun A Posteriori) ou MLLR (Maximun likelhood Linear Regression), permettent d'obtenir un modèle qui a des performances similaires au modèle obtenu avec un apprentissage classique tout en nécessitant une quantité de donnée moins grande que celle utilisée pour un apprentissage classique. Cependant, cette quantité de donnée reste assez importante.

Le principe de la représentation relative des locuteurs a été initialement appliqué en reconnaissance de la parole dans des techniques d'adaptation rapide.

Ces techniques reposent sue le principe d'utiliser des connaissances à priori obtenues à partir d'en ensemble de locuteur de référence. Les principales techniques sont : RMP(Regression-Based model Prediction) ,Speaker Clustring, RSW (Reference Speaker Weithing) et les voix propres (eigenvoices) [14].

### 2.4.1. Clustering des locuteurs

Le clustering repose sur le principe de création plusieurs clusters de référence représentant l'ensemble des locuteurs .Ces clusters peuvent être déterminés avec des mesures de similarité sur les coefficients acoustique des locuteurs ou sur leur modèles. Chaque Cluster contient des locuteurs similaires[14].

### 2.4.2. Les voix propres

Cette approche[26]a été développée initialement dans le cadre de l'adaptation en locuteur. Elle s'inspire largement des concepts des eigenfaces. Les voix propres sont générées par des algorithmes de réduction de dimensionnalité. A partir de la matrice des paramètres HMM du locuteur, on applique ces algorithmes et on ne conserve que les axes à grande inertie. Les voix propres sont aussi calculés par une méthode itératives appelées MLLES (paragraphe 6.2) [14]qui consiste à rechercher un espace optimale en maximisant la vraisemblances de données. Les locuteurs sont localisés par maximum de vraisemblance. En identification du locuteur, on applique la distance euclidienne ou l'angle entre les vecteurs de coordonnées. [14].

# III. Décision et Mesure de performance

Un module de reconnaissance fournit en sortie un score. La nature de ce score varie selon les modules de reconnaissance utilisés. Il s'agit la plupart du temps d'une distance, d'une probabilité ou d'une vraisemblance. Un module de décision doit, à partir de ce score, fournir une décision qui constituera la réponse finale du système de reconnaissance.

Le module de décision décrit dans le paragraphe précédent reçoit, en entrée et pour chaque test, un score. Celui-ci résulte de la comparaison entre les caractéristiques biométriques de l'utilisateur testé et la référence apprise lors de la phase d'enrôlement. Un score élevé signifiera que la probabilité pour que l'utilisateur testé corresponde à l'identité qu'il annonce est élevée et un score faible signifiera que cette probabilité est faible. La décision binaire qui constitue la sortie du module résulte de la comparaison de ce score avec un seuil défini à l'avance. Si le score est supérieur au seuil, l'utilisateur est accepté et s'il est inférieur au seuil, l'utilisateur est rejeté.

## 1. Vérification du locuteur

Un système de vérification d'identité peut être confronté deux types de tests[27] :

- Test de client lors desquels l'échantillon biométrique présenté au système correspond à l'identité clamée.
- Test imposteur lors desquels l'échantillon biométrique présenté au système provient d'un individu inconnu du système.

Le système automatique doit répondre à chaque tentative d'authentification auquel il fait face par une décision binaire. Il peut donc engendrer deux types d'erreurs :

Faux rejet **(**FR**)** erreur commise lorsque le système rejette, à tort, un client légitime (i.e. erreur commise lors d'un test client) ;

Fausse acceptation **(**FA**)** erreur commise lorsqu'un imposteur est mal encontre sèment accepté en tant qu'utilisateur légitime (i.e. erreur commise lors d'un test imposteur).

Le choix d'un seuil a une incidence directe sur les performances du système. Pour un système idéal, les scores obtenus par les clients seront tous plus élevés que les scores obtenus par les imposteurs. Dans ce cas, le seuil à fixer se situe entre le score imposteur le plus élevé et le score client le plus faible, assurant ainsi une authentification parfaite (Fig 8).

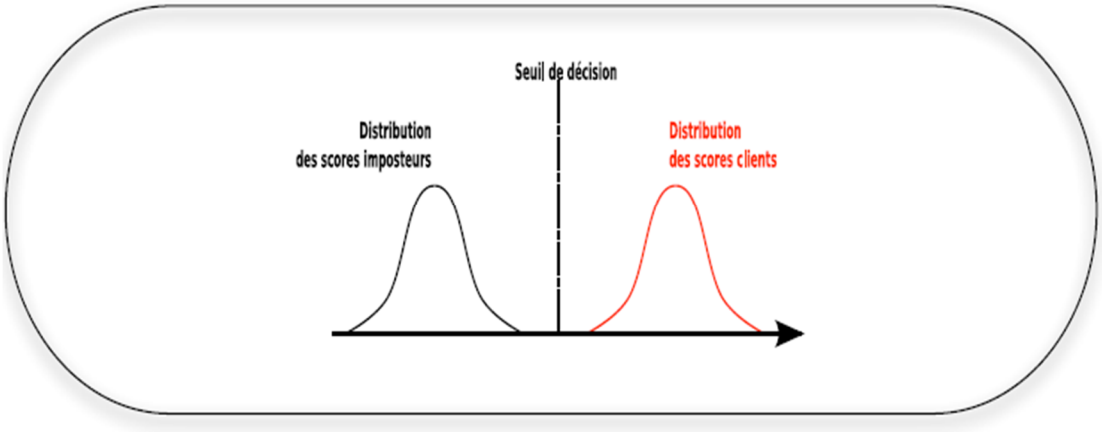



**Figure 8: Répartition des scores clients et imposteurs et seuil de décision d'un système**

En pratique, les distributions des scores clients et imposteurs se superposent partiellement. Ce cas ne permet pas une authentification parfaite et des erreurs de type faux rejets et fausses acceptations apparaissent. Le choix du seuil influe sur le taux de faux rejets et de fausses acceptations. Cet effet est illustré par la figure9.

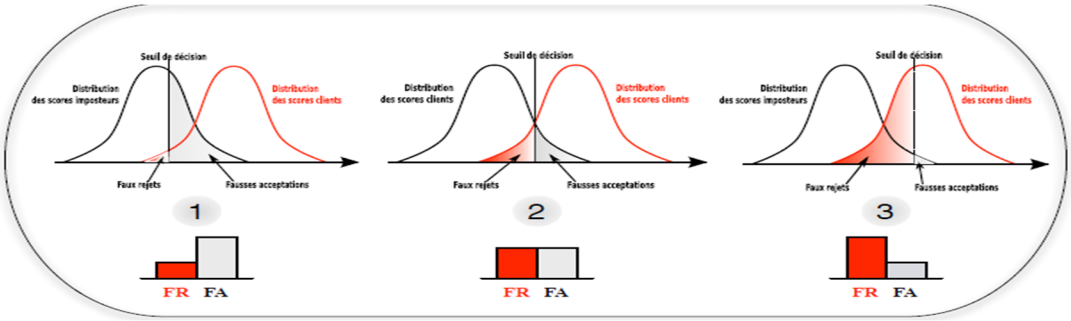



**Figure 9: Influence du seuil de décision sur les erreurs d'un système de reconnaissance biométrie.**

1. Seuil de décision choisi dans le but de réduire le nombre de faux rejets.
2. Seuil de décision choisi pour obtenir autant de faux rejets que de fausse d'acceptation.
3. Seuil de décision choisi dans le but de réduire le nombre de fausse d'acceptation

Pour un seuil de décision fixé les taux de faux rejet p(FR) et de fausse acceptation p(FA) que l'utilisation de ce seuil occasionne peuvent être calculés a posteriori.

$$p(FA) = \frac{nombre\ de\ tests\ dont\ resulte\ une\ fausse\ acceptation}{nombre\ de\ test\ imposteur}$$

$$p(FR) = \frac{nombre\ de\ Test\ dont\ resulte\ un\ faux\ rejet}{nombre\ de\ Tests\ client}$$

À chaque valeur de seuil est associé un couple $(p(FA), p(FR))$ et l'ensemble des couples obtenus peut être représenté sous la forme d'une courbe ROC (Receiver Operating Characteristic) [28] ou, comme sur la figure 8, d'une courbe DET (Detection Error Tradeoff). La courbe DET diffère principalement de la courbe ROC par l'échelle basée sur une distribution normale qui se substitue à l'échelle linéaire [29].

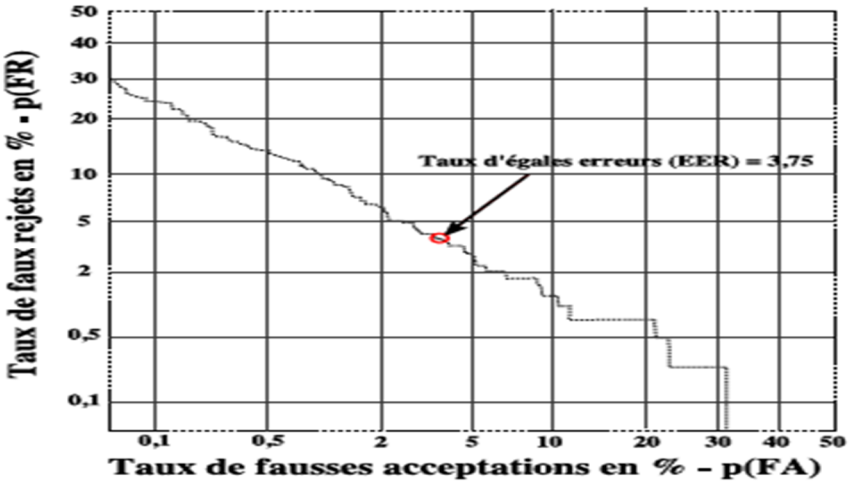



**Figure 10: Exemple de représentation des performances d'un système de vérification d'identité par une courbe DET**

## 2. Identification du locuteur

Consiste à reconnaître un locuteur parmi un ensemble de locuteurs en comparant son identité vocale à des références connues. Les performances du système d'identification sont données en termes de taux d'identification correcte $I_c$ ou incorrecte $I_i$.

$$I_c = \frac{Nombre\ de\ Tests\ correctement\ identifiés}{Nombre\ total\ de\ tentative}$$

$$I_i = \frac{Nombre\ de\ Tests\ mal}{Nombre\ total\ de\ tentative}$$

# Conclusion

Dans ce chapitre nous avons introduit le principe de la reconnaissance automatique du locuteur ainsi que la déférente étape du système. La reconnaissance automatique du locuteur est probablement la méthode la plus ergonomique pour résoudre les problèmes d'accès. Cependant, la voix ne peut être considérée comme une caractéristique biométrique d'une personne compte tenu de la variabilité intra-locuteur.

Un système de reconnaissance de locuteur procède généralement en trois étapes : l'analyse acoustique de signal de parole, la modélisation du locuteur et une dernière étape de décision. En analyse acoustique, les MFCC sont les coefficients acoustiques les plus répandus. Quant à la modélisation, l'approche GMM constitue l'état de l'art en RAL, en mode indépendant du texte.la décision d'un système de reconnaissance automatique du locuteur est basée sur les deux processus d'identification et/ou vérification de locuteur, et cela quelle que soit l'application ou la tâche visée.

# Chapitre II

Reconnaissance du locuteur par mélange du Gaussiennes

# I. L'approche statistique GMM-UBM en RAL

## Introduction

Les approches génératives utilisées en reconnaissance du locuteur reposent essentiellement sur le paradigme GMM/UBM. Cette partie présente ce paradigme de façon détaillée tout en se focalisant sur son utilisation dans le cadre de la reconnaissance du locuteur indépendante du texte [27].

### 1. Schéma général

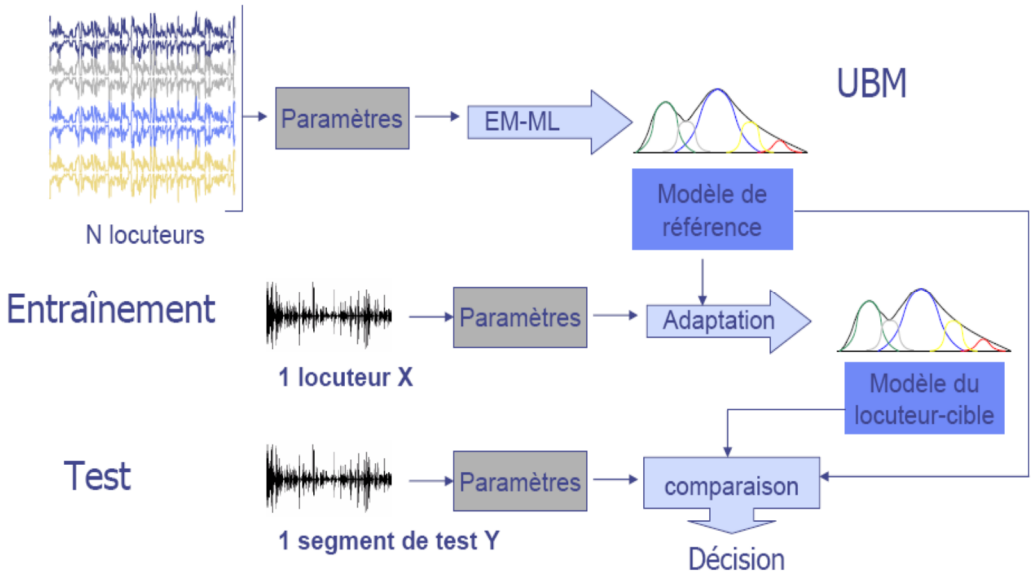



**Figure 11: Shéma de la méthode GMM-UBM pour la VAL indépendant du texte**

Le schéma de fonctionnement est représenté sur la figure 11 Les différents modules représentés sont :

– le module « Paramètres ». Il permet d'extraire les paramètres du signal de parole pertinents pour la VAL.

– les modules «Modèle de référence et modèle du locuteur-cible», qui estiment, à partir des données d'apprentissage, les modèles statistiques des locuteurs.

– le module «Comparaison», qui calcule la mesure de similarité entre l'échantillon de test et le modèle de locuteur cible. Il fournit la décision de vérification. La suite de ce chapitre décrit chacun de ces modules.

### 2. Les Mélanges de Gaussiennes en RAL

La reconnaissance du locuteur s'appuie sur une représentation discrète du signal de parole. Celui-ci est transformé en une séquence de vecteurs de paramètres, dont la fréquence d'échantillonnage est généralement 100Hz.

Considérons que chaque vecteur de paramètres extrait d'un signal de parole est une réalisation d'une variable aléatoire multidimensionnelle. Les approches génératives en reconnaissance du locuteur reposent sur l'hypothèse qu'il existe une fonction injective de l'ensemble des locuteurs dans l'espace des fonctions de densité de probabilité. Cette hypothèse suppose, plus précisément, que les vecteurs de paramètres provenant d'un locuteur suivent une loi de probabilité propre à ce locuteur [27].

La complexité de ces fonctions de densité nous conduit à rechercher une approximation suffisant à la résolution du problème de reconnaissance du locuteur. Dans les Méthodes Statistiques du Second Ordre (MSSO) [30],les locuteurs sont représentés par une loi Gaussienne, c'est à dire un doublé $(\mu, \Sigma)$, où $\mu$ est le vecteur moyen de la Gaussienne et $\Sigma$ la matrice de covariance, estimée à partir de la séquence acoustique d'apprentissage$X$. Nous avons souligné la simplicité de la modélisation des locuteurs par MSSO et le fait qu'elle limite la granularité de modélisation des variations acoustiques.

L'utilisation de mélanges de Gaussiennes (GMM) permet d'obtenir une approximation plus précise de la fonction de densité de probabilité caractéristique des locuteurs, tout en restant relativement simple à estimer [31],[32],[33]. La densité de probabilité d'un mélange de $N$distributions Gaussienne est :

$$p(\mathcal{X}|\Theta) = \sum_{i=1}^{N} \gamma_i \, \mathcal{N}(\mathcal{X}, \mu_i, \Sigma_i)$$

Telle que,$\sum_i \gamma_i = 1 \; et \; \forall i, \gamma_i \geq 0. \gamma_i, \mu_i \; et \; \Sigma_i$ sont respectivement le poids, le vecteur moyens et la matrice de covariance de la distribution $i$ dans la mixture. $\Theta = [\mu, \Sigma, \gamma]^T$ est le vecteur de paramètres global de mixture de Gaussienne. La densité de probabilité gaussienne $\mathcal{N}(\mathcal{X}, \mu, \Sigma)$ est défini par l'équation suivante[27] :

$$\frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(\mathcal{X}-\mu)^T\Sigma^{-1}(\mathcal{X}-\mu)\right]}$$

En reconnaissance du locuteur, la matrice de covariance est généralement supposée diagonale. La vraisemblance pour qu'un vecteur de paramètres $X$ait été produit par le GMM de vecteur de paramètres $\Theta$ est[27] :

$$f(X|\Theta) = \sum_{i=1}^{N} \gamma_i \, \mathcal{N}(X, \mu_i, \Sigma_i)$$

La valeur moyenne de la log-vraisemblance pour une séquence $X$ de paramètre $X_t$,$t \in [1; T]$ et un GMM $\Theta$,que notons $\ell(X|\Theta) \; est$ [27]:

$$\ell(X|\Theta) = \log[f(X|\Theta)] = \frac{1}{T}\sum_t \log f(X_t|\Theta)$$

## 3. La densité d'une mélange de gaussienne

La détermination du $\Theta$ pour une collection de trames $\mathcal{X}$ s'effectue par l'algorithme d'apprentissage EM (Expectation Maximisation) [1]. Cet algorithme itératif effectue à chaque étape deux phases Expectation et Maximisation destiné à augmenter la vraisemblances des donnée d'apprentissage au modèle de gaussiennes(d'où le suffixe ML Maximum Likelihood ajouté à son nom).L'algorithme garantit à chaque itération la croissance d'une fonction objective de vraisemblances des paramètres sachant $\mathcal{X}$.Il converge vers un maximum de vraisemblance, mais seulement locales, dans le champs d'optimisation de la fonction de densité [34] .

Le paramètre $\Theta = \{\gamma, \mu, \Sigma\}$ de la mixture contient $G[\frac{F(F+1)+F+1}{2}]$ valeurs à estimer. Or la collection de trames $\mathcal{X}$ pour un segment de voix de durée initiale allant de 30 secondes à quelques minutes contient un effectif de trames de l'ordre de 5 à 20 000 trames après VAD. Pour une dimension de l'espace acoustique $F = 50$ et une mixture G=64 gaussienne, le paramètre $\Theta$ contient déjà 81664 valeurs à estimer. Et ici le nombre G de 64 s'avère assez loin de la quantité minimale empirique nécessaire pour façonner une mixture de segment de voix de vraisemblances satisfaisante. L'estimation EM-ML conduit alors, par sous-apprentissage, à un modèle médiocre[34].

L'alternative consiste à la matrice de covariance $\Sigma$ la contrainte de diagonalité dans l'algorithme EM. La matrice $\Sigma$ de chaque gaussienne est seulement rempli avec la diagonale $\sigma^2$des variances. Cette contrainte peut apparaitre restrictive, voire peut réaliste, mais en réduisant de $G[\frac{F(F+1)+F+1}{2}]$ à $G(2F + 1)$ le nombre de valeurs à estimer, elle permet l'accroissement du nombre G de gaussiennes de la mixture et donc de la précision locale de l'estimation[34].

## 4. Mesure de vraisemblances

Etant donné la collection de trames $\mathcal{X}$d'un énoncé de voix et un locuteur présumés, le système doit déterminer la probabilité de l'hypothèse locuteur $H_0$ : «cet énoncé de voix est prononcé par s » cette probabilité s'ecrit :

$$P(H_0|\mathcal{X}) = \frac{P(\mathcal{X}|H_0)P(H_0)}{P(\mathcal{X})}$$

Etant définie une densité pour la loi du modèle de s, le facteur de vraisemblance $P(\mathcal{X}|H_0)$est alors estimé par la valeur de cette densité pour$\mathcal{X}$.Sous l'hypothèse d'indépendance des trames $x$ de$\mathcal{X}$, ce facteur est le produit des vraisemblances$P(x|H_0)$.

Le problème du calcul du dénominateur est levé en introduisant l'hypothèse$H_1$,"non locuteur", complémentaire de$H_0$. La probabilité de l'équation précédente s'écrit :

$$P(H_0|\mathcal{X}) = \frac{P(\mathcal{X}|H_0)P(H_0)}{P(\mathcal{X}|H_0)P(H_0) + P(\mathcal{X}|H_1)P(H_1)}$$

Il est alors possible de calculer $P(H_0|\mathcal{X})$à condition d'avoir défini la densité de la loi des "imposteurs" de s (tous les locuteurs hormis s) et estimé les fréquences a priori $P(H_1)$et $(H_0)$.

La comparaison des probabilités $P(\mathcal{X}|H_0)$ et $P(\mathcal{X}|H_1)$permet de mesurer le risque associé à la décision d'acceptation. Le ratio des hypothèses complémentaires (likelhood ratio) est défini par :

$$LR(H_0, H_1|\mathcal{X}) = \frac{P(H_0|\mathcal{X})}{P(H_1|\mathcal{X})} = \frac{P(\mathcal{X}|H_0)P(H_0)}{P(\mathcal{X}|H_1)P(H_1)}$$

En VAL, une décision binaire d'acceptation ou rejet est obtenue en fixant un seuil de décision $\Omega$à$LR(H_0, H_1|\mathcal{X})$.Après incorporation des probabilités a priori $P(H_0)$ et $P(H_1)$à ce seuil, la décision dépend seulement de la valeur$\frac{P(H_0|\mathcal{X})}{P(H_1|\mathcal{X})}$ :

$H_0$Acceptée si $\frac{P(H_0|\mathcal{X})}{P(H_1|\mathcal{X})} > \Omega$,rejetée sinon.

Dans le cadre de la modélisation par GMM, la vraisemblance $P(\mathcal{X}|H_0)$ est évaluée par la densité de $\mathcal{X}$suivant la mixture de gaussiennes du locuteur s. La vraisemblance $P(\mathcal{X}|H_1)$ nécessite l'estimation d'un modèle GMM des imposteurs de s.

## 5. L'algorithme EM (Expectation Maximisation)

L'algorithme EM fait intervenir à la fois des observations $\mathcal{X}$ et des variables manquantes (l'indice de la gaussienne $m = 1, \dots, M$). Cet algorithme maximise de façon itérative, la fonction de la vraisemblance. Cette maximisation n'est pas directe, elle fait intervenir la fonction auxiliaire $Q(\theta, \theta^{(t)})$ qui était défini comme étant l'espérance mathématique du logarithme de la vraisemblance jointe (incluant les variables observées et les variables cachés) sur l'ensemble complet des variables d'entrainement calculés sur la base des paramètres courants :

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N} p(m|x_n, \theta^{(t)}) \log p(x_n, m|\theta)$$

Où $\theta$ désignent des paramètres à estimer $(\mu_m, \Sigma_m, \gamma_m)$ et $\theta^{(t)}$ l'ensemble des paramètres à estimer à l'itération $t$. Ce qui donne après le calcul :

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N} \pi_{n,m}^{(t)} \left[\log \gamma_m - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_m|\right] - \sum_{m=1}^{M}\sum_{n=1}^{N} \pi_{n,m}^{(t)}[\frac{1}{2}(x_n - \mu_m)^t \Sigma_m^{-1}(x_n - \mu_m)]$$

Où $\pi_{n,m}^{(t)}$ est une probabilité a posteriori estimée à l'itération $t$ :

$$\pi_{n,m}^{(t)} = \frac{\gamma_m^t p(x_n|\mu_m^{(t)}, \Sigma_m^{(t)})}{\sum_{k=1}^{M} \gamma_k^t \, p(x_n|\mu_n^{(t)}, \Sigma_n^{(t)})}$$

En supposant que $p(x_n|\theta)$ sont des densités gaussiennes à matrices de covariances diagonales, l'expression de la fonction auxiliaire devient :

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M}\sum_{n=1}^{N} \pi_{m,n}^{(t)} \log(\gamma_m) - \frac{1}{2}\sum_{m=1}^{M}\sum_{n=1}^{N} \pi_{m,n}^{(t)}[Cte + \log \sigma_m^2 + \frac{(x_n - \mu_m)^2}{\sigma_m^2}]$$

Où $\sigma_m^2$ est un élément diagonal de la matrice de covariance.

Les paramètres sont estimés en annulant la dérivée partielle de la fonction auxiliaire $Q$ par rapport à chacun de ceux-ci .Le cas des poids de composantes de mélanges $\gamma_m$ est assez simple puisqu'il s'agit de paramètres scalaires. Ceci dit, il faut tenir le compte de la contrainte de qui existe sur des paramètres ($\sum_{m=1}^{M} \gamma_m = 1$).La maximisation sous contrainte résout simplement en introduisant un multiplicateur de Lagrange associé à cette contrainte et l'on obtient :

$$\gamma_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \pi_{n,m}^{(t+1)}$$

En ce qui concerne les vecteurs moyennes, on montre que les formules d'estimation sont données par :

$$\mu_m^{(t+1)} = \frac{\sum_{n=1}^{N} \pi_{n,m}^{(t)} x_n}{\sum_{n=1}^{N} \pi_{n,m}^{(t)}}$$

Et pour les variances :

$$\sigma_m^{2(t+1)} = \frac{\sum_{n=1}^{N} \pi_{n,m}^{(t)} (x_n - \mu_m^{(t)})^2}{\sum_{n=1}^{N} \pi_{n,m}^{(t)}}$$

## 6. Le modèle GMM-UBM

Les paramètres GMM des imposteurs de s, nécessaires au calcul de $P(\mathcal{X}|H_1)$peuvent être estimés à partir des trames d'une cohorte d'imposteurs, de modèles "proches" de s[35, 36]. Mais la difficulté à déterminer la liste de ses imposteurs pour un locuteur donné, ainsi qu'à gérer des énoncés "éloignés" à partir d'imposteurs proches, a conduit à élaborer un modèle GMM unique appelé modèle du monde [7].

La stratégie adoptée consiste à collecter un nombre considérable de trames, issus de sessions de voix différentes et de locuteurs distincts, pour entraîner une mixture de gaussiennes générique. L'initialisation des paramètres peut être aléatoire, arbitraire ou assistée (par exemple par un dictionnaire VQ).

Le modèle de mixtures de gaussiennes obtenu porte le nom de modèle du monde (GMM-UBM Universel Background Model). Le GMM-UBM ne fait pas que modéliser une hypothèse "non-locuteur". Il structure l'espace acoustique[37] en régions de densité élevée, donc en classes probabilistes, par une technique d'apprentissage, hors toute considération phonologique ou linguistique. Le résultat obtenu peut donc être considéré, si la collection est suffisamment vaste et variée, comme une classification probabiliste de l'espace acoustique, son nom UBM[34].

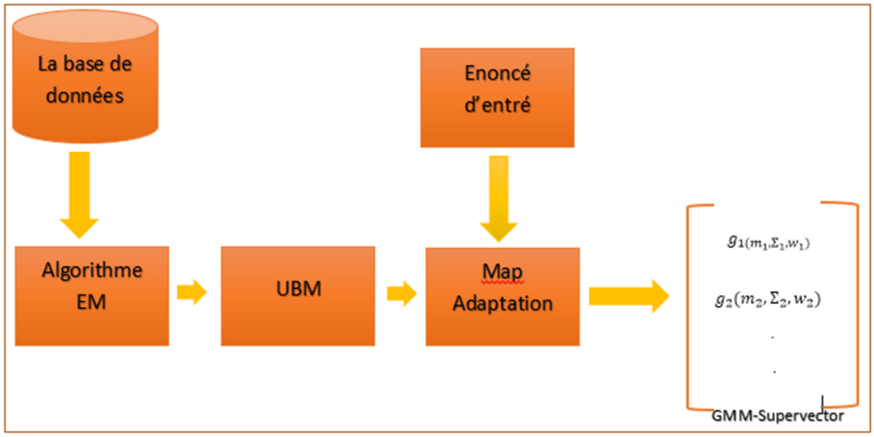



**Figure 12 : Processus d'extraction des supervecteurs GMM à partir d'un énoncé**

Le choix des données d'apprentissage du GMM-UBM est primordial dans la qualité d'un système de reconnaissance. Quantité et variabilité des données sont des facteurs prévisibles de qualité. Les modélisations par GMM-UBM privilégient les segments bien renseignés (longue durée, grand nombre de locuteurs). L'apport d'informations plus pauvres (enregistrements bruités, durée insuffisante,...) ne contribue que de manière minime à la qualité du résultat.

## 7. Adaptation à Postériori MAP

Malgré la réduction du nombre de paramètres GMM à estimer lorsque ses matrices de covariance par composante sont diagonales, celui-ci reste élevé pour un segment de voix. Pour un nombre de gaussiennes G de 512 valeur réaliste pour espérer constituer un GMM-UBM robuste, le nombre de valeurs à estimer, dans l'exemple LIA choisi, est de $512(2 \times 50 + 1) = 51712$Disposant d'une collection de trames ne dépassant pas en général un effectif de 20 000 après VAD, le paramètre $\Theta$ d'un énoncé de voix reste trop volumineux pour espérer une approximation par GMM satisfaisante. Dans le cas de segments courts (moins de 20 secondes), il s'avère même tout à fait inadapté.

La solution consistant à réduire la durée de la fenêtre de Hamming de découpage en trames (habituellement fixée à une durée de 20 millisecondes avec un pas de 10 millisecondes), afin d'augmenter le nombre de trames fournies à l'algorithme EM, n'est pas pertinente : une "unité élémentaire" de parole couvre une durée de 50 à 150 millisecondes pour les occlusives[38], autour de 150 millisecondes pour les voyelles . S'il s'agit de parole, l'émiettement du signal instantané n'apporte plus d'informations à son analyse.

Etant donné un modèle GMM-UBM, la représentation d'un segment de voix s'obtient par adaptation de ce modèle aux données contingentes du segment. La méthode utilisée s'appuie sur la notion statistique de Maximum à Posteriori (MAP)[39]et [7]pour son application en reconnaissance du locuteur). Le paramètre $\Theta_{UBM} = \{\gamma, \mu, \Sigma\}$du GMM-UBM est adapté aux nouvelles données contingentes aux trames du segment à représenter, formant un paramètre adapté$\widehat{\Theta} = \{\hat{\gamma}, \hat{\mu}, \hat{\Sigma}\}$.

L'adaptation selon le critère MAP s'effectue par algorithme EM. A chaque itération t, la phase d'expectation met à jour la variable latente y par son espérance :

$$\forall x \in \mathcal{X}, y_g^{(t)} = \frac{\gamma_g^t f(x|\mu_g^{(t)}, \Sigma_g^t)}{\sum_{g'=1}^{G} \gamma_{g'}^t f(x|\mu_{g'}^{(t)}, \Sigma_{g'}^t)}$$

Et celle de maximisation évalue alors (itération t+1) les paramètres $\hat{\gamma}, \hat{\mu}, \hat{\Sigma}$ par maximum de vraisemblance :

$$\gamma_g^{(t+1)} = \frac{n_g}{n_g + \tau_g^{[w]}} \left( \frac{1}{n} \sum y_g^{(t)} \right) + \frac{\tau_g^{[w]}}{n_g + \tau_g^{[w]}} \gamma_g$$

Où $n$taille de la collection $\mathcal{X}$ et $n_g = \sum y_g^{(t)}$

$$\mu_g^{(t+1)} = \frac{n_g}{n_g + \tau_g^{[\mu]}} \left( \frac{\sum y_g^{(t)} x}{\sum y_g^{(t)}} \right) + \frac{\tau_g^{[\mu]}}{n_g + \tau_g^{[\mu]}} \mu_g$$

$$\Sigma_g^{(t+1)} = \frac{n_g}{n_g + \tau_g^{[\Sigma]}} \left( \frac{\sum y_g^{(t)} \left( x - \mu_g^{(t+1)} \right) \left( x - \mu_g^{(t+1)} \right)}{\sum y_g^{(t)}} \right) + \frac{\tau_g^{[\Sigma]}}{n_g + \tau_g^{[\Sigma]}} \Sigma_g$$

Nous détaillons ici ces formules pour y souligner le rôle des facteurs $\tau$.Qualifiables de facteurs de confiance (relevant factor : facteur latent) ou de pertinence, ils contrôlent le degré d'adaptation de chaque paramètre UBM aux données. En théorie, ils dépendent du paramètre à estimer. En pratique, ils sont confondu, qu'il s'agisse de poids, moyenne ou matrice de covariance, en un seul vecteur $(\tau_g)_{g=1}^{G}$.Lorsque le nombre $\tau_g$ tend vers l'infini, les estimations $\gamma_g^{(t+1)}, \mu_g^{(t+1)}, \Sigma_g^{(t+1)}$ tendent ves les paramètres $\gamma_g, \mu_g, \Sigma_g$ de l'UBM.Lorsque ce nombre tend vers 0,ces estimations sont seulement déterminées par la collection de trames du segment à représenter (algorithme EM-ML).Le problème du manque de données dans la seule collection de trames d'un segment de voix est donc évité, par balance entre un modèle du monde robuste et les nouvelles observations.

Comme le montrent les formules précédentes, un facteur$\tau_g$indique une quantité d'informations associée à chaque composante du GMM-UBM. Le prolongement empirique de l'adaptation MAP consiste à ramener ces valeurs$\tau_g$à des échelles qui les rendent comparables aux effectifs $n_g$des segments utilisés et à conserver les valeurs fournissant les meilleures performances.

Il s'est avéré que la seule adaptation des moyennes suffisait à atteindre les performances optimales et cette technique s'est généralisée dans les systèmes état de l'art. L'adaptation MAP fournit alors en sortie une batterie de vecteurs de moyennes adaptées $\widehat{\mu_g} \in \mathbb{R}^F$(un vecteur par composante gaussienne du GMM).Ces $G$vecteurs peuvent être alors exprimés, par concaténation, sous forme d'un supervecteurs (figure 13) de dimension $GF$:

$$s = [\widehat{\mu_1}, \widehat{\mu_2}, \dots, \widehat{\mu_G}] \in \mathbb{R}^{GF}$$

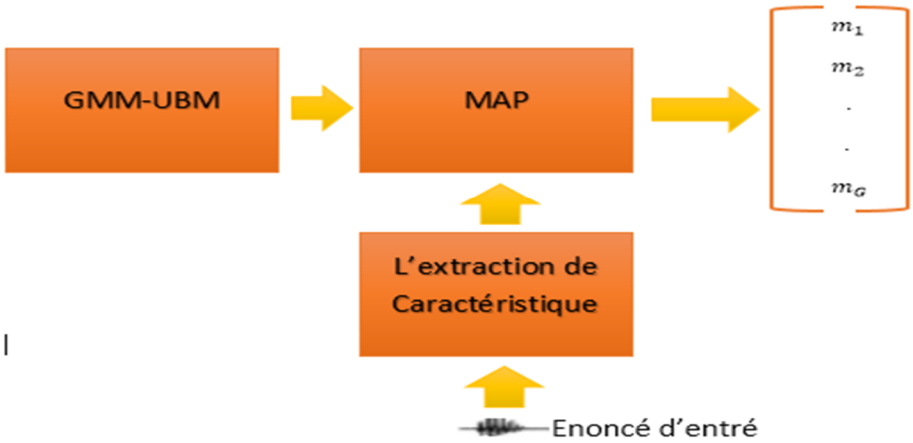



**Figure 13: Processus de génération des supervecteurs via MAP**

Ce supervecteur constitue la statistique d'ordre 1 issue du GMM-UBM. Les procédures succédant à la production de modèles adaptés nécessitent d'ajouter à la représentation la statistique d'ordre 0 des trames du segment considéré. Il s'agit du vecteur n des effectifs de trames associées (probabilistiquement) à chaque gaussienne :

$$n = (n_g)_{g=1}^{G} \in \mathbb{R}^G$$

Où $n_g$ effectif de trames affectées à la gaussienne g, est égal à la somme sur toutes les trames de la collection des probabilités d'occupation de cette gaussienne.

La représentation par adaptation du GMM-UBM fournit donc en sortie, dans sa version la plus courante, une numérisation du signal de parole sous forme vectorielle de dimension $G(F + 1)$.

## 8. Adaptation par MLLR

MLLR (Maximum likelihood lineaire regression) sont utilisés par des chercheurs comme des entrés au SVM. Initialement MLLR a été développé pour l'adaptation du locuteur dans la reconnaissance de la parole, il a également été utilisé dans la reconnaissance du locuteur comme une alternative à l'adaptation maximale a posteriori (MAP) de l'UBM.

Il existe deux types d'expansion qui sont produit par l'adaptation de MLLR. Supposons le gaussien :

$$g(x) = \sum_{g=1}^{G} \gamma_i \mathcal{N}(x, \mu_i, \Sigma_i)$$

Avec $\mu_i$ est le moyen $\Sigma_i$ est la matrice covariance, en adaptant les moyens UBM par MLLR à une énoncé donnée $utt_\alpha$, après l'adaptation, une matrice de Transformation $A$ et un vecteur $b$ qui sont utilisés pour calculer des nouveaux moyens $\mu_g^\alpha$ :

$$\mu_g^\alpha = A\mu_g + b$$

$\mu_g$ Représente le vecteur moyen UBM et $\mu_g^{\alpha}$ est le vecteur moyen adapté, les paramètres $A$ et $b$ définit les transformations linéaire, ils sont estimé par le Maximum de Vraisemblance.

La première expansion est le super vecteur GMM $m$ qui est construit en empilant les moyens de modèle adapté. La seconde expansion est le vecteur de transformation MLLR τ qui consiste à empiler les lignes transposés de la matrice A, ces lignes sont séparés par entrées correspondant du vecteur b. la figure suivante montre ce processus :

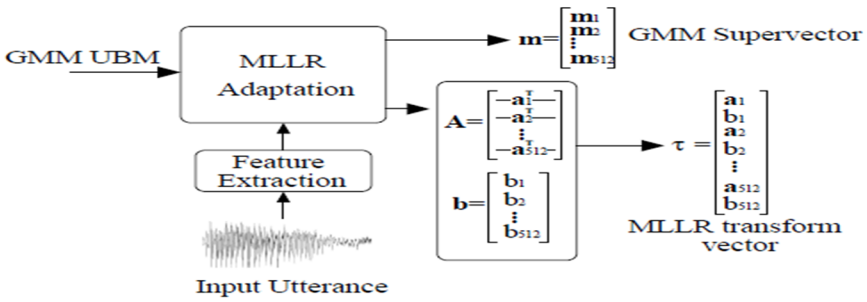



**Figure 14: Adaptation MLLR**

## 9. Calcul de score

Une fois que des objets à comparer ont été numérisés, il est toujours possible de mesurer leur proximité deux à deux et d'en déduire une décision, lorsqu'une variable latente à déterminer est supposée maintenir proche ces productions. Dans le cas de la reconnaissance du locuteur, l'identité de la personne constitue la variable cachée. La mesure de similarité entre deux représentations (scoring) gagne à s'appuyer sur des modèles préliminaires dans lesquels la machine a été renseignée sur l'identité du locuteur et des hypothèses a priori ont été avancées.

Le champ d'investigation des mesures de similarité n'est malheureusement pas borné et nous développerons seulement, dans cette section et par la suite, autour des modèles et scorings actuellement appliqués aux représentations issues du GMM-UBM.

Plusieurs alternatives sont possibles pour comparer segment de test et modèle locuteur :

- un modèle peut être réalisé à partir du segment de test, puis ce modèle est comparé au modèle du locuteur-cible. La mesure s'effectue alors directement entre les représentations vectorielles issues des collections de trames.

- les trames du segment de test sont présentées une à une au modèle du locuteur cible et ses comparaisons synthétisées dans une valeur unique. La comparaison peut s'effectuer par mesure de vraisemblance de la trame sur le modèle a priori du locuteur-cible et la moyenne de ces mesures renvoyée comme score. Plutôt qu'une vraisemblance d'une moyenne des trames, c'est la moyenne de leur vraisemblance qui est produite.

Cette seconde alternative, jugée plus précise, a été le plus souvent élue. Dans la logique d'une compétition affrontant modèle-cible et modèle-imposteur, le ratio de vraisemblance de la trame entre ces deux modèles (Log Likelihood Ratio by Frame, LLR-by frame) retournera une mesure plus homogène. Etant donnée la collection $\mathcal{X}$ de trames d'un segment de test et $s$ le locuteur-cible, la mesure de vraisemblance de $\mathcal{X}$ sous l'hypothèse que ses trames sont issues du modèle $s$ s'écrit :

$$P(\mathcal{X}|s) = \prod_{x\in\mathcal{X}} P(x|s).$$

Les trames $x$ de $\mathcal{X}$ étant supposées indépendantes.

Sous l'hypothèse inverse (notée s), le GMM-UBM est utilisé comme modèle imposteurs et l'on peut écrire :

$$P(\mathcal{X}|\bar{s}) = P(\mathcal{X}|UBM) = \prod_{x\in\mathcal{X}} P(x|UBM).$$

Notant $n$ le nombre de trames de la collection, ces vraisemblances sont normalisées pour obtenir un score homogène, en calculant leur moyenne (géométrique, pour tenir compte d'un produit et non d'une somme). Le score LLR-by-frame est défini par :

$$score(\mathcal{X}, s) = \log \frac{P(\mathcal{X}|s)^{\frac{1}{n}}}{P(\mathcal{X}|UBM)^{\frac{1}{n}}} = \frac{1}{n}\{\sum_{x\in\mathcal{X}} (\log \sum_{g=1}^{G} w_g^{(s)} \mathcal{N}\left(x\middle|\mu_g^{(s)}, \Sigma_g^{(s)}\right) - \log \sum_{g=1}^{G} w_g \mathcal{N}(x|\mu_g, \Sigma_g))\}$$

où les paramètres $w_g^{(s)}, \mu_g^{(s)}, \Sigma_g^{(s)}$ (resp. $w_g, \mu_g, \Sigma_g$) Sont issus de l'adaptation du GMMUBM à s (resp. sont ceux du GMM-UBM).

La représentation pratique la plus usuelle n'adaptant que la moyenne, ce score se réduit à :

$$score(\mathcal{X}, s) = \frac{1}{n}\{\sum_{x\in\mathcal{X}} (\log \sum_{g=1}^{G} w_g \mathcal{N}\left(x\middle|\mu_g^{(s)}, \Sigma_g\right) - \log \sum_{g=1}^{G} w_g \mathcal{N}(x|\mu_g, \Sigma_g))\}$$

## Conclusion

Les mélanges de Gaussiennes constituent l'état de l'art en reconnaissance automatique du locuteur, en mode indépendant du texte. Il existe plusieurs techniques pour apprendre un modèle GMM. En premier lieu, l'algorithme EM permet d'estimer les paramètres du modèle tout en offrant un formalisme théorique et une preuve de convergence. Malheureusement, ses limites apparaissent lorsqu'on dispose de peu de donnée. Dans ce cas, il est plus judicieux d'utiliser un apprentissage par adaptation MAP.

# Chapitre III

Machine à Vecteur supports (SVM)

# I. Machines à vecteurs supports

Les Support Vector Machines (SVM) sont des nouvelles techniques discriminantes dans la théorie de l'apprentissage statistique. Elles ont été proposées en1995 par V. Vapnik dans son livre « The nature of statistical learning theory ».Elles permettent d'aborder plusieurs problèmes divers et variés comme la régression la classification, la fusion etc (Kharroubi, 2002).

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 vecteur appartenant aux différentes classes se trouvent de différents côté de l'hyperplan.

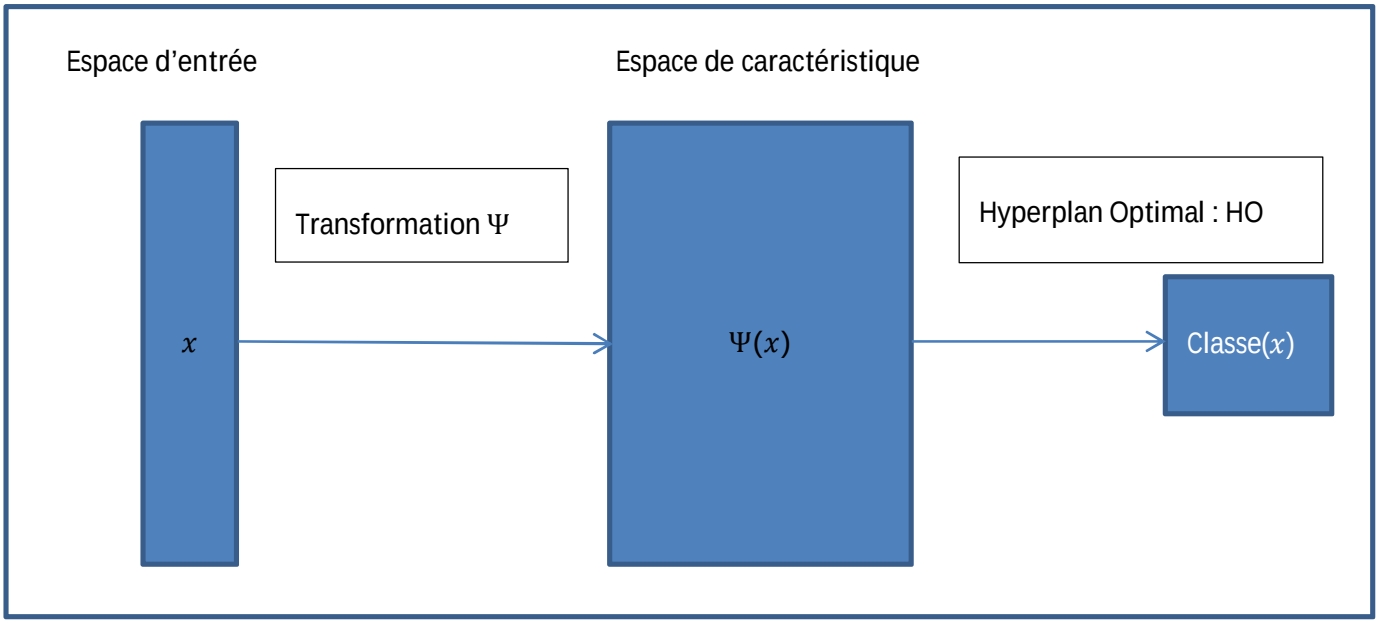



**Figure 15: Principe des techniques SVM**

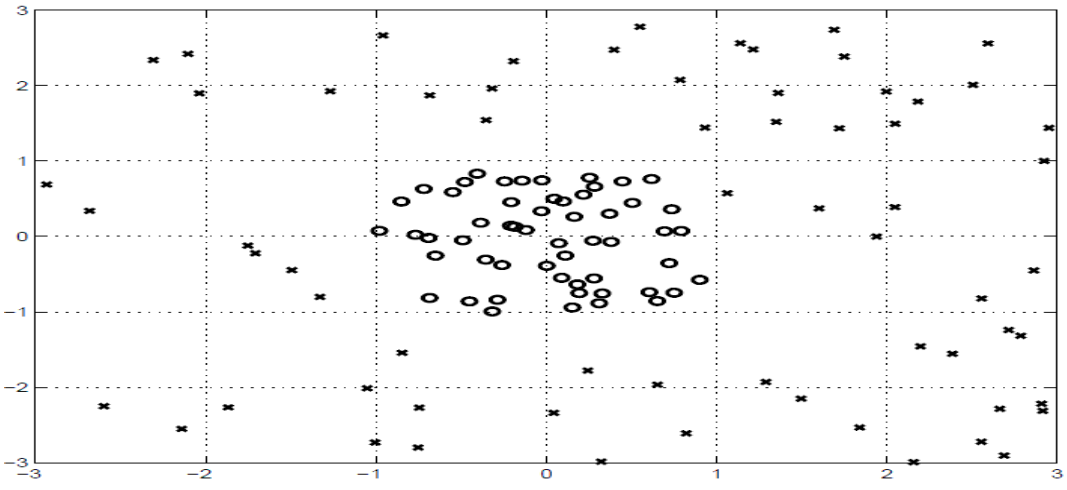
- La plus petite distance entre les vecteurs et l'hyperplan (lsa marge) soit maximale.

Exemple :

Pour savoir une idée plus claire sur SVM,voici un exemple inspirée de travail de B.Schlkof[41]qui met en pratique le principe des SVM. Dans cet exemple, les données non linéairement séparable dans $\mathbb{R}^2$deviennent linéairement séparable dans $\mathbb{R}^3$ grâce à la transformation $\psi$ défini par:

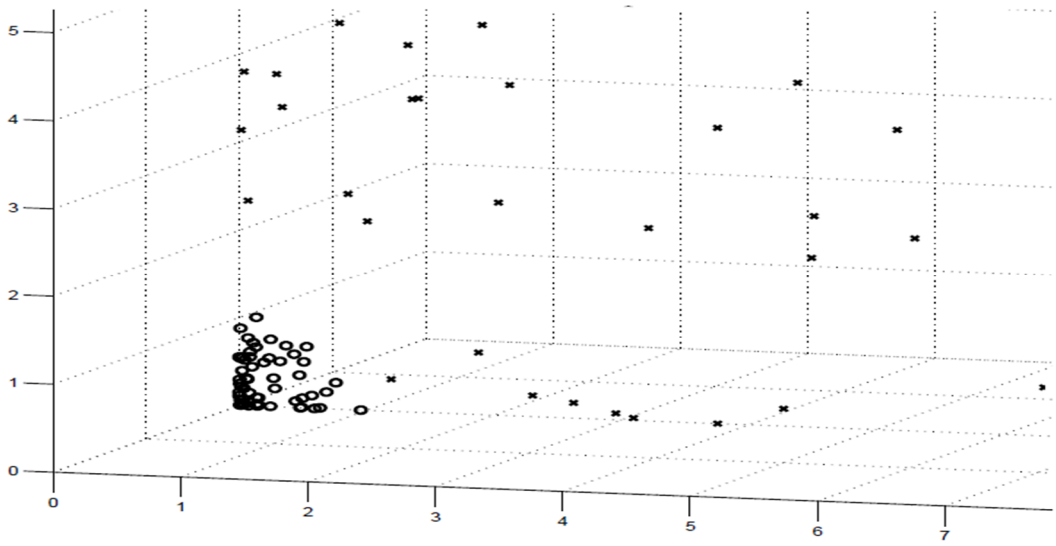
$$(x_1, x_2) \rightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2) \in \mathbb{R}^2 \times \mathbb{R}^3$$

La figure 15 représente une simulation de cette transformation que nous avons réalisée par Matlabe. Les données de la figure A ont été tiré aléatoirement en $\mathbb{R}^2$ et la figure B représente l'image de ces données dans $\mathbb{R}^3$ suivant la transformation $\psi$.




A :

A : Exemple tirés aléatoirement dans $\mathbb{R}^2$ appartenant à 2 classes non linéairement séparable




B : l'image des exemples de la figure 15.A dans $\mathbb{R}^3$ en utilisant la transformation $\psi$

**Figure 16: Exemple montrant l'efficacité d'une transformation dans un espace de plus grande dimension pour faciliter le classement**

# 1. Construction de l'hyperplan optimal

Pour bien décrire la technique de construction de l'hyperplan optimale séparant des données appartenant à 2 classes différentes dans deux cas différents : le cas des données linéairement séparable et le cas des données non linéairement séparable. Soit $D$ une base de donnée de $m$ points appartenant à 2 classes différentes $\{-1,1\}$ défini par (Kharroubi, 2002) :

$$D = \{(x_i, y_i), x_i \in \mathbb{R}^n, y_i \in \{-1,1\}, i = 1 \dots m\}$$

## 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 16 donne une représentation visuelle de l'HO dans le cas des données linéairement séparables.

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

$$H(x) = \begin{cases} w.x_i + b \geq 1, & y_i = 1 \\ w.x_i + b \leq 1, & y_i = -1 \end{cases}$$

Ce qui équivalents à :

$$y_i(w.x_i + b) \geq 1 pour\ i = 1 \dots m \quad (*)$$

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 équivalente à maximiser la somme des distances des deux classes par rapport à l'hyperplan. Ainsi, la marge a l'expression mathématique suivante :

$$M = \min_{x_i | y_i = 1} \frac{w.x + b}{||w||} - \max_{x_i | y_i = -1} \frac{w.x + b}{||w||} = \frac{1}{||w||} - \frac{-1}{||w||} = \frac{2}{||w||}$$

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

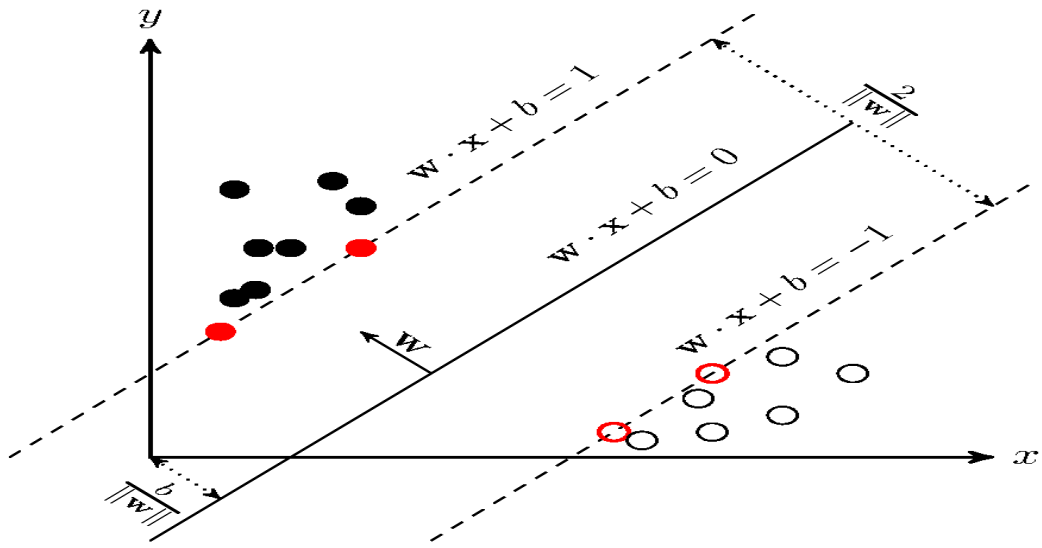



**Figure 17: Processus de dessiner l'hyperplan optimal**

**Principe de Fermat(1638) :**

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

**Principe de Lagrange(1788) :**

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

$$L(z,\alpha) = g(z) + \sum_{i=1}^{m} \alpha_i C_i^g(z)$$

Où $\alpha_i = \alpha_1, \alpha_2, \dots, \alpha_m$ sont des constants appelés coefficient de Lagrange.

**Principe de Kuhn-Tucker (1951) :**

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

$$\min_z L(z,\alpha^*) = L(z_0,\alpha) = \max_{\alpha \geq 0} L(z_0,\alpha)$$

Pour plus de détails sur ces trois principes, le lecteur peut consulter les références [42, 43].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 :

$$L(w,b,\alpha) = \frac{1}{2} w.w - \sum_{i=1}^{m} \alpha_i [y_i[(x_i.w) + b] - 1] \quad (1)$$

Le lagrangien doit être minimal par rapport à $w, b$ et maximale par rapport$\alpha \geq 0$.

$L(w, b, \alpha)$ est minimale par rapport à $b$:

$$\frac{\partial L(w, b, \alpha)}{\partial b} = 0 \Leftrightarrow \sum_{i=1}^{m} \alpha_i y_i = 0 \ (2)$$

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

$$\frac{\partial L(w, b, \alpha)}{\partial w} = 0 \Leftrightarrow w - \sum_{i=1}^{m} \alpha_i x_i y_i = 0 \ (3)$$

$L(w, b, \alpha)$ est maximale par rapport à $\alpha \geq 0$ :

En remplaçant $(2) et$ (3) dans le lagrangiens (1) :

$$L(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i . x_j) \ (4)$$

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

$$\sum_{i=1}^{m} \alpha_i y_i = 0; \alpha_i \geq 0$$

Soit la solution $\alpha^0 = (\alpha_1^0, \alpha_2^0, \dots \alpha_m^0)$,D'après le théorème de Kuhn-Tucker une condition nécessaire et suffisante pour que $\alpha^0$ soit optimal est :

$$\alpha_i^0 [y_i [(w_0 . x_0) + b_0] - 1] = 0 \, pour \, i = 1 \dots m$$

Ce qui veut dire que $\alpha_i^0$ où $y_i [(w_0 . x_0) + b_0] = 1$.

**Définition**

On définit les vecteurs de supports VS tout vecteur $x_i$tel que $y_i [(w_0 . x_0) + b_0] = 1$.Ce qui est équivalents à :

$$VS = \{x_i | \alpha_i > 0\} \, pour \, i = 1, \dots, m$$

Ainsi on peut facilement calculer $w_0, b_0$:

$$w_0 = \sum_{VS} \alpha_i^0 y_i x_i$$

$$b_0 = -\frac{1}{2} [(w_0 . x^*(1))] + [(w_0 . x^*(-1))]$$

La fonction de classement $class(x)$ est défini par :

$$class(x) = sign[(w_0 . x) + b_0] = sign[\sum_{x_i \in VS} \alpha_i y_i (x_i . x) + b_0]$$

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

## 1.2. Cas des données non-linéairement séparables

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

- 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 pénalités non-négative $\xi_i \; pour \; i = 1, \dots, m$ appelées variables d'écarts .Ces variables transforment l'inégalité (1) comme suit :

$$y_i(w.x_i + b) \geq 1 - \xi_i \; pour \; i = 1, \dots, m$$

L'objectif est de minimiser la fonction suivante :

$$\psi(w, \Xi) = \frac{1}{2} w.w + C \sum_{i=1}^{m} \xi_i$$

Où $C$ est un paramètre de régularisation. Elle 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 donné par l'équation(4) par rapport à $\alpha_i$ sous les contraintes suivantes (Kharroubi, 2002) :

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \; avec \; 0 \leq \alpha_i \leq C \; pour \; i = 1, \dots, m$$

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

# 2. Principe des SVM

Les classifieurs SVM utilise l'idée de l'HO (Hyperplan Optimal) pour calculer une frontière entre des nuages des points. Elles projettent les données dans l'espace de caractéristique en utilisant des fonctions non linéaires.

Dans cette espace on construit l'HO qui sépare les données transformées. L'idée principale est de construire une surface de séparation linéaire dans l'espace de caractéristiques qui correspond à une surface non linéaire dans l'espace d'entrées.

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 $D$, définit dans la section précédente, par la transformation$\psi$.

$$D' = \{(\psi(x_i), y_i) \in \mathbb{R}^p \times \{-1,1\} pour i = 1, \dots, m | p \geq n\}$$

En construisant un HO dans l'espace de caractéristique suivant la technique expliquée dans la section 1.On aura la fonction de classement suivante :

$$class(x) = sign[(w_0.x) + b_0] = sign[\sum_{x_i \in VS} \alpha_i y_i(\psi(x_i).\psi(x)) + b_0]$$

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 $\psi$ satisfasse la condition suivante :

$$\psi(u).\psi(v) = K(u,v)$$

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 :

**Théorème (Mercer) :**

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

$$K(u,v) = \sum_{k=1}^{+\infty} \beta_k \psi_k(u).\psi_k(v)$$

Tels que les $\beta_k > 0$(i.e $K(u,v)$ décrit un produit interne dans l'espace de caractéristique) il est nécessaire et suffisant que la condition suivante soit satisfaite

$$\iint K(u,v)g(u)g(v)dudv \geq 0$$

$pour\ toute\ fonction\ g \neq 0\ avec$ :

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

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) :

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

Où $\sigma$ à déterminer.

Maintenant que nous avons défini ce qu'est un noyau, la fonction de classement devient

$$class(x) = sign[\sum_{x_i \in VS} \alpha_i^0 y_i K(x_i, x) + b_0]$$

Reprenons l'exemple qu'on a évoqué au début de ce chapitre. Donc on a la transformation $\psi$ tel que :

$$\psi: \mathbb{R}^2 \to \mathbb{R}^3$$
$$X = (x_1, x_2) \to (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

D'où

$$\begin{aligned} K(U,V) &= \psi(U).\psi(V) \\ &= (u_1^2, \sqrt{2}u_1u_2, u_2^2).(\sqrt{2}v_1v_2) \\ &= u_1^2v_1^2 + 2u_1v_1u_2v_2 + u_2^2v_2^2 \\ &= (u_1v_1 + u_2v_2)^2 \\ &= \left[(u_1, u_1)\begin{pmatrix} v_1 \\ v_2 \end{pmatrix}\right]^2 = (U.V)^2 \end{aligned}$$

Comme on peut le constater le noyau correspondant à la transformation $\psi$ de notre exemple proposé au début de ce chapitre n'est autre qu'un noyau polynomiale de degré 2.

Dans la section suivante, nous présentons en détails le principe de noyau et nous allons discuter le noyau entre des vecteurs et le noyau entre des séquences.

# II. Noyau de vecteur et de séquence

## 1. Noyaux de vecteur

Dans ce chapitre, nous présentons le concept de noyaux, à la base d'un groupe de méthode appelées « méthodes à noyau » (kernel methods), parmi lesquelles on compte les SVMs. Nous donnons les éléments essentiels qui permettent de construire et choisir un noyau pour un problème de classification. Nous intéressons particulièrement à la classification de données numériques, c'est-à-dire de vecteurs et séquences de vecteurs.

### 1.1. L'astuce de Noyau

En pratique, l'astuce du noyau consiste à réécrire un algorithme où toutes les relations entre données d'entrée peuvent s'écrire sous forme de produits scalaires, en remplaçant ce produit scalaire par une fonction scalaire de deux variables (noyau). L'astuce du noyau permet ainsi de généraliser un algorithme linéaire manipulant des vecteurs :

- pour traiter les vecteurs de façon non linéaire (parce que les données présentent des non linéarités qu'il est utile d'exploiter pour le problème visé) ;
- pour manipuler d'autres types d'objets que les vecteurs.

Concernant le second point, l'astuce du noyau permet par exemple traiter des objets symboliques comme les chaines de caractères ou des objets structurés comme les graphes ou les automates. La figure suivante contient la liste les noms de noyaux couramment utilisés en anglais selon le type d'objets manipulés. Par exemple, pour les noyaux de séquences de données discrètes, il est de coutume de parler de « String Kernels», en bioinformatique, catégorisation de texte et reconnaissance de la parole.

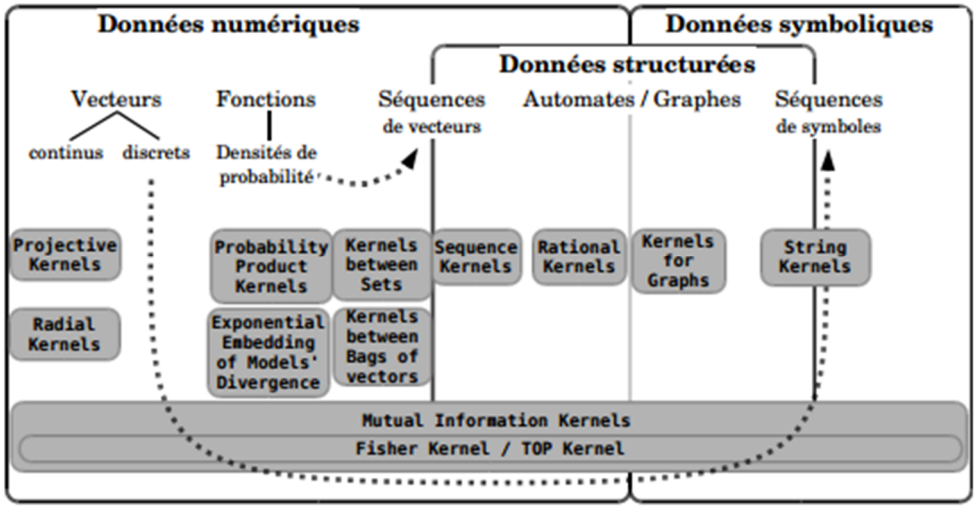



**Figure 18: Liste de noyaux, selon l'objet manipulés en entrée**

D'un point de vue qualitatif, le noyau peut être vu une mesure de similarité, qui permet de comparer deux objets d'un même type. Pour appliquer les méthodes à noyau sur un ensemble de données, il suffit en pratique de connaitre les valeurs de noyaux pour tous les couples de cet ensemble. Par exemple, pour dérouler l'algorithme d'apprentissage SVM, il suffit de connaitre les valeurs de noyaux estimées sur le corpus d'apprentissage, Ces valeurs sont habituellement mémorisées dans une matrice carrée : la « matrice de Gram ».

Définition (Matrice de Gram) :

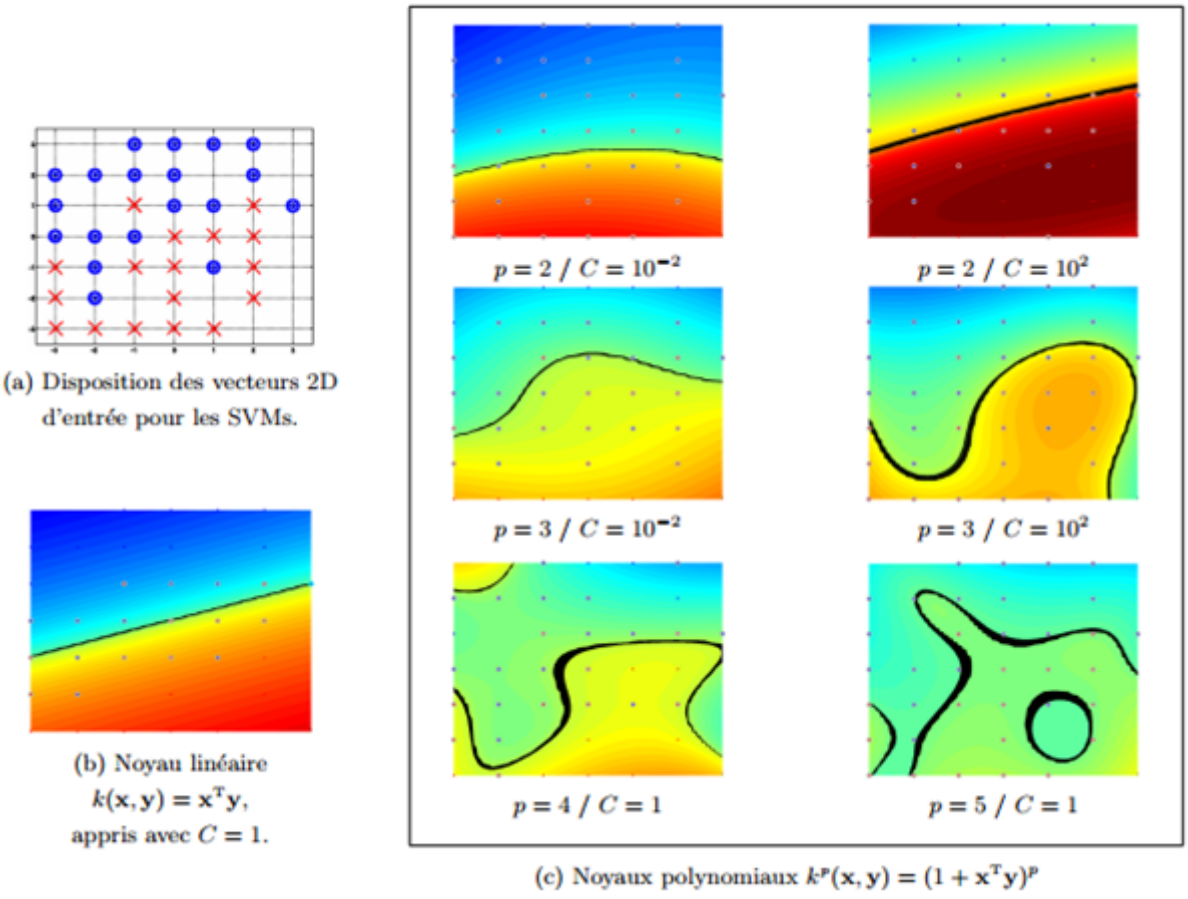
Soit une fonction noyau :$k: X^2 \rightarrow \mathbb{R}$ et un ensemble de donnée $\mathcal{A} = \{a_i\}_{i=1,\dots,N}$ de taille $N$.La matrice de Gram $K_{\mathcal{A}}$ est définie par la matrice carrée $N \times N$ contenant les valeurs du noyau sur les couples :$(K_{\mathcal{A}})_{i,j} = k(a_i, a_j)$

## 1.2. Le noyau entre des vecteurs (Le noyau projectif et radial)

Il existe deux types de noyau, le noyau projectif et radial(ou métrique), le noyau projectif se base sur le produit scalaire de deux vecteurs comme le noyau linéaire et polynomiale(les formules mathématiques de ces noyau sont appelés dans la section précédente). Le noyau radial s'appuie sur le produit scalaire par lui-même de la différence de deux vecteurs, soit la distance euclidienne au carrée ou norme $L_2$ au carrée. Parmi les noyaux radiaux existe, nous citons : Noyau Gaussien, Noyau de Cauchy, Noyau de la place et le noyau radial uniforme. Ces deux types de noyau sont des noyaux entre vecteur.

Dans les figures suivantes, nous illustrons les types de frontière que construisent les classifieurs SVM à noyau vectoriel sur un cas d'école bidimensionnel représenté en Fig.19(a). Les valeurs des fonctions discriminantes apprises par minimisation du risque régularisé sont représentées par les intensités de couleur. La frontière de décision est représentée par une courbe noire.

Fig.19 montre l'allure des frontières de décisions obtenues par apprentissage SVM avec les noyaux polynomiaux. On peut constater que même si augmenter le degré du polynôme permet de complexifier la modélisation, le comportement de la frontière devient très instable à partir d'un certain degré. Non seulement des effets de bord apparaissent, mais en plus les fonctions discriminantes atteignent des valeurs très faibles autour des vecteurs d'apprentissage. La prise de décision souffre de l'imprécision numérique apportée par ce second phénomène (mauvaise localisation de la frontière).



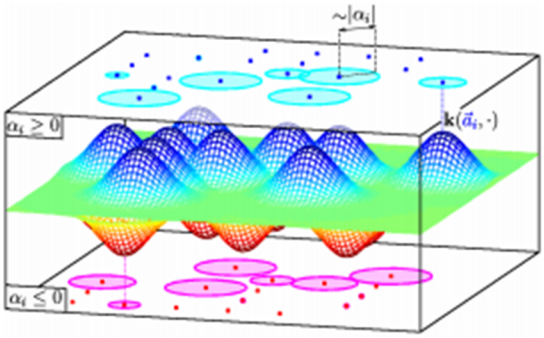

(a) Disposition des vecteurs 2D d'entrée pour les SVMs.

(b) Noyau linéaire $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$, appris avec $C = 1$.

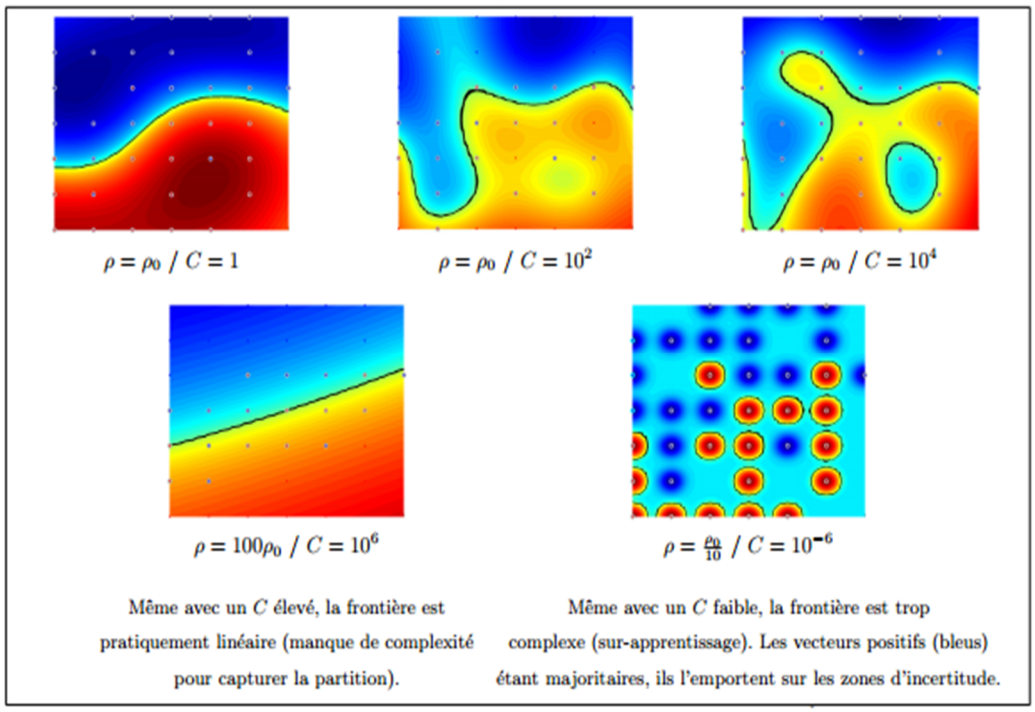
(c) Noyaux polynomiaux $k^p(\mathbf{x}, \mathbf{y}) = (1 + \mathbf{x}^T\mathbf{y})^p$

**Figure 19: Allure des modèles SVM vectoriel : noyaux projectifs**

Fig.20 représente l'allure des solutions d'un SVM obtenus avec les noyaux RBF Gaussiens. L'effet du sur-apprentissage par les noyaux radiaux lorsque la zone d'influence ρ est trop faible est visible dans Fig.20(b) (en bas à droite). Lorsqu'au contraire ρ est élevé, la frontière de décision devient quasiment linéaire (même figure, en bas à gauche). Fig.20(a) illustre comment la fonction discriminante est construite à partir du noyau et des vecteurs de support.




(a) : Fonction « représentante »$k(a_i, .)$ et illustration des poids de Lagrange $\beta_i$Associés aux vecteurs de supports, avec $\rho = \rho_0/C = 10^2$




(b)Modèle SVM appris avec un noyau Gaussien $k(x, y) = e^{-\frac{||x-y||^2}{2\rho^2}}$

**Figure 20: Allure des modèles SVM vectoriels: noyau Gaussiens**

## 2. Noyau de séquence

Pour appliquer SVM à une application de reconnaissance du locuteur ou de la langue, nous avons besoin d'une méthode de calcul des opérations du noyau sur les entrées de parole. Pour la reconnaissance de langue ou du locuteur, nous devons trouver un moyen qui nous permet de prendre une séquence de vecteurs caractéristique qui représente des vecteurs d'entées, ces vecteurs sont extraits à partir d'un énoncé$\{x_i\}$.

L'objectif d'utiliser la méthode de noyau de séquence est de trouver une fonction de comparaison de séquences de manière probabiliste, Alors nous avons besoin d'une fonction qui, étant donnée deux paramètres (notre cas : deux énoncé :$X_i \; et \; Y_i$), et elle produit une mesure de similarité entre ces deux paramètres. En générale le noyau de séquence permet de comparer tous les vecteurs caractéristiques d'une séquence avec tous les vecteurs caractéristique d'une autre séquence et produire une mesure de similarité. Parmi les noyaux utilisés, nous citons : le noyau GLDS et le noyau di ficher.

### 2.1. Le noyau GLDS (Generalized Linear Discriminant Scoring)

Le GLDS un des plus simple super vecteur SVM.GLDS est un noyau de séquence, utilisé pour des séquences de taille variable. Ce noyau est proposé par Campbell en 2002.La méthode GLDS crée des super vecteur en utilisant une expansion polynomial de vecteur. Cette expansion polynomiale est notée par $b(x)$, elle sert à former des monômes entre composante des vecteurs d'entrée, jusqu'à un degré $p$.

Nous prenons par exemple une expansion de degré 2, pour un vecteur $X = (x_1, x_2)$ de dimension 2,alors l'expansion polynomiale est définit par :

$$b(x) = \{x_1, x_2, x_1x_2, x_1^2, x_2^2\}$$

Pendant l'enregistrement,les locuteur de fond(background speakers),et les locuteurs cible (target speakers) $X = (x_1, x_2, x_3, \dots, x_T)$ sont représnteés comme la moyenne de vecteurs caractéristiques :

$$b_{avg} = \frac{1}{T}\sum_{t=1}^{T} b(x_t)$$

Ces vecteurs moyens sont des variances normalisées, ils sont affectés à l'étiquètes appropriés pour la phase d'entrainement de SVM (+1=Target speaker,-1=background speaker). L'optimisation de SVM (Campbell a utilisé un noyau linéaire standard) donne un ensemble de vecteurs de support $b_i$, le poid $\alpha_i$ et le biais $d$ :

$$w = \sum_{i=1}^{L} \alpha_i \, t_i b_i + d$$

Où $d = (d, 0,0,0, \dots ,0)^T$ ,$t_i \in \{-1, +1\}$ sont des sorties (étiquète de classe de vecteur de support) et $L$ est le nombre de support vecteur.

De cette manière nous pouvons le modèle locuteur par un unique super vecteur. Le vecteur de modèle projeté $w$ est normalisé en utilisant les énoncé de fond (background utterances), et il sert comme un modèle pour le locuteur cible (Target speaker).l'attribution de score(le match de score) est calculé comme un produit scalaire :

$$S = w_{target}^{t}.b_{test}$$

$w_{target}^{t}$: Vecteur de modèle normalisé du locuteur cible.

$b_{test}$ : Vecteurs Caractéristique moyens de l'énoncé test.

Le noyau GLDS a cependant une limitation pratique et théorique. La première est que l'utilisation d'expansions polynomiales au-delà d'un degré 3 n'est pas envisageable pour des problèmes de grande dimensionnalité. La seconde vient du fait qu'il n'est pas possible de généraliser l'approche en l'état à des expansions infinies. Mais le principal inconvénient est qu'il est difficile de contrôler la dimension de super vecteur. C'est pourquoi il est préférable d'utiliser une expansion polynomiale de degré 3.

Supposons que D est la dimension de l'expansion polynomiale, d est la dimension de vecteur d'entrées et p le degré de l'expansion polynomiale, Alors :

$$D = \frac{(d + p)!}{d!\,p!}$$

| **Le degré $p$** | **La dimension $d = 15$** |
|---|---|
| **2** | D=136 |
| **3** | 816 |
| **4** | 3876 |
| **5** | 15504 |

**Tableau 1: l'influence de degré p sur la dimension de l'expansion polynomiale**

La croissance exponentielle de cette dimension avec le degré explique pourquoi en pratique, le noyau GLDS n'est implémenté qu'avec $p = 3$. Ce manque de flexibilité est un frein à la fois pour les performances de la modélisation et pour l'utilisation du noyau dans d'autres contextes.

## 2.2. Le noyau de Fisher Discriminant (Kernel Fisher Discriminant (KFD))

Avant de présenter ce noyau, il est nécessaire de savoir quelques informations sur le noyau LDA (Linear Discriminant Analysis).Intuitivement, l'idée de LDA est de trouver une projection où la séparation des classes est Maximisée. Etant donné deux ensembles de données étiquetées,$C_1$ et $C_2$ et nous définissons les moyens de classe $m_1$ $et$ $m_2$ par :

$$m_i = \frac{1}{l_i}\sum_{n=1}^{l_i} x_n^i$$

Où $l_i$ est le nombre des exemples de la classe $C_i$. Le but de l'analyse discriminante linéaire est de donner une grande séparation de la classe des moyens tout en gardant la variance petite. Ceci est formulé par :

$$J(w) = \frac{w^T S_B w}{w^T S_w w}$$

Où $S_B$ est la matrice de between-class et $S_w$ est la matrice de within-class.

$$S_B = (m_1 - m_2)(m_1 - m_2)^T$$

$$S_w = \sum_{1,2}\sum_{n=1}^{l_i}(x_n^i - m_i)(x_n^i - m_i)^T$$

Mais afin de résoudre des problèmes non linéaires LDA n'est pas suffisant, alors afin de résoudre de problèmes non-linéaires, les données peuvent être mappées à un nouvel espace de caractéristique, F, via une fonction$\phi$ . Dans ce nouvel espace de caractéristique, la fonction qui doit être maximisé est :

$$J(w) = \frac{w^T S_B^{\phi} w}{w^T S_w^{\phi} w}$$

## Conclusion

Les machines à vecteur de support ou séparateur à vastes marges est un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Les SVM sont une généralisation des classifieurs linéaires.

Les SVM ont été développés dans les années 1990 à partir des considérations théoriques de Vladimir Vapnik sur le développement d'une théorie statistique de l'apprentissage : la théorie de Vapnik-Chervonenkis. Les SVM ont rapidement été adoptés pour leur capacité à travailler avec des données de grandes dimensions, le faible nombre d'hyper paramètres, leurs garanties théoriques, et leurs bons résultats en pratique.

Les SVM ont été appliqués à de très nombreux domaines (bio-informatique, recherche d'information, vision par ordinateur,...). Selon les données, la performance des machines à vecteurs de support est de même ordre, ou même supérieure, à celle d'un réseau de neurones ou d'un modèle de mélanges gaussiens.

# Chapitre IV

SVM pour l'identification du locuteur en mode indépendant du texte

# I. Historique

Depuis que les SVM ont vu le jour en 1995, plusieurs chercheurs du domaine de la reconnaissance de formes ont commencé à s'y intéresser, Principalement, en traitement d'image, on peut citer les brillants travaux réalisés sur la reconnaissance de lettres et de chiffres manuscrits et sur la détection du visage, ainsi que le travail de B. Schôlkopf en reconnaissance d'objet en 3D.

S. Benyacob s'est intéressé aux SVM pour faire la fusion des données de différents experts pour l'identification biométrique. Les résultats intéressants obtenus par ces applications ont incité les chercheurs d'autres disciplines comme la reconnaissance de locuteur à s'intéresser aux SVM. Sachant que les SVM exigent des vecteurs d'entrée de taille fixe, leur adaptation au RAL est moins évidente que dans le cas du traitement d'image. Si une image peut être facilement représentée par un vecteur fixe que ce soit en 2D et 3D, le signal de parole est difficilement représentable par un vecteur fixe puisqu'il est non délimité dans le temps.

La durée d'un signal de parole varie de quelques secondes à plusieurs minutes. Ainsi pour adapter les SVM à toute application utilisant le signal de parole, il faudrait trouver une nouvelle représentation de données qui permette de fournir un vecteur de taille fixe quelle que soit la longueur du signal de parole à traiter.

La première tentative d'application des SVMs en identification du locuteur a été effectué par M. Schmidt et H.Gish en 1996, Dans cette publication Schmidt a été utilisé directement les trames obtenues en phase de paramètrisation comme vecteurs d'entrées pour SVM, Mais cette méthode ne donne pas des bonnes résultats car il est connu que ces vecteurs contiennent simultanément un certaines nombre d'information sur le canal, la parole, les émotions, etc ,ce qui rend la tâche difficiles aux SVMs pour extraire uniquement des informations pertinentes des locuteurs directement de ces vecteurs sans passer par la modélisation.

Des figures 21 et 22 représentent le système proposé dans cette application dans les deux phases apprentissage et test. Dans le système proposé par M. Schmidt et H. Gish un modèle SVM est construit pour chaque client $\lambda$ de la base de données en utilisant toutes les trames calculées à partir des segments de signal de parole destinés pour l'apprentissage contre toutes les trames des segments de signal de parole d'apprentissage de tous les autres locuteurs de la base de données. Dans la phase de Test, et pour chaque client $\lambda$, un score $S_\lambda$ est calculé par l'équation suivante :

$$S_\lambda^X = \sum_{t \in X} f_\lambda(t)$$

Où $X$ ets le segment de test, t une trame de segment de test X,$\lambda$ est le client et $f_\lambda$ est le calssifieur SVM du client $\lambda$.Ainsi le locuteur qui maximisent les scores $S_\lambda^X$ est retenu comme bon locuteur.

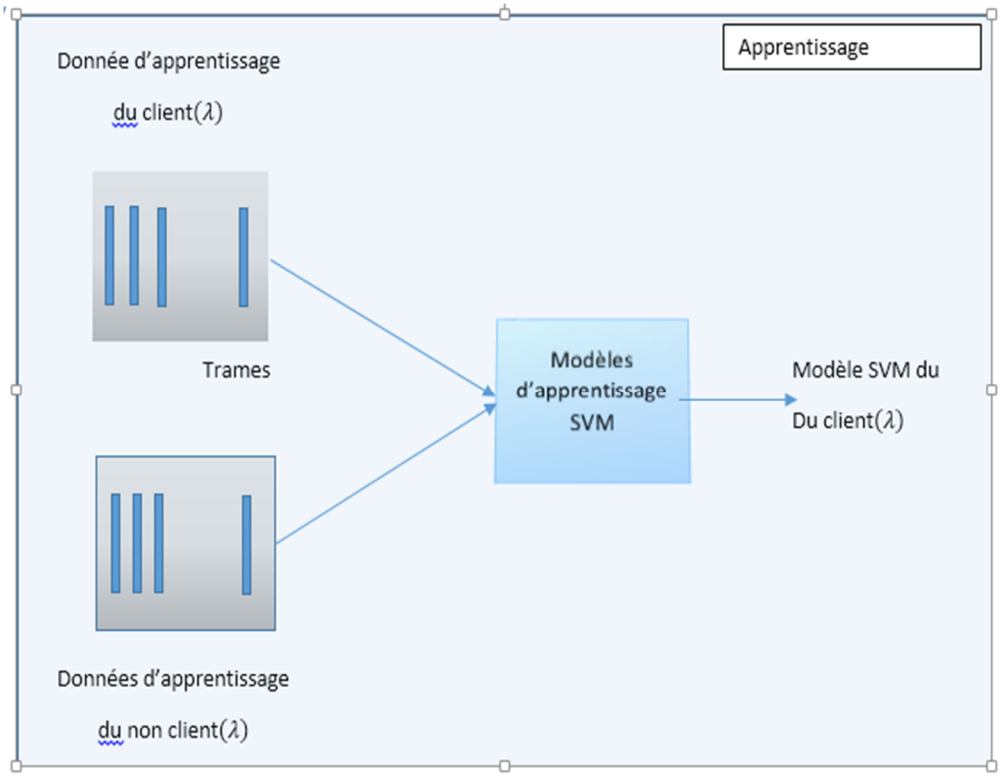



**Figure 21: Phase de l'apprentissage**

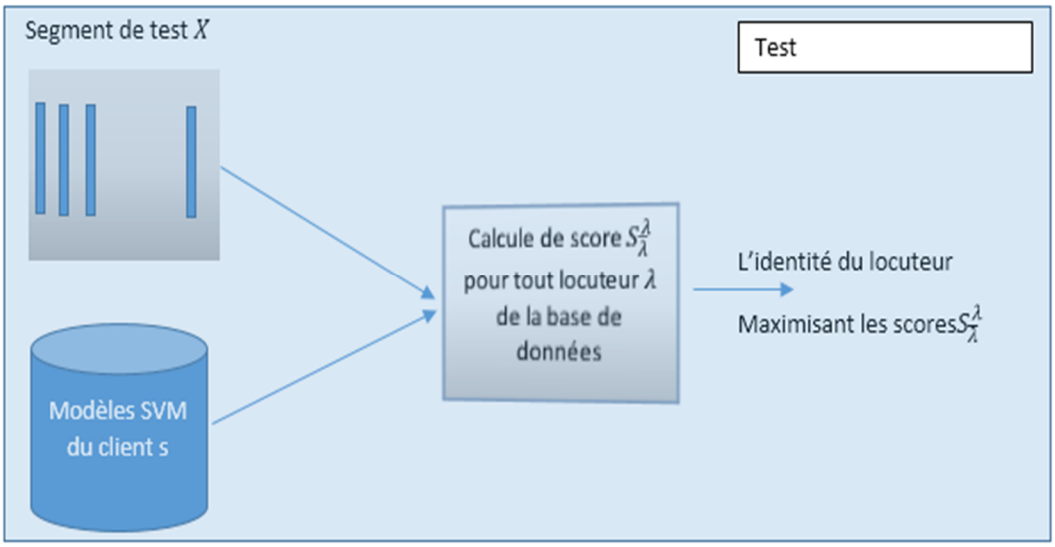



**Figure 22: Phase de Test**

Cette étude s'étale sur l'ensemble de 26 locuteurs (15 masculins, 11 féminins), chaque locuteur a parlé une durée de 10 secondes lors de la phase de l'entrainement .Pour la phase de Test, il existe 468 segments de test, ce qui montre que chaque locuteur prononce 18 phrases, Certains locuteurs sont utilisés dans la phase de test et de l'entrainement. Le Tableau suivant montre les résultats obtenus lors de cette expérience :

| Le degré de polynôme | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| **L'identification correcte** | 51% | 55% | 54% | 53% |

Si nous allons comparer ces résultats avec des études récents, nous trouvons que ces résultats ne sont pas bien et ça résulte grâce à l'utilisation directe de trames obtenues dans la phase de paramètrisation.

Suite à ce travail, d'autres groupes de recherche travaillant sur la reconnaissance de locuteur dont l'ENST fait partie se sont intéressés à ces techniques. L'équipe d'IBM a publié à Eurospeech2001 un travail intéressant sur l'adaptation des SVM en identification du locuteur.

Le système qu'ils ont proposé utilise les SVM comme système supplémentaire pour l'aide à la décision qui entre en action seulement quand le score obtenu par le système de base utilisant les GMM et LLR n'est pas fiable. Dans ce système, IBM a utilisé un nouveau noyau de séquence nommé le noyau de Fisher (chapitre 3 : Noyau de séquence).

# II. Approche hybride GMM-SVM

## 1. Description du système

Nous présentons dans cette section notre approche qui consiste à utiliser un système hybride GMM-SVM pour l'identification du locuteur en un milieu fermé en mode indépendant du texte. L'idée principale de cette approche est décrite dans la figure 23.

En premier lieu dans la phase de paramétrisation, nous avons cherché à identifier l'espace de caractéristique. Nous avons utilisé une base de donnée contient 40 locuteurs (26 masculins, 24féminins), l'espace de caractéristique présente une concaténation des coefficients MFCC.

En deuxième lieu, lors de la phase de modélisation, nous avons cherché à discriminer les différentes classes des locuteurs, nous utilisons le modèle de mélange Gaussienne qui nous permet de créer les supers vecteur qui sont utilisés comme des entrées pour le classifeur SVM.

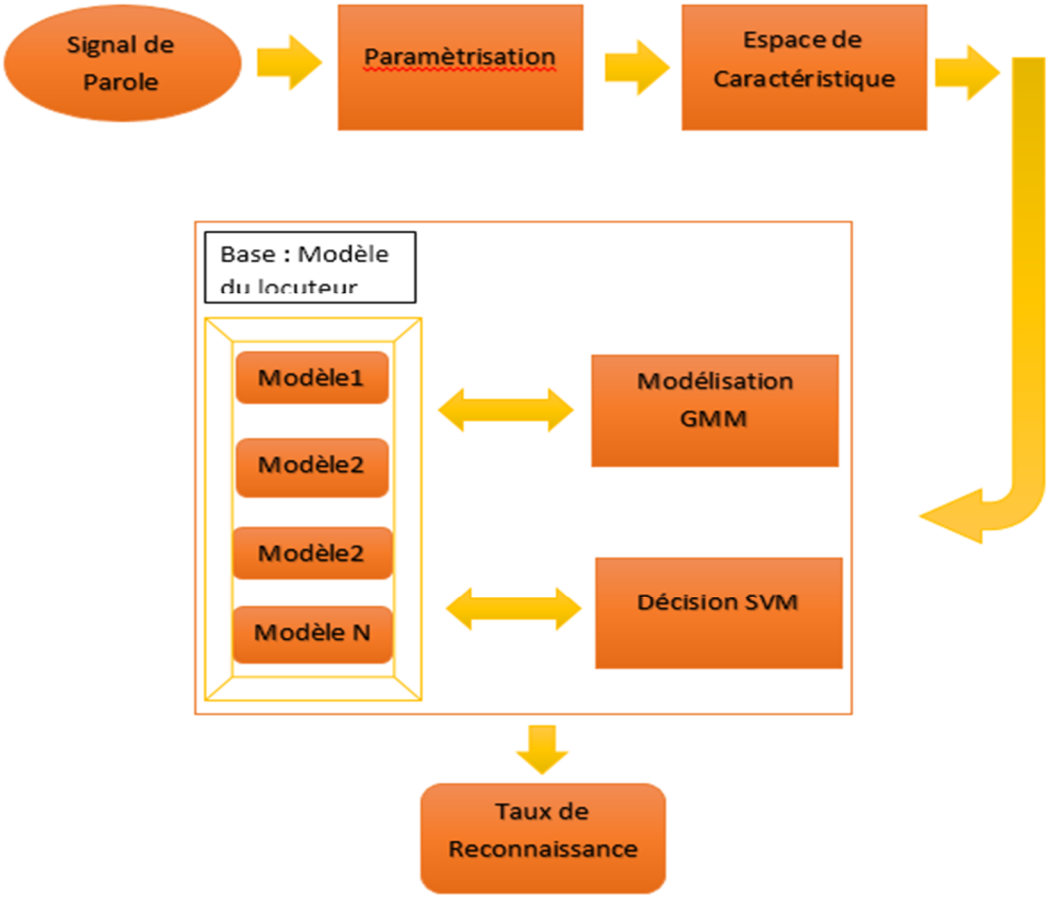



**Figure 23: Système hybride GMM-SVM pour identification du locuteur**

La modélisation GMM consiste à représenter chaque client par un gaussien, ce dernier est caractérisé par son poids, un vecteur moyen $\mu$ de dimension $d$ et une matrice de variance $\Sigma$ de dimension$d \times d$.

L'apprentissage des GMM, caractérisé par l'apprentissage des paramètres des modèles du locuteurs $(p_i, \mu_i, \Sigma_i)$, s'est fait par l'algorithme EM (Expectation Maximisation) qui assure la convergence vers une solution optimale. L'algorithme EM permet de régler les paramètres d'un modèle de distribution GMM pour atteindre un maximum de vraisemblance. Après cette tâche, des paramètres résultants sont appelés des super vecteur, la figure suivante montre ce processus

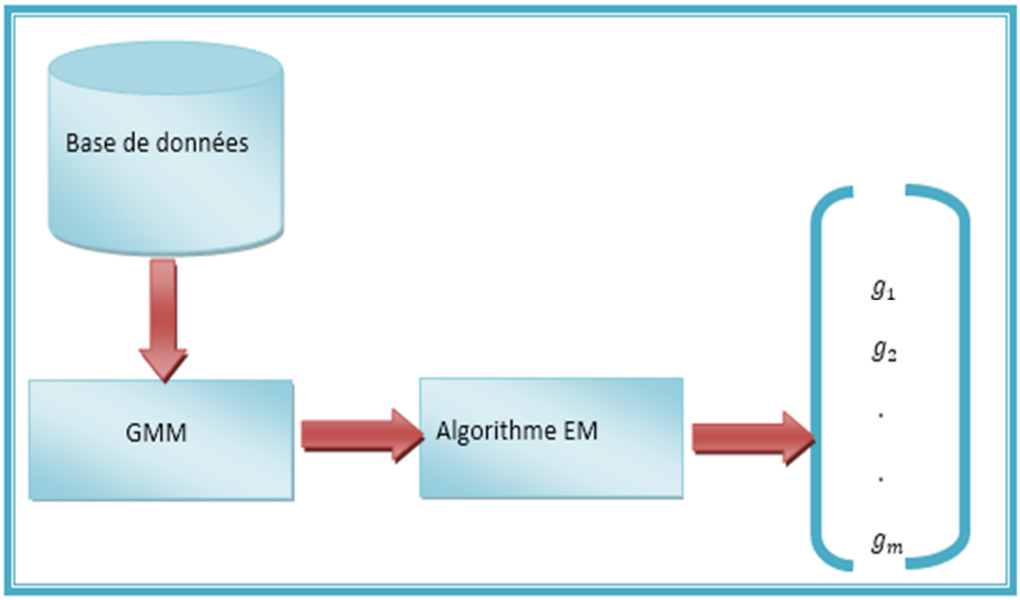



**Figure 24: Notre processus pour générer des supers vecteurs**

Chaque $g_i$ est caractérisé par le triple$(p_i, \mu_i, \Sigma_i)$, pour les entrées, nous utilisons juste les vecteurs moyens, comme il a illustré dans le chapitre 2, la figure 13.l'utilisation du vecteurs moyens comme entrées au SVM est proposés par Campbell en 2002.

En général, nous considérons qu'un super vecteur est défini comme la concaténation des composantes des moyennes des gaussiennes du GMM. Maintenant et après la création des super vecteurs, il faut utiliser ces dernier pour créer un modèle SVM pour chaque locuteur, comme il est illustré dans la figure 21.

En troisième lieu, lors de la phase de Test, nous utilisons le même démarche apparait dans la figure 22, pour chaque client $\lambda$, un score $S_\lambda$ est calculé par l'équation suivante :

$$S_\lambda^X = \sum_{t \in X} f_\lambda(t)$$

Le locuteur qui maximise les scores est retenu comme bon locuteur.

# 2. Protocole expérimentale

## 2.1. Base de données

La base de données utilisée contient 40 locuteurs, 23 masculins, 17 féminins, l'âge de ces locuteurs est compris entre 18 et 30 ans. Chaque locuteur a enregistré au moins de deux sessions séparées par un décalage de 2 à 3 semaines. Les phrases prononcées sont en arabe, français et anglais. Les phrases sont indépendantes pour tous les locuteurs.

Les enregistrements audio ont été recueillis des locuteurs à partir d'Internet entant que messages vocaux via Skype, qui est considéré comme le plus grand fournisseur de voix sur Internet. Afin de couvrir une large gamme d'environnements acoustiques réels, ils ont recommandé aux locuteurs de faire des appels à partir de nombreux endroits différents, par exemple, maison, bureau, etc. En outre, différents types de matériel d'enregistrement ont été utilisés pour le codage (ordinateurs portables, tablettes et smartphones ...). Les messages vocaux sont numérisés à 16 kHz avec une résolution de 16 bits et stockés en format wav, qui est considéré comme le type de fichier le plus couramment utilisé.

### 2.2. Paramétrisation

En phase de paramétrisation, nous avons spécifié l'espace de caractéristique utilisé. Cet espace est définit par des vecteurs de tailles fixes. En effet, comme le signal audio est dynamique et variable, nous avons présenté les séquences d'observation de tailles variables par des vecteurs de taille fixe. Chaque vecteur est donné par la concaténation des coefficients mel cepstre MFCC (19 coefficients).

### 2.3. Modélisation GMM

En la phase de modélisation, nous avons déterminés des nouvelles représentations discriminant les différentes classes des locuteurs de l'espace de caractéristiques. Ces nouvelles représentations sont basées sur une modélisation GMM des locuteurs portée sur l'algorithme de type expectation maximisation(EM) estime d'une manière itérative les paramètres de chaque modèle au sens du maximum de vraisemblance.il assigne une distribution de probabilité pour chaque enregistrement et indique sa probabilité d'appartenance à chaque modèle. 40 modèles sont générés permettant des spécifiés les 40 classes du locuteur. Les modèles générés caractérisent les vecteurs d'entrées pour la phase de décision.

### 2.4. Décision

En phase de décision, nous avons utilisé les représentations des 40 modèles déterminés par la phase de modélisation. Ces modèles présentent les vecteurs d'entrée pour le support vecteur machine. Pour nos expériences, nous avons varié le noyau de notre système hybride GMM-SVM : noyau linéaire, noyau polynomiale de degré 3 et noyau RBF avec $\sigma = 0.1$, et aussi nous avons varié le nombre de gaussien pour voir l'influence de nombre de gaussien sur notre système hybride GMM-SVM.

## 3. Résultats et Evaluation

Nous avons testé différents types de noyaux SVM tout en variant le nombre de gaussiennes afin de visualiser l'influence de l'ordre des GMM sur les performances du système.

. Pour la phase de test, nous utilisons 15 locuteurs, chaque locuteur a prononcé 6 phrases. Les tableaux 2, 3,4 montrent l'effet de l'augmentation du nombre de gaussiennes sur le taux de reconnaissance de notre système hybride dans le cas d'un noyau linéaire, polynomial et gaussien.

Pour les tableaux suivants : G désigne le nombre de gaussien et T.R le taux de reconnaissance.

| G | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |
|---|---|---|---|---|---|---|---|---|
| T.R | 29.82% | 32.71% | 30.45% | 31.03% | 31.60% | 34.46% | 34.91% | 34.94% |

Tableau 2: Noyau linéaire

| G | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |
|---|---|---|---|---|---|---|---|---|
| T.R | 26.14% | 29.13% | 30.79% | 32.84% | 34.13% | 35.38% | 36.53% | 38.47% |

Tableau 3: Noyau polynomial d=3

| G | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |
|---|---|---|---|---|---|---|---|---|
| T.R | 18.71% | 18.65% | 29.86% | 18.99% | 30.51% | 31.32% | 39.13% | 40.9386% |

Tableau 4: Noyau gaussien $\sigma = 0.1$

Afin d'étudier la similarité entres les 40 modèles des locuteurs, nous avons varié le type de noyau du modèle SVM et le nombre de gaussien utilisé pour la modélisation GMM. Nous avons choisi de présenter les résultats donnés par le système de noyau linéaire (tableau 1), polynomial de degré 3(2) et RBF (tableau 3).L'influence de nombre de gaussiens sur le taux de reconnaissance est évident pour tous les noyaux. Le meilleur résultat est obtenu par le noyau RBF pour G=1024.

## Conclusion

Dans ce chapitre, nous avons présenté le système hybride GMM-SVM pour l'identification du locuteur en mode indépendant du texte. Ce système présente la capacité de modélisation des GMM basé sur l'algorithme EM et l'efficacité de décision des supports vecteurs machine. Dans ce travail, nous avons passé essentiellement par trois phases : la paramétrisation, la modélisation et la décision. Dans la phase de paramétrisation, nous avons préparé notre espace de caractéristiques. Cet espace est donné par la concaténation des paramètres MFCC. Pour la phase de Modélisation, nous avons cherché à discriminer l'espace caractéristique en utilisant la modélisation GMM de type EM pour générer des super vecteurs qui sont utilisés comme des entrées pour la machine SVM pour la phase de décision.

# Conclusion et Perspectives

Ce travail s'inscrit dans le domaine de l'identification du locuteur en mode indépendant de texte. Un système d'identification du locuteur consiste à déterminer, à partir d'un ensemble de locuteurs référencés dans le système, L'identité du locuteur présente dans un signal vocal. Pour cela le système calcule des mesures de similarité entre ce signal et tous les modèles des locuteurs de la base. Deux types d'identification sont connus : milieu fermé, milieu ouvert. En milieu fermé, chaque accès de test est comparé à tous les modèles des locuteurs référencés dans le système. Mais pour le milieu ouvert, le locuteur ne possède pas un modèle de référence, c'est-à-dire le locuteur n'appartient pas à la base de référence.

L'objectif de ce travail est d'introduire le modèle discriminant SVM 'support Vector Machine' dans le système IAL (Identification Automatique du locuteur).Mais l'utilisation de ce modèle pour l'IAL est très difficile car les SVMs ont besoin en entrés des vecteurs de tailles fixe.La communauté de reconnaissance du locuteur a récemment redécouvert un moyen robuste pour présenter des énoncés en utilisant un seul vecteur, dite super vecteur. D'une part, ces supers vecteurs peuvent être utilisés comme entrées pour soutenir la machine de vecteur (SVM).La question reste, comment nous pouvons créer ces supervecteurs ?

Il existe deux méthodes pour créer ces supervecteurs : soit en utilisant le noyau de séquence soit en utilisant le modèle de mélange gaussien(GMM).Dans le cadre de ce projet nous avons décrit notre approche qui consiste à utiliser le système hybride GMM-SVM pour l'identification du locuteur en mode indépendant du texte, l'idée principale de ce système est de modéliser des locuteur par la méthode de modélisation GMM et dans la phase de décision exploite les représentations données par la modélisation GMM .

Pour des résultats obtenus par ce système dans cette étude, nous constatons que les résultats ne sont pas satisfaisant, alors comme perspectives, nous proposons : d'améliorer ces résultats par l'inclusion d'autres algorithmes, d'appliquer ce système sur un système IAL en milieu ouvert, et ensuite nous créons une application d'authentification en utilisant ce système hybride.

# Référence


1. Preti, A., *Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur*, in *ACADÉMIE D'AIX-MARSEILLE UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE*. 10 décembre 2008, Université d'Avignon: France. p. 197.

2. Boujelbene, S.Z., D. Mezghani, and N. Ellouze, *Improved Feature Data for Robust Speaker Identification Using Hybrid Gaussian Mixture Models-Sequential Minimal Optimization System.* International Review on Computers & Software, 2009. **4**(3).

3. Van Vuuren, S. *Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch*. in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. 1996. IEEE.

4. BOUJELBENE, S.Z., D.B.A. MEZGHANI, and N. ELLOUZE, *Identification du Locuteur par Système Hybride GMM-SMO,5th International Conference: Sciences of Electronic,Technologies of Information and Telecommunications,tunisia*2009

5. Jourani, R., *Reconnaissance automatique du locuteur par des GMM à grande marge*. 2012, Université de Toulouse, Université Toulouse III-Paul Sabatier; Université Mohammad V-Agdal de Rabat.

6. Furui, S., *Cepstral analysis technique for automatic speaker verification.* Acoustics, Speech and Signal Processing, IEEE Transactions on, 1981. **29**(2): p. 254-272.

7. Reynolds, D.A., *Speaker identification and verification using Gaussian mixture speaker models.* Speech communication, 1995. **17**(1): p. 91-108.

8. Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech.* the Journal of the Acoustical Society of America, 1990. **87**(4): p. 1738-1752.

9. Campbell, J.P., D.A. Reynolds, and R.B. Dunn. *Fusing high-and low-level features for speaker recognition*. in *INTERSPEECH*. 2003.

10. Fredj, I.B., K. Ouni, and E.S. de Technologie, *Optimisation de la paramétrisation pour la reconnaissance floue des signaux de parole Optimization of features parameters for fuzzy speech recognition.*

11. Fredouille, C., *Reconnaissance du locuteur et approche statistique: Information dynamiques et normalisation bayésienne des vraisemblances*. 2000, Thèse de l'Université d'Avignon.

12. Harb, H., *Classification du signal sonore en vue d'une indexation par le contenu des documents multimédias.* Manuscrit de thèse, thèse sous la supervision de Prof. Liming Chen, Lab. LIRIS, Ecole Centrale de Lyon. Αναφοράστηνεργασία [40], 2003.

13. Arcienega, M. and A. Drygajlo. *A Bayesian network approach for combining pitch and reliable spectral envelope features for robust speaker verification*. in *Audio-and Video-Based Biometric Person Authentication*. 2003. Springer.

14. Mami, Y., *Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence*. 2003, Télécom ParisTech.

15. Booth, I., M. Barlow, and B. Watson, *Enhancements to DTW and VQ decision algorithms for speaker recognition.*Speech Communication, 1993. **13**(3): p. 427-433.

16. SAKKA, Z., et al., *RECONNAISSANCE DU LOCUTEUR PAR LA TECHNIQUE DE LA QUANTIFICATION VECTORIELLE.*

17. Matsui, T. and S. Furui, *Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMM's.* Speech and Audio Processing, IEEE Transactions on, 1994. **2**(3): p. 456-459.

18. Yu, K., J. Mason, and J. Oglesby, *Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation.* IEE Proceedings-Vision, Image and Signal Processing, 1995. **142**(5): p. 313-318.

19. Rabiner, L.R. and B.-H. Juang, *Fundamentals of speech recognition*. Vol. 14. 1993: PTR Prentice Hall Englewood Cliffs.

20. Savic, M. and S.K. Gupta. *Variable parameter speaker verification system based on hidden Markov modeling*. in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*. 1990. IEEE.

21. Rosenberg, A.E., C.-H. Lee, and S. Gokcen. *Connected word talker verification using whole word hidden Markov models*. in *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*. 1991. IEEE.

22. Webb, J. and E. Rissanen. *Speaker identification experiments using HMMs*. in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*. 1993. IEEE.

23. Oglesby, J. and J. Mason. *Speaker recognition with a neural classifier*. in *Artificial Neural Networks, 1989., First IEE International Conference on (Conf. Publ. No. 313)*. 1989. IET.

24. Artières, T. and P. Gallinari, *Approches prédictives neuronales pour l'identification.* XXème Journées d'Etudes sur la parole (JEP), Trégastel, France, 1994: p. 275-280.

25. Homayounpour, M.M. and G. Chollet. *Neural net approaches to speaker verification: comparison with second order statistic measures*. in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*. 1995. IEEE.

26. Kuhn, R., et al., *Rapid speaker adaptation in eigenvoice space.* Speech and Audio Processing, IEEE Transactions on, 2000. **8**(6): p. 695-707.

27. Larcher, A. *Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée*. 2009. Avignon.

28. Oglesby, J., *What's in a number? Moving beyond the equal error rate.* Speech communication, 1995. **17**(1): p. 193-208.

29. Martin, A., et al., *The DET curve in assessment of detection task performance*. 1997, DTIC Document.

30. Bimbot, F., I. Magrin-Chagnolleau, and L. Mathan, *Second-order statistical measures for text-independent speaker identification.* Speech communication, 1995. **17**(1): p. 177-192.

31. Reynolds, D.A. and R.C. Rose, *Robust text-independent speaker identification using Gaussian mixture speaker models.* Speech and Audio Processing, IEEE Transactions on, 1995. **3**(1): p. 72-83.

32. Reynolds, D.A., T.F. Quatieri, and R.B. Dunn, *Speaker verification using adapted Gaussian mixture models.* Digital signal processing, 2000. **10**(1): p. 19-41.

33. Bimbot, F., et al., *A tutorial on text-independent speaker verification.*EURASIP journal on applied signal processing, 2004. **2004**: p. 430-451.

34. Bousquet, P.-M. *Bénéfices et limites des représentations en facteur de variabilité totale pour la reconnaissance du locuteur*. 2014. Avignon.

35. Higgins, A., L. Bahler, and J. Porter, *Speaker verification using randomized phrase prompting.* Digital Signal Processing, 1991. **1**(2): p. 89-106.

36. Rosenberg, A.E., et al. *The use of cohort normalized scores for speaker verification*. in *Second international conference on spoken language processing*. 1992.

37. Scheffer, N. and J.-F. Bonastre. *Ubm-gmm driven discriminative approach for speaker verification*. in *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*. 2006. IEEE.

38. Marcha, A., *La production de la parole.*2007.

39. Gauvain, J.-L. and C.-H. Lee, *Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains.*Speech and audio processing, ieee transactions on, 1994. **2**(2): p. 291-298.

40. Kharroubi, J., *Etude de techniques de classement" Machines à vecteurs supports" pour la vérification automatique du locuteur*. 2002, Télécom ParisTech.

41. Schölkopf, B., C.J. Burges, and A.J. Smola, *Advances in kernel methods: support vector learning*. 1999: MIT press.

42. Ciarlet, P.G., *Introduction à l'analyse numérique matricielle et à l'optimisation*. 1988.

43. Fletcher, R., *Practical methods of optimization*. 2013: John Wiley & Sons.