REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Université de Mohamed El-Bachir El-Ibrahimi - Bordj Bou Arreridj

Faculté *des Sciences et de la technologie*

Département d'Electronique

Mémoire

Présenté pour obtenir

LE DIPLOME DE MASTER

FILIERE : **Télécommunications.**

Spécialité : Systèmes des télécommunications.

Par

- **Chadi Salim**
- **Halis Aymen**

Intitulé

Vérification automatique du locuteur par GMM/SVM

Évalué le : ……

Par la commission d'évaluation composée de* :

Nom & Prénom	*Grade*	*Qualité*	*Etablissement*
M.S.MEZAACHE	*…*	*Président*	*Univ-BBA*
M.ASBAI NASSIM	*…*	*Encadreur*	*Univ-BBA*
M.S.AIDEL	*….*	*Examinateur*	*Univ-BBA*

Année Universitaire 2022/2023

Dédicaces

Mes parents, qui m'ont encouragé à aller de l'avant et qui m'ont donné tout leur amour
pour prendre mes études. Aux quels je dois ce que je suis. Que dieu les protège.

A mes frères HOUSSAM et NABIL et mes sœurs et toute ma famille

A mes amis HALIS AYMEN et ROUAG KHALED DJIHED

Et mes amis en générale

CHADI SALIM

Dédicaces

A mes très chers parents qui ont toujours été là pour moi, et qui m'ont donné un magnifique modèle de labeur et de persévérance.
J'espère qu'ils trouveront dans ce travail toute ma reconnaissance et tout mon amour.
A mon frère HICHEM et ma sœur ANHAR et a toute ma famille
Mes amis CHADI SALIM et ROUAG KHALED DJIHED et DALI AYOUB et ZATCHI ABD ERRAZAK

HALIS AYMEN

Remerciements

En premier lieu, nous tenons à remercier Dieu tout puissant qui .m'a donné la force de continuer et .D'achever ce travail
Nous remercions fortement notre encadreur M. Assbai Nassim de .nous avoir orienté par ses conseils judicieux dans le but de mener à .bien ce travail
Nos remerciements et notre respect aux membres du jury qui nous feront l'honneur d'évaluer notre .Travail

Résumé :

Cette étude porte sur la vérification automatique du locuteur en utilisant une approche hybride combinant le modèle de mélange de Gaussiennes (GMM) et les machines à vecteurs de support (SVM). L'objectif de cette combinaison est de déterminer l'identité d'un locuteur en se basant sur les caractéristiques acoustiques de sa voix.

En utilisant à la fois le GMM et les SVM, nous avons pu tirer parti des avantages de chaque méthode. Le GMM a permis de modéliser les variations individuelles de la voix des locuteurs, tandis que les SVM ont été utilisées pour la classification et la reconnaissance des locuteurs en exploitant la séparation optimale des classes.

Dans notre travail, nous avons pris en compte différents paramètres pour améliorer les performances du système. Parmi ces paramètres, on trouve le nombre de MFCCs, les paramètres prosodiques, le type de SVM et les composantes du modèle UBM, jouent un rôle crucial dans l'amélioration des performances du système.

Les résultats de notre étude ont démontré que l'approche hybride basée sur le GMM et les SVM est efficace pour la vérification automatique du locuteur. Les résultats obtenus confirment l'importance de combiner les modèles de mélange de Gaussiennes et les machines à vecteurs de support, ainsi que l'utilisation du modèle de fond universel pour améliorer la précision et la robustesse du système.

Abstract :

This study focuses on the automatic verification of the speaker using a hybrid approach combining the Gaussian mixture model (GMM) and support vector machines (SVM). The objective of this combination is to determine the identity of a speaker based on the acoustic characteristics of his voice.

By using both GMM and SVM, we were able to leverage the benefits of each method. The GMM was used to model individual variations in speaker voice, while the SVM was used for speaker classification and recognition by exploiting optimal class separation.

In our work, we have taken into account various parameters to improve system performance. These parameters, including the number of MFCCs, the prosodic parameters, the type of SVM, and the components of the UBM model, play a crucial role in improving system performance.

The results of our study demonstrated that the hybrid approach based on GMM and SVM is effective for automatic verification of the speaker. The results obtained confirm the importance of combining Gaussian mixing models and support vector machines, as well as the use of the universal bottom model to improve the accuracy and robustness of the system.

ملخص :

تركز هذه الدراسة على التحقق التلقائي من السماعات باستخدام نهج هجين يجمع بين نموذج خليط غاوسي (GMM) وآلات ناقلات الدعم (SVM). الهدف من هذا التخصص هو تحديد هوية المتحدث بناءً على الخصائص الصوتية لصوته.

باستخدام كل من GMM و SVMs ، تمكنا من الاستفادة من فوائد كل طريقة. تم استخدام GMM لنمذجة الاختلافات الفردية لأصوات المتحدثين ، بينما تم استخدام SVM لتصنيف السماعات والتعرف عليها من خلال استغلال الفصل الأمثل للفئة.

لقد أخذنا في الاعتبار في عملنا معايير مختلفة لتحسين أداء النظام. من بين هذه المعلمات ، نجد أن عدد MFCCs ، والمعلمات العامة ، ونوع SVM ومكونات نموذج UBM ، تلعب دورًا مهمًا في تحسين أداء النظام.

أظهرت نتائج دراستنا أن النهج الهجين القائم على GMM و SVM فعال للتحقق التلقائي من السماعات. تؤكد النتائج التي تم الحصول عليها على أهمية الجمع بين نماذج الخليط الغاوسي وآلات ناقلات الدعم ، وكذلك استخدام نموذج الخلفية العالمية لتحسين دقة ومتانة النظام.

Table des matières

Liste des figures

Liste des tableaux

Abréviations

Liste des figures

Liste des tableaux

Abbreviations

DET	Detection Error Tradeoff
DFT	transformée de Fourier discrète
EER	Equal Error Rate
FA	Fausse Acceptation.
FR	Faux Rejet
GMM	Gaussian Mixture Models
IAL	Identification Automatique du Locuteur
MFCC	Mel Frequency Cepstral Coefficients
RAL	Reconissance Automatique du locuteur
STFT	La Transformée de Fourier à court terme
SVM	Support Vector Machines
UBM	Universal Background Model
VAD	Détection de l'Activité Vocale
VAL	Vérification Automatique du Locuteur

Introduction général

Introduction générale

La parole est un moyen essentiel de communication utilisé par les êtres humains. Comprendre les mécanismes de production et d'audition de la parole est crucial pour le développement de techniques de traitement automatique de la parole. L'approche acoustique est une approche couramment utilisée dans l'analyse et le traitement de la parole, qui se concentre sur les propriétés acoustiques du signal vocal.[9]

L'identification et la vérification automatique du locuteur sont des tâches essentielles dans le domaine du traitement automatique de la parole. Ces techniques visent à déterminer l'identité d'un locuteur en analysant les caractéristiques acoustiques de sa voix. La vérification du locuteur trouve de nombreuses applications pratiques dans des domaines tels que la sécurité, les services bancaires et d'autres applications nécessitant une identification précise du locuteur.

l'apprentissage des données du locuteur commence par l'extraction des MFCC à partir des enregistrements vocaux. Les MFCC sont des coefficients qui représentent les caractéristiques acoustiques de la voix humaine, en se basant sur l'échelle de fréquence Mel pour mieux représenter la perception auditive humaine.

Parmi les approches d'apprentissage des données acoustiques, les approches basées sur les Mélanges de Gaussiennes (GMM) sont couramment utilisées. Les GMM permettent de modéliser les distributions de probabilité des caractéristiques acoustiques des locuteurs. Le modèle du mélange est utilisé pour représenter statistiquement les variations des caractéristiques acoustiques propres à chaque locuteur. L'adaptation du modèle par Maximum A Posteriori (MAP) permet d'améliorer la précision et la robustesse des modèles GMM.

Le système de vérification automatique du locuteur (VAL) est conçu en utilisant des techniques basées sur les modèles de mélange gaussien (GMM) et les machines à vecteurs de support (SVM). Ces approches sont largement utilisées dans le domaine de la reconnaissance de locuteurs et de la biométrie. [1]

Le modèle de mélange gaussien (GMM) est utilisé pour modéliser les distributions de probabilité des caractéristiques vocales extraites de la voix d'un locuteur. Il permet de représenter statistiquement les variations des caractéristiques acoustiques propres à chaque locuteur. D'autre part, les machines à vecteurs de support (SVM) sont utilisées pour séparer et

classifier les caractéristiques vocales des différents locuteurs. Cette combinaison de GMM et de SVM offre une approche robuste et précise pour la vérification automatique du locuteur.[2] les modèles de mélange gaussien (GMM) sont utilisés pour modéliser les distributions statistiques des MFCC pour chaque locuteur. Les GMM sont des modèles probabilistes qui représentent la probabilité de chaque MFCC appartenant à chaque locuteur. Ces modèles sont entraînés à l'aide d'un ensemble de données d'entraînement qui contient des enregistrements vocaux de différents locuteurs.
Une fois que les modèles GMM sont entraînés, les machines à vecteurs de support (SVM) sont utilisées pour la classification des locuteurs. Les SVM sont des algorithmes d'apprentissage supervisé qui construisent des frontières de décision pour séparer les différentes classes de locuteurs. Les MFCC des enregistrements vocaux inconnus sont comparés aux modèles GMM et utilisés comme entrées pour les SVM, qui déterminent la classe de locuteur correspondante.

Cet mémoire est organisée comme suit :

Dans le premier chapitre, nous décrirons les concepts de base du signal de parole ainsi que certaines méthodes de Traitement et analyse de la parole pour la vérification du locuteur, tandis que dans le deuxième chapitre, nous expliquerons en détail l'apprentissage et la classification des donnes du locuteur par GMM/SVM.

Dans le troisième chapitre, nous présenterons une simulation pratique d'un système de Vérification automatique du locuteur par GMM/SVM en utilisant les métriques : MFCC , les paramètres prosodique, le noyau , gamma et c et le nombre des mixtures et locuteurs.

Chapitre1

Traitement et analyse de la parole pour la vérification du locuteur

1.1 Introduction

La vérification du locuteur est une tâche de traitement automatique de la parole qui vise à vérifier l'identité d'un locuteur en analysant les caractéristiques acoustiques de sa voix. Cette tâche est largement utilisée dans les systèmes de sécurité, les services bancaires et autres applications nécessitant une identification précise du locuteur**.**

Dans ce chapitre, nous présentons les mécanismes de production de la parole et l'analyse spectrale et temporelle de la parole utilisés pour concevoir un système de vérification du locuteur (VAL) .

1.2 Mécanismes de production et de l'audition de la parole

1.2.1 Production de la parole

La production de la parole est un processus complexe qui implique la coordination de plusieurs organes et muscles dans le corps humain, notamment le larynx, les cordes vocales, la bouche,le nez et les poumons [3].

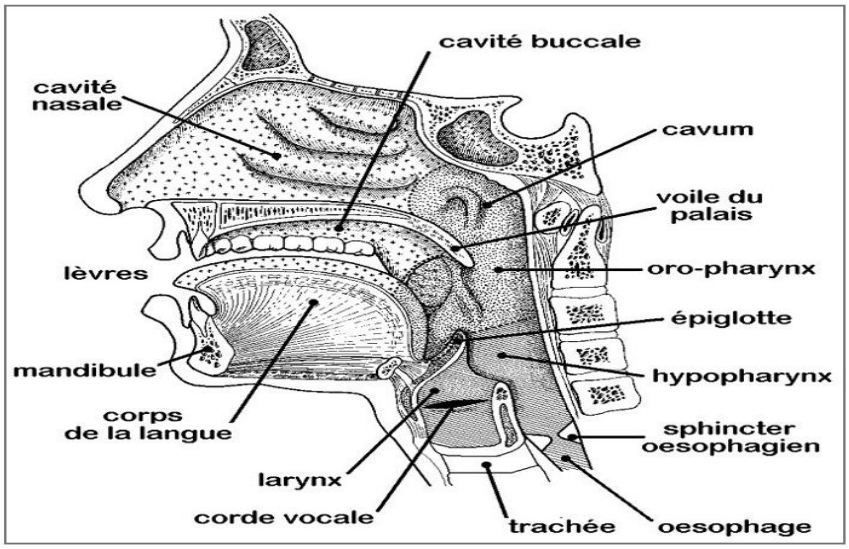

Figure1.1 Coupe schématique du conduit vocal et des principaux organe s'implique dans la production de la parole

Le processus commence par la respiration, où l'air est inspiré dans les poumons. Ensuite, l'air est expiré à travers les cordes vocales dans le larynx, qui produit des sons. Les sons sont ensuite modifiés à mesure qu'ils passent à travers la bouche et le nez pour produire des phonèmes, qui sont les sons de base de la langue [4].

Les phonèmes sont ensuite combinés pour former des mots, qui sont ensuite utilisés

pour construire des phrases et communiquer des idées. Tout au long du processus, le cerveau joue un rôle crucial en coordonnant les mouvements des organes impliqués dans la production de la parole et en choisissant les mots et les phrases appropriés à utiliser. [5]

1.2.2 L'audition de la parole

Est un processus complexe qui implique la perception, le traitement et la compréhension des sons de la parole. Le processus commence par l'oreille, où les sons sont captés par le pavillon de l'oreille et transmis à travers le canal auditif jusqu'au tympan. Les vibrations du tympan sont alors transmises aux osselets de l'oreille moyenne, qui amplifient le signal sonore [6].

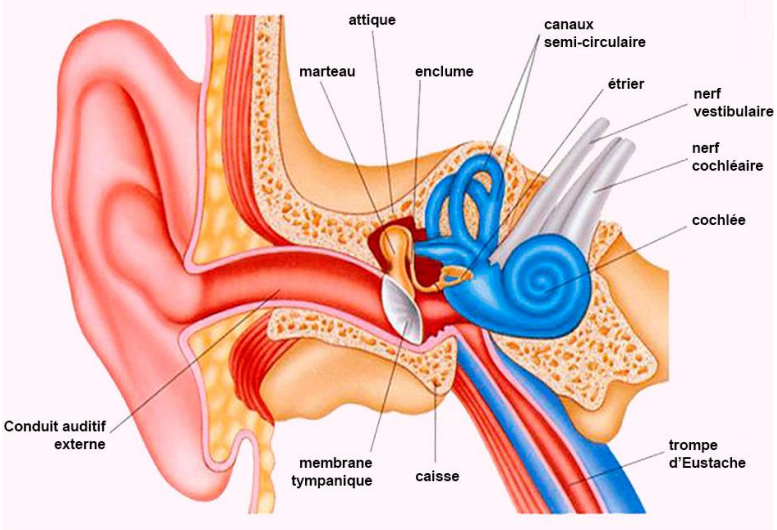

Figure 1.2 Section schématique de l'oreille

Le signal sonore amplifié est ensuite transmis à l'oreille interne, où il est converti en un signal électrique par les cellules ciliées de la cochlée. Ce signal électrique est ensuite transmis au cerveau par le nerf auditif. Le cerveau traite ensuite ce signal sonore pour en extraire les informations linguistiques pertinentes, telles que les phonèmes, les mots et les phrases. Le traitement de la parole dans le cerveau implique l'activation de plusieurs aires cérébrales, notamment le cortex auditif, le cortex préfrontal et le cortex temporal [7].

1.3 L'approche acoustique

L'approche acoustique de la parole est une méthode d'analyse qui se concentre sur les caractéristiques acoustiques du son de la parole. Cette approche s'intéresse principalement aux propriétés physiques des ondes sonores produites par la voix humaine, ainsi qu'aux paramètres acoustiques qui peuvent être extraits à partir de ces ondes sonores [8].

1.3.1 Propriétés du signal de parole[9]

Les signaux de parole sont des ondes acoustiques complexes contenant des informations cruciales sur la parole humaine, avec des propriétés spécifiques telles que la périodicité, la variabilité temporelle, le spectre de fréquence variable et la non-stationnarité. Ces propriétés ont des implications importantes pour l'analyse, la reconnaissance automatique et la modélisation de la parole, ainsi que pour le développement de technologies telles que la reconnaissance vocale, la synthèse de parole et la conversion de la parole.

- **Redondance**

Le signal de parole contient souvent des répétitions et des redondances, ce qui peut faciliter la détection de mots et la compréhension du message vocal. Cette propriété est également utilisée dans certaines techniques de codage de la parole pour réduire la quantité de données nécessaires pour transmettre un signal de parole.

- **La continuité :**

Le signal de parole est généralement un signal continu, sans pause ou interruption significative, sauf lorsque le locuteur fait des pauses ou respire. Cette propriété est essentielle pour la compréhension de la parole, car les mots et les phrases sont souvent liées les uns aux autres de manière fluide.

- **La variabilité du signal :**

Le signal de parole peut varier considérablement d'un locuteur à l'autre, ainsi que dans différents contextes et situations. Cette variabilité peut être due à des différences de tonalité, de vitesse, d'accent et d'autres facteurs qui peuvent affecter la perception et la compréhension du message vocal.

- **Le non stationnarité du signal :**

Le signal de parole est souvent considéré comme un signal non-stationnaire, car il peut varier considérablement en termes de contenu spectral, de durée et de niveau sonore. Cette propriété peut rendre la tâche de traitement du signal de parole plus complexe, en particulier

pour la reconnaissance automatique de la parole.

1.3.2 L'analyse acoustique du signal de parole

Est une méthode qui permet d'extraire des informations à partir des signaux acoustiques produits par la voix humaine. Cette analyse est souvent utilisée pour des applications telles que la reconnaissance automatique de la parole, la synthèse de la parole, et l'analyse de la prosodie de la parole

Prétraitements acoustiques

Les prétraitements acoustiques sont des techniques appliquées aux signaux acoustiques (tels que les signaux de parole) pour améliorer leur qualité et/ou faciliter leur traitement ultérieur. Ces techniques incluent notamment la filtrage, la normalisation, le préaccentuation, le fenêtrage et la segmentation.

Acquisition

L'acquisition de signaux de parole est le processus de capture des ondes acoustiques qui représentent la parole humaine et de conversion de ces ondes en signaux numériques pour l'analyse et le traitement. Les signaux de parole sont généralement capturés à l'aide de microphones, qui transforment les ondes acoustiques en signaux électriques. Les signaux électriques sont ensuite échantillonnés à une fréquence suffisamment élevée pour capturer l'ensemble du spectre de fréquence de la parole, typiquement à une fréquence d'échantillonnage de 8 kHz ou plus. Les signaux échantillonnés sont ensuite stockés sous forme numérique pour l'analyse et le traitement ultérieurs, tels que la reconnaissance automatique de la parole, la synthèse de la parole et la modification de la parole. Des techniques de prétraitement, telles que le filtrage, l'élimination du bruit et le fenêtrage, peuvent également être appliquées pour améliorer la qualité des signaux de parole avant l'analyse et le traitement.

- **Théorème de Shannon :**

  Le théorème de Shannon, également connu sous le nom de théorème d'échantillonnage de Nyquist-Shannon, est un résultat fondamental en théorie de l'information et en traitement du signal. Il établit que pour qu'un signal continu puisse être parfaitement reconstruit à partir de ses échantillons numériques, la fréquence d'échantillonnage doit être au moins deux fois

supérieure à la fréquence maximale présente dans le signal.

- **Préaccentuation [10]**

Est une technique de prétraitement appliquée au signal audio pour augmenter l'amplitude des hautes fréquences et améliorer ainsi la qualité sonore. Elle consiste à filtrer le signal à l'aide d'un filtre passe-haut avant son enregistrement ou sa transmission. Cette technique permet d'atténuer les effets de la distorsion et du bruit dans les hautes fréquences, qui peuvent affecter la clarté du signal.

$$X(z) = 1 - \alpha z^{-1} \quad (1.1)$$

- **Fenêtrage**

Est une technique de traitement de signal qui consiste à multiplier le signal par une fonction de fenêtre pour extraire des segments de données du signal continu. Cette technique est utilisée pour limiter l'effet des discontinuités du signal, comme les bords de la fenêtre. Elle est souvent utilisée en conjonction avec la transformée de Fourier pour extraire les composantes fréquentielles d'un signal donné. [11]

L'opération de fenêtrage consiste à multiplier le signal $x(n)$ par un autre signal $h(k)$ possédant N échantillons unités.

$$s(n) = \sum_{k=1}^{N} x(n).h(k) \quad (1.2)$$

Avec :

$s(n)$: Signal résultant

$x(n)$: Signal à fragmenter

$h(k)$: Fenêtre de Hamming, $k = 1, \ldots\ldots, N$

$$h(k) = \alpha + (1 - \alpha)\cos(\frac{2\pi k}{N})$$

$$\alpha = 0.54$$

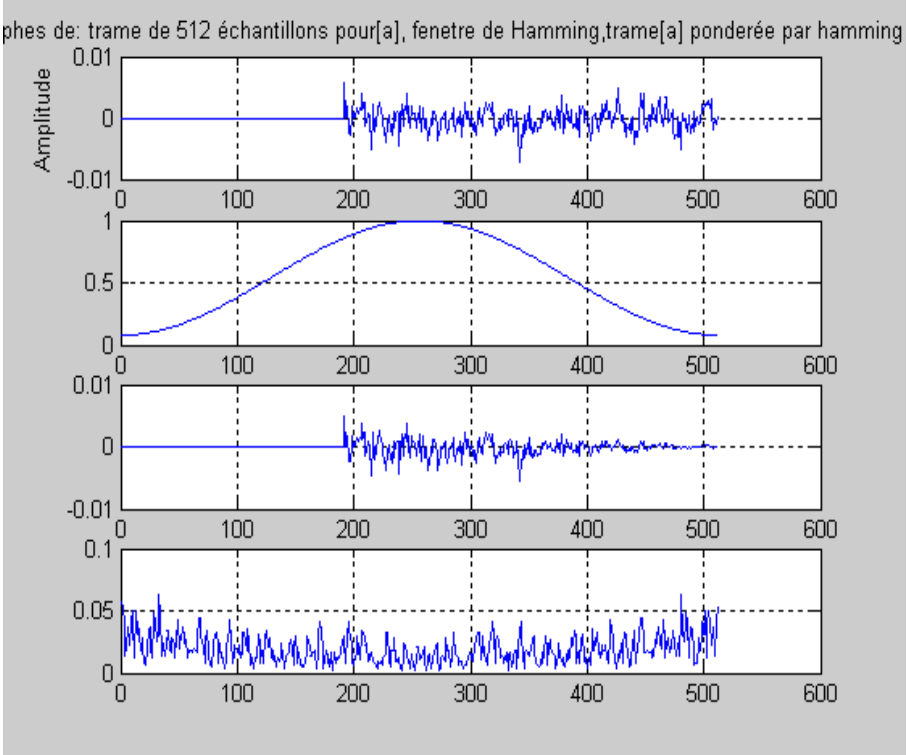

Figure1.3 Fenêtrage

- **Détection de l'activité vocale (VAD en anglais)**

Est une tâche de traitement du signal qui consiste à détecter les segments d'un enregistrement audio contenant des informations vocales (parole ou chant), tout en rejetant les segments qui ne contiennent que du bruit ou des sons non vocaux. Cette tâche est importante pour de nombreuses applications telles que la reconnaissance de la parole, la compression de la parole, la transcription de la parole. [12]

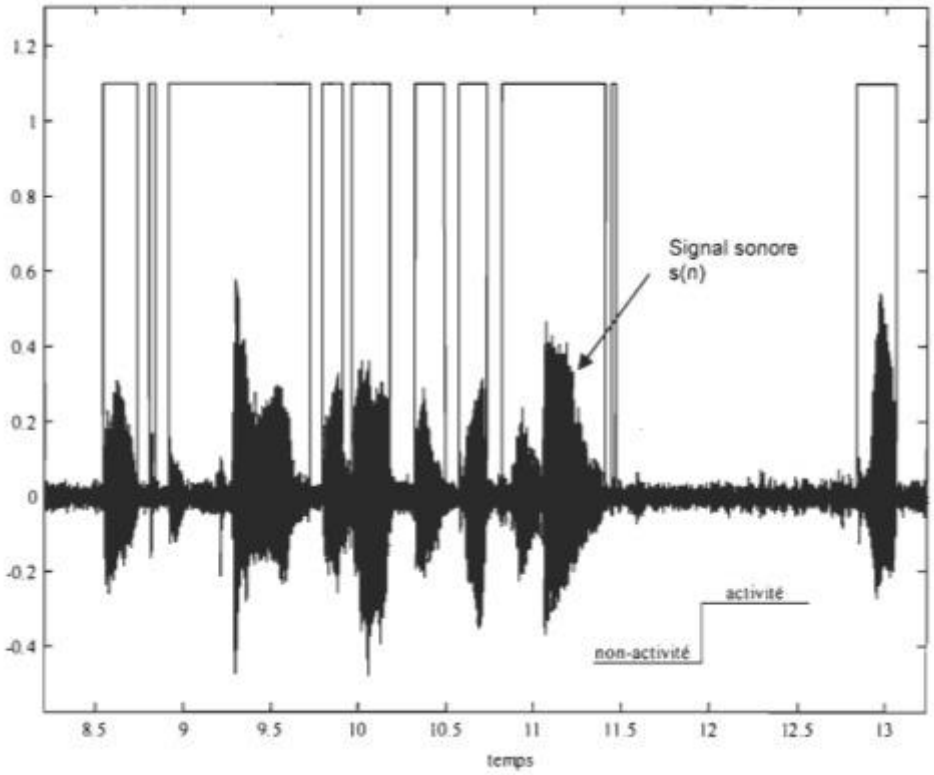

Figure 1.4 Exemple illustrant le principe de la détection d'activité vocale

1.4 L'analyse spectrale et temporelle de la parole

Est un processus qui permet de décomposer les signaux acoustiques de la parole en différentes composantes pour étudier leur structure et leur contenu

1.4.1 L'analyse spectrale

L'analyse spectrale de la parole consiste à décomposer un signal acoustique en ses composantes fréquentielles à l'aide de la transformée de Fourier. Cette analyse permet d'obtenir un spectre de fréquence qui révèle la composition fréquentielle du signal [12].
L'équation de la transformée de Fourier d'un signal continu $f(t)$ est donnée par :

$$F(\omega) = \int f(t) \exp(-i\omega t)\, dt \qquad (1.3)$$

$F(\omega)$est la transformée de Fourier du signal $f(t), \omega$ est la fréquence angulaire en radians par seconde, i est l'unité imaginaire

En traitement de la parole, l'analyse spectrale est utilisée pour extraire des caractéristiques acoustiques qui sont utilisées pour la reconnaissance automatique de la parole.

Les caractéristiques spectrales telles que les formants, la fréquence fondamentale et le rapport signal-bruit sont utilisées pour identifier les phonèmes et les mots de la parole [13].

1.4.2 L'analyse temporelle

L'analyse temporelle de la parole consiste à examiner les caractéristiques temporelles des signaux acoustiques de la parole. Elle est utilisée pour étudier des aspects tels que la durée des segments de parole, la fréquence d'articulation, les pauses et les phénomènes suprasegmentaux tels que l'intonation et le rythme

- **Energie**

L'énergie à court terme est utilisée pour détecter les périodes de silence dans un signal. Elle est élevée en présence d'un son et faible en l'absence de son, c'est-à-dire pendant les périodes de silence. En outre, les sons voisés présentent généralement une énergie plus élevée que les sons non voisés.

$E = \sum_{n=1}^{N} s^2(n)$(1.4)

$s(n)$: le n-iéme échantillon de la trame considérée.

N : nombre d'échantillons de la fenêtre considéré.

- **Fréquence fondamentale (F0)**

La fréquence fondamentale (F0), également appelée tonalité, est une propriété acoustique du signal de parole qui représente la fréquence de vibration des cordes vocales lors de la production de la voix. Elle est mesurée en Hertz (Hz) et correspond à la hauteur perçue d'un son. La F0 est particulièrement importante pour la perception de la mélodie de la parole et pour distinguer les voix de différents locuteurs. Elle est également utilisée dans la reconnaissance automatique de la parole pour la détection des changements de tonalité, tels que les contours de phrases et les questions. La F0 varie considérablement entre les locuteurs et peut être influencée par des facteurs tels que l'âge, le sexe et l'émotion. Elle peut être mesurée de différentes manières, notamment par des algorithmes d'analyse de la période glottale ou de la transformée de Fourier à court terme.[14]

- **Taux de passage par zéro (TPZ)**

Le taux de passage par zéro est une mesure de la fréquence à laquelle le signal de parole traverse l'axe horizontal, ce qui correspond à la fréquence de changement de signe dans le signal. Cette mesure est utile pour détecter les transitions entre les sons voisés (où le taux de passage par zéro est relativement faible) et les sons non-voisés (où le taux de passage par zéro est relativement élevé). Le taux de passage par zéro est souvent utilisé en conjonction avec

d'autres mesures pour l'analyse de la parole, telle que l'énergie à court terme. Il est également utilisé dans la reconnaissance automatique de la parole pour la détection des régions de transition entre les sons et pour la segmentation des mots dans le signal de parole. [15]

Le taux de passage par zéro est défini par l'expression suivante :

$$TPZ = \frac{i.100}{N}\% \quad (1.5)$$

i : le nombre de passage par zéro

N : la taille de la fenêtre d'analyse

Le taux de passage par zéro des sons non voisés est supérieur à celui des sons voisés

1.5 L'analyse spectrographique de la parole

Est largement utilisée dans différents domaines, tels que la phonétique, la reconnaissance automatique de la parole, la thérapie de la parole et la synthèse de la parole. Elle peut être utilisée pour mesurer les différences entre les sons de la parole produits par des locuteurs différents, ou pour identifier les caractéristiques acoustiques qui sont importantes pour la reconnaissance de la parole par un système informatique.[16]

Le spectrogramme est souvent utilisé pour l'analyse de la parole car il permet de visualiser les différentes caractéristiques acoustiques du signal sonore, telles que les formants, les transitions consonant-voyelle, les pauses et les changements d'intensité. Ces caractéristiques peuvent être utilisées pour identifier les sons de la parole et pour comprendre la manière dont la parole est produite [17]

L'équation pour l'analyse spectrographique de la parole consiste à prendre une transformée de Fourier à court terme (STFT) d'un signal de parole. Ceci peut être représenté mathématiquement comme suit : [18]

$$X(t,f) = \int x(\tau)w(\tau - t)e^{(-j2\pi f\tau)}d\tau \quad (1.6)$$

$X(t,f)$est la valeur du spectrogramme complexe au temps t et à la fréquence f. $x(\tau)$est le signal de parole.$(w(\tau - t))$est une fonction de fenêtre qui est généralement appliquée au signal de parole avant d'effectuer la STFT pour réduire la fuite spectrale.$e^{(-j2\pi f\tau)}$est l'exponentielle complexe utilisée pour pondérer le signal vocal à chaque case de fréquence.L'intégrale est prise sur une fenêtre de temps court centrée sur l'instant t

1.6 Extraction des paramètres MFCCs

Les coefficients MFCC (Mel-FrequencyCepstral Coefficients) sont un ensemble de paramètres largement utilisés dans le domaine de la reconnaissance de la parole et du traitement du signal audio. Les étapes suivantes décrivent comment extraire les paramètres MFCC à partir d'un signal audio :

1.6.1 Calcul des coefficients cepstraux (MFCC)

Le calcul des paramètres MFCC se réalise de la façon suivante :

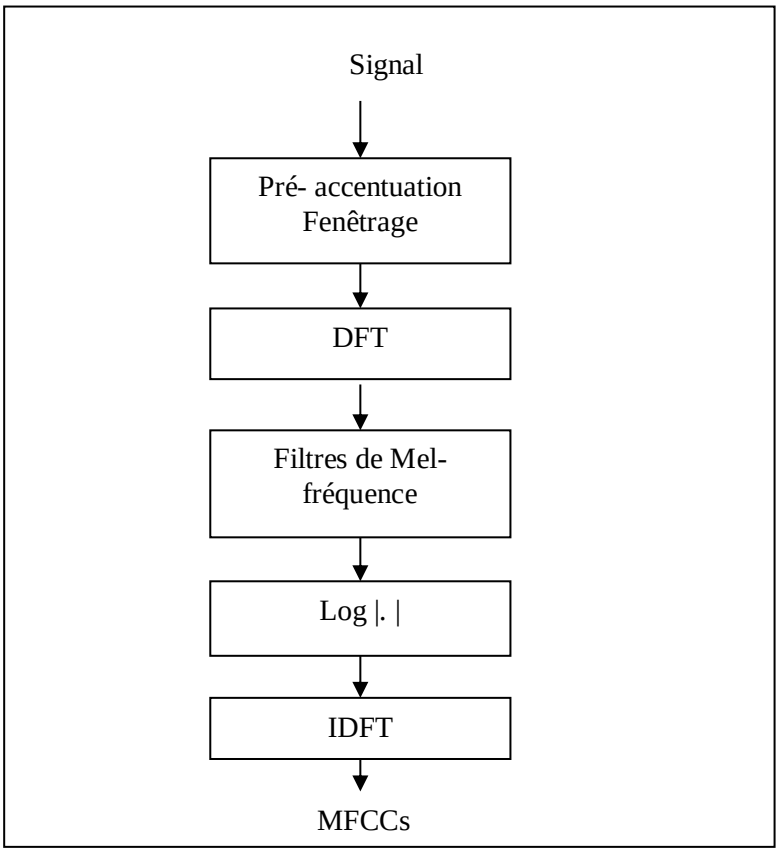

Figure 1.5calcul des MFCCs

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

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

$S[k]$ est obtenue par :

$$s[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\Pi nk/N} \qquad (1.7)$$

Le spectre du signal est multiplié avec des filtres triangulaires (voir Fig.1.6) dont les bandes passantes sont équivalentes en domaine Mel-fréquence.Les points frontières *B[m]* des filtres en mel-fréquence sont calculés ainsi :

$$B[m] = B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1} \qquad 0 \le m \le M+1 \qquad (1.8)$$

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

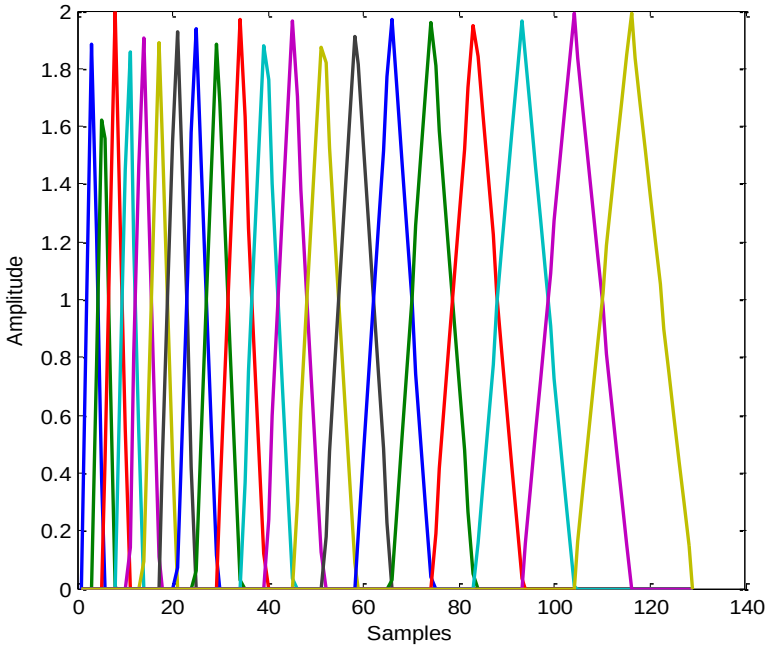

Figure 1.6 Mel filterbanks

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

$$f[m] = \left(\frac{N}{F_s}\right) B^{-1}\left(B(f_l) + m\frac{B(f_h) - B(f_l)}{M+1}\right) \qquad (1.9)$$

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

Le coefficient $H_m[k]$ de chaque filtre est déterminé par le système suivant :

$$H_m[k] = \begin{cases} 0 & si \quad k \le f[m-1] \\ \dfrac{k - f[m-1]}{f[m] - f[m-1]} & si \quad f[m-1] \le f[m] \\ \dfrac{f[m+1] - k}{f[m+1] - f[m]} & si \quad f[m] \le k \le f[m+1] \\ 0 & si \quad k \ge f[m+1] \end{cases} \tag{1.10}$$

Pour un spectre lissé et stable, à la sortie des filtres un logarithme de spectre d'amplitude est calculé :

$$E[m] = \log\left[\sum_{k=0}^{N-1} |S[k]|^2 H_m[k]\right] \qquad 0 \le m \le M \tag{1.11}$$

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

$$c[n] = \sum_{m=0}^{M-1} E[m] \cos\left(\frac{\pi n(m+\frac{1}{2})}{M}\right) \qquad 0 \le n \le M \tag{1.12}$$

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

1.7 La reconnaissance automatique de locuteurs (RAL)

La RAL est une technologie qui permet d'identifier de manière automatique les locuteurs dans un enregistrement vocal. La RAL peut être utilisée dans plusieurs domaines, tels que la sécurité, l'authentification, la surveillance ou encore la retranscription de conversations[19].

Le processus de RAL consiste en plusieurs étapes, notamment la segmentation de l'enregistrement en segments contenant chacun la parole d'un seul locuteur, l'extraction de caractéristiques vocales de chaque segment, la comparaison de ces caractéristiques avec celles d'autres segments afin de déterminer si le locuteur est déjà connu ou inconnu, et enfin l'identification du locuteur en utilisant des techniques de reconnaissance de modèles ou de machine learning [20].

La reconnaissance automatique de locuteurs (RAL) implique généralement deux tâches distinctes

1.7.1 L'identification de locuteur

L'Identification Automatique du Locuteur (IAL) est le processus qui consiste à déterminer, parmi une population de locuteurs connus, la personne ayant prononcé un message donné. D'un point de vue schématique, une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est comparée à une référence caractéristique du locuteur : identité du locuteur dont la référence est la plus proche de la séquence de parole est donnée en sortie du système d'IAL. Deux modes sont proposés en IAL :
- l'identification en ensemble fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du système ;
- et l'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu.

En mode «ensemble ouvert», le système d'IAL doit décider de la fiabilité de son jugement en acceptant ou rejetant identité qu'il a trouvée. De par son principe - déterminer une identité parmi les identités potentielles - les performances des systèmes d'IAL se dégradent généralement au fur et à mesure que la population de locuteurs augmente [21].

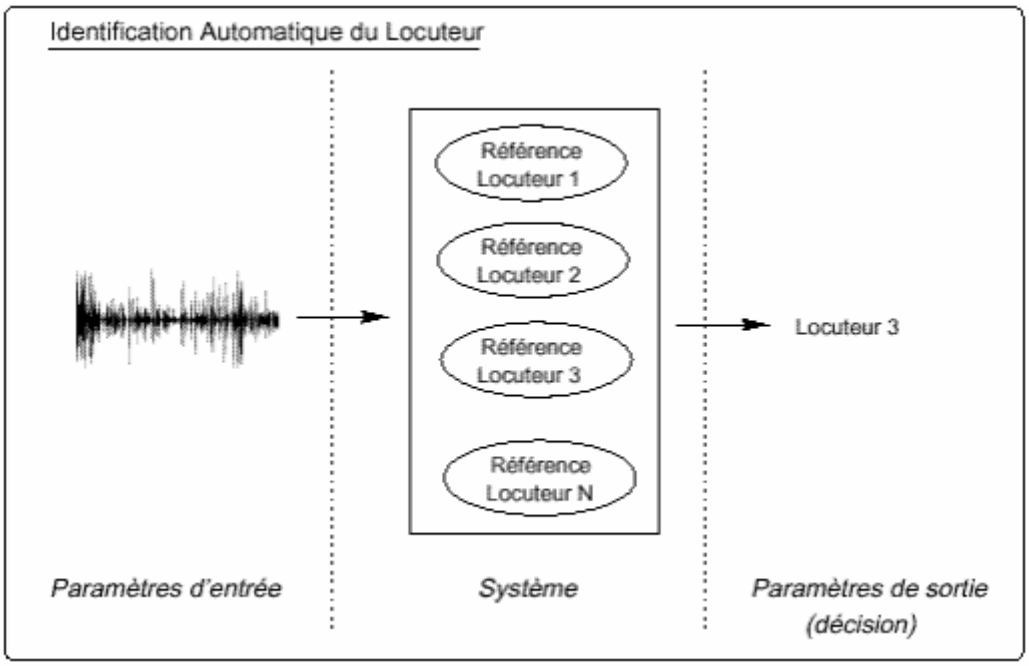

Figure 1.7 Principe de base de la tâche d'Identification Automatique du Locuteur

1.8 La Vérification Automatique du Locuteur (VAL)

La VAL est une technique de reconnaissance de locuteur qui vise à déterminer si l'identité revendiquée par un locuteur est correcte ou non, en comparant les caractéristiques acoustiques

de sa voix à celles d'un modèle enregistré préalablement.

La VAL est utilisée dans de nombreux domaines, tels que la sécurité, l'authentification, la surveillance ou encore la retranscription de conversations. Elle permet de garantir l'identité d'un locuteur en s'assurant que la voix enregistrée correspond bien à celle du locuteur revendiqué

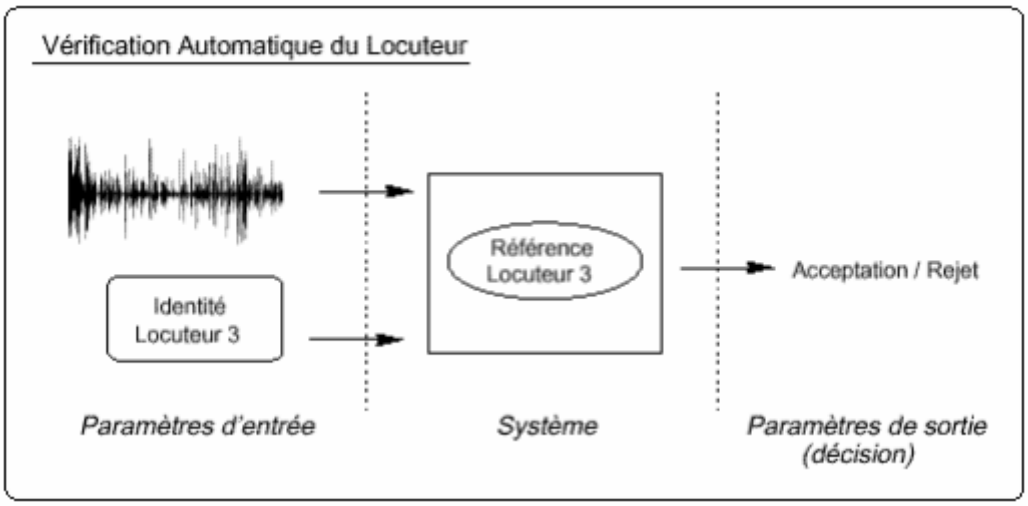

Figure 1.8 Principe de base de la tâche de Vérification Automatique du Locuteur

1.9 Conclusion

Ce chapitre porte un aperçu sur le système de production de la parole, Nous avons brièvement parcouru les différentes techniques utilisées pour l'analyse de la parole, ainsi quelles étapes à suivre pour extraire les paramètres MFCCs, en vue de leur utilisation dans la vérification du locuteur.

Dans le prochain chapitre ,nous aborderons l'apprentissage des données du locuteur (MFCCs) par le classificateur GMM/SVM.

Chapitre 2

L'apprentissage et la classification des données du locuteur par GMM/SVM

2.1.Introduction

L'apprentissage et la classification des données du locuteur par GMM/SVM sont des techniques courantes dans le domaine de la reconnaissance de locuteurs et de la biométrie. Les modèles de mélange gaussien (GMM) sont utilisés pour modéliser les distributions de probabilité des caractéristiques vocales du locuteur, tandis que les machines à vecteurs de support (SVM) sont utilisées pour séparer les caractéristiques des différents locuteurs [22].

Ce chapitre va être consacré au développent des concepts et fondements mathématiques dédiés aux classifier Bayésien GMM-UBM ainsi qu'aux modèle binaire SVMs.

2.2. Modes d'Apprentissage

2.2.1 Apprentissage non supervisé

L'apprentissage non supervisé est une technique d'apprentissage automatique qui consiste à extraire des motifs et des structures à partir de données non étiquetées, c'est-à-dire des données qui ne sont pas classées ou regroupées en classes prédéfinies. Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé ne nécessite pas de données d'entraînement annotées par un superviseur. [23]

2.2.2 Apprentissage supervisé

L'apprentissage supervisé est une technique d'apprentissage automatique qui consiste à entraîner un modèle à prédire une sortie pour des données d'entrée en se basant sur un ensemble de données d'entraînement étiquetées. Contrairement à l'apprentissage non supervisé, l'apprentissage supervisé nécessite des données d'entraînement annotées par un superviseur.

2.3.Les approches d'apprentissage des données acoustiques

2.3.1 Les approches basées sur les Mélanges de Gaussiennes (GMM)

L'outil des mélanges de Gaussiennes (appelé Gaussien mixture, mixture of Gaussians, GMM, GM ou MoG selon les sources) est largement répandu dans les domaines de la littérature et de l'ingénierie informatique. Il est couramment employé pour modéliser des données numériques ou effectuer le clustering d'un ensemble d'individus. Le recours à un modèle GMM est justifié principalement par l'interprétation des classes du mélange. Il est certain que les vecteurs de paramètres seront distribués de manière différente en fonction des caractéristiques du son de parole considéré, qu'il soit voisé ou non voisé. Chaque composante du modèle va représenter des ensembles sous-jacents de classes acoustiques, chacune de ces classes

représentant des événements acoustiques tels que les voyelles, les nasales, etc. Ces classes permettent de caractériser l'espace acoustique propre à chaque locuteur. [24]

2.3.2 Modèle du mélange

La densité d'un individu x représenté dans l'espace vectoriel Rd selon la loi normale (ou Gaussienne) avec une moyenne μ et une matrice de covariance Σ peut être exprimée de la manière suivante :

$$f(x/\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \quad (2.1)$$

En superposant et pondérant *M* Gaussiennes, on définit un mélange de Gaussiennes. On Note habituellementπ_m, respectivement $\mu_m$et $\Sigma_m$le poids (sous les contraintes et$\pi_m >$ ($\sum_{m=1}^{M}\pi_m = 1$)respectivement la moyenne et la matrice de covariance de la m-ième composante. On note également$\theta_m = \{\pi_m, \mu_m, \Sigma_m\}$ ainsi que $\theta = \{\theta_m\}$. La densité d'un individu $x$selon la distribution de probabilité paramétrée par θ est donnée par :

$p(x/\theta) = \sum_{m=1}^{M} \pi_m f(x/\mu_m, \Sigma_m)$ (2.2)

La probabilité jointe d'un échantillon$X = (x_1, x_2, \dots, x_N)^T$, où chaque ligne caractérise un individux_n,et en supposant que l'échantillon est indépendamment et identiquement distribué (i.i.d), peut être exprimée de la manière suivante :

$p(X/\theta) = \prod_{n=1}^{N} p(x_n/\theta)$ (2.3)

Pour des raisons pratiques, on préfère souvent utiliser le logarithme de la probabilité jointe de l'échantillon, appelée vraisemblance.

2.3.3 Adaptation du modèle par Maximum *A Posteriori* (MAP)

L'adaptation bayésienne, également appelée MAP (Maximum A posteriori), est une technique d'apprentissage qui permet de prendre en compte des contraintes probabilistes sur les paramètres des modèles. Cette méthode est appliquée aux modèles qui ont déjà été entraînés et pour lesquels on dispose de données a priori. Elle permet de créer de nouveaux modèles dépendants d'un locuteur en particulier, à partir d'un modèle initial indépendant. L'adaptation

bayésienne se décompose en deux étapes. Dans la première étape, on calcule les paramètres statistiques des trames d'apprentissage par rapport au modèle UBM. Dans la pratique, on adapte uniquement les moyennes du GMM, les poids et les variances restant inchangés. La seconde étape consiste à combiner les nouveaux paramètres, obtenus lors de la première étape, avec les paramètres du modèle UBM en utilisant des coefficients de pondération. Cette étape permet de faire varier l'influence des données a priori en fonction du nombre de données d'apprentissage pour chaque gaussienne du modèle. Ainsi, seules les gaussiennes occupées par un nombre important de trames d'apprentissage seront modifiées, les paramètres des autres gaussiennes restant inchangés par rapport à leurs valeurs a priori. Cette technique permet donc une adaptation fine des modèles en fonction des données d'apprentissage disponibles [24][25].Etant donné un signal de parole représenté par une séquence de vecteurs acoustiques $X = \{x_1, x_2, \dots, x_N\}$,les formules suivantes sont appliquées uniquement aux vecteurs Moyennes $\mu_i$du modèle UBM (*M* gaussiennes) pour obtenir les vecteurs moyennes adaptés

$$\begin{aligned} \bar{\mu}_i &= \alpha_i E_i(X) + (1+\alpha_i)\mu_i, i = 1, \dots, M \\ \alpha_i &= \frac{n_i(X)}{n_i(X)+r} \\ n_i(X) &= \sum_{j=1}^{N} P(i/x_j) \\ E_i(X) &= \frac{1}{n_i} \cdot \sum_{j=1}^{N} P(i/x_j) x_j \\ P(i/x_j) &= \frac{\pi_i p_i(x_j)}{\sum_{k=1}^{M} \pi_k p_k(x_j)} \end{aligned} \qquad (2.4)$$

2.4.La fonction Noyau

la fonction noyau est une fonction mathématique utilisée pour transformer les données d'entrée dans un espace de dimension supérieure. Elle permet de résoudre des problèmes de classification non linéaires en introduisant des relations non linéaires entre les données.[26]

Les fonctions noyau couramment utilisées dans GMM/SVM comprennent le noyau linéaire, le noyau polynômial, le noyau gaussien (RBF - Radial Basis Fonction), etc. Chaque type de noyau a ses propres caractéristiques et peut être choisi en fonction des caractéristiques des données et des performances souhaitées.

2.5.Les approches d'apprentissage des données acoustiques

Nous avons un ensemble d'apprentissage de l'ensemble $\{(\boldsymbol{x1}, \boldsymbol{y1}), (\boldsymbol{x2}, \boldsymbol{y2}), \ldots\ldots, (\boldsymbol{xi}, \boldsymbol{yi})\}$ où $\boldsymbol{xi} \in \boldsymbol{R^n}$ avec $\boldsymbol{i} = \boldsymbol{1} \ldots \boldsymbol{l}$et $\boldsymbol{yi} \epsilon \{\pm \boldsymbol{1}\}$. Pour classer cet ensemble, nous utilisons une famille de fonctions linéaires définies par l'équation $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + \boldsymbol{b} = \boldsymbol{0}$, où $\boldsymbol{w} \epsilon \boldsymbol{R^n}$et $\boldsymbol{b} \epsilon \boldsymbol{R}$. Cette famille de fonctions linéaires est utilisée pour déterminer si un vecteur appartient à l'une des deux classes.

$$f(x) = \text{sign}\,(\langle w, x \rangle + b = 0) \quad (2.5)$$

2.5.1 Cas de données linéairement séparables

Afin de séparer de manière optimale les deux classes, nous allons créer un hyperplan H qui satisfait l'équation $\langle w, x \rangle + b = 0$. Cet hyperplan sera situé à mi-distance entre les deux hyperplans H1 et H2 parallèles à H, qui sont définis respectivement par les équations :

$$H1: \langle w, xi \rangle + b = +1 \quad (2.6)$$

$$H2: \langle w, xi \rangle + b = -1 \quad (2.7)$$

Telle que les deux conditions suivantes soient respectées:

Condition1

Il n'y a aucun point qui se situe entre H_1et H_2.Cette contrainte se traduit par les inégalités:

$$\langle w, xi \rangle + b \geq +1 \text{pour} yi = +1 \quad (2.8)$$

Et

$$\langle w, xi \rangle + b \geq -1 \text{pour} yi = -1 \quad (2.9)$$

Ces deux inégalités peuvent être combinées en une seule :

$$yi(\langle w, xi \rangle + b) \geq +1 \quad (2.10)$$

Condition 2

La distance ou la marge maximale entre $H1$ et $H2$ dans le cas de la séparation par hyperplan est donnée par $M = \frac{2}{\|w\|}$.Pour maximiser M, cela revient à minimiser $\|w\|$ou à minimiser$\|w\|^2$, où $\|w\|^2$représente le carré de la norme euclidienne du vecteur W. Ainsi, le problème de séparation par hyperplan optimal peut être formulé comme suit : Minimiser $\|w\|^2$, sujet à la contrainte de classification correcte :

$$\begin{cases} \min\limits_{w \in \Re^N} \frac{1}{2} \| w \|^2 \\ y_i(\langle w, x_i \rangle + b) \geq +1 \forall i = 1, \ldots, l \end{cases} \quad (2.11)$$

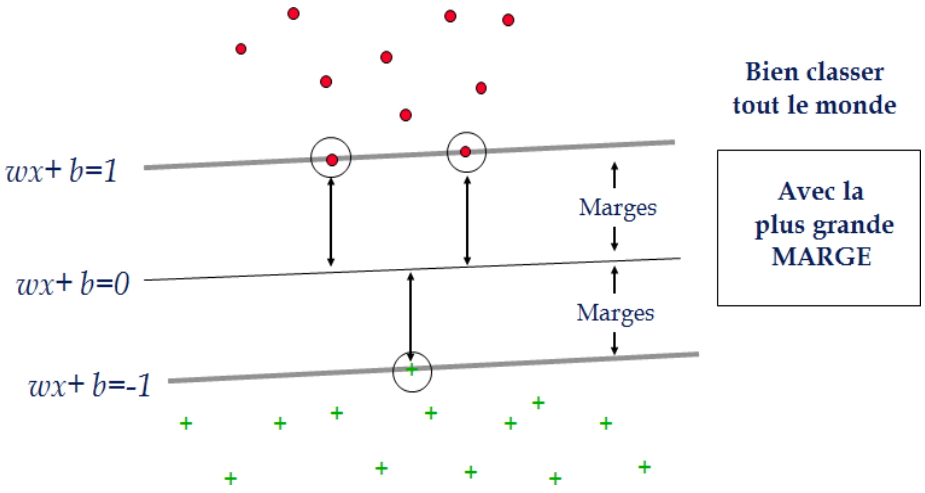

Figure 2.1 Données linéairement séparables

Le Lagrangien associé au problème d'optimisation précédent peut être formulé de la manière suivante :

$$f(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i (y_i(\langle w, x_i \rangle + b) - 1) \quad (2.12)$$

Le Lagrangien doit être minimisé par rapport à w et b et maximisé par rapport à α

$$\frac{\partial L}{\partial w} = 0 \quad (2.13)$$

$$\frac{\partial L}{\partial b} = 0 \quad (2.14)$$

Dans l'étape de prétraitement acoustique, nous utilisons la méthode VAD (Voice Activity Detection) pour détecter la présence de fragments sonores et supprimer le silence du flux de parole. Le module de paramétrage fournit des vecteurs propres de 14 coefficients MFCC toutes les 15 ms, en utilisant une fenêtre de Hamming de 25 ms et les $a_i \geq 0$

A partir des relations (2.13) et (2.14), nous pouvons déduire :

$$w = \sum_{i-1}^{l} \alpha_i y_i x_i \quad (2.15)$$

$$\sum_{i-1}^{l} \alpha_i y_i = 0 \quad (2.16)$$

En les remplaçant dans $L(w, b, \alpha)$, on obtient le problème dual :

$$L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{,j=1}^{l}{}_{i} \alpha_i \alpha_j y_i y_j x_i x_j \quad (2.17)$$

À maximiser sous les contraintes

$\sum_{i-1}^{l} \alpha_i y_i = 0 \text{et} a_i \geq 0 i = 1, \ldots., l$ (2.18)

La fonction de décision est alors :

$f(x) = \text{sgn}(\sum_{i=1}^{l} \alpha_i y_i \langle x_i, x \rangle + b)$ (2.19)

La fonction de décision des SVM est influencée uniquement par les points correspondants aux coefficients α non nuls. Ces points sont appelés vecteurs de support. Lorsque les données sont linéairement séparables, les vecteurs de support sont les points les plus proches de la limite de décision, c'est-à-dire ceux qui se trouvent exactement à une distance égale à la marge.

Cette caractéristique des SVM est particulièrement intéressante, car seuls les vecteurs de support sont nécessaires pour décrire la limite de décision. Le nombre de vecteurs de support dans le modèle optimal est généralement faible par rapport au nombre total de données d'entraînement.

En d'autres termes, les SVM utilisent une approche sélective où seuls les vecteurs de support, qui sont les points critiques pour définir la frontière de décision, sont pris en compte. Cela permet d'avoir un modèle plus efficace avec une représentation plus concise de la décision, même lorsque le nombre de données d'entraînement est élevé.

2.5.2 Cas de données non-linéairement séparables

Condition 1

La distance entre les vecteurs bien classés et l'hyperplan doit être maximal.

Condition 2

La distance entre les vecteurs mal classés et l'hyperplan doit être maximal aussi. Pour formaliser le principe de la marge souple, on introduit des variables de pénalité non-négatives ε_i , $i = 1, \ldots.,$ appelées variables d'écart. Les contraintes de l'équation (2.11) sont alors transformées de la manière suivante : $y_i(\langle w, x_i \rangle + b) \geq +1 - \varepsilon_i$, $i = 1, \ldots., l$ (2.20)

Avec cette introduction de termes de pénalité, la fonction objective est modifiée comme suit :

$\min_{w,b,\varepsilon} \left(\frac{1}{2} w^T w\right) + C \sum_{i=1}^{l} \varepsilon_i$, $C \geq 0$ (2.21)

Le paramètre C est défini par l'utilisateur. Il représente la tolérance au bruit de classificateur et la pénalité pour la violation des contraintes. Une valeur plus élevée de C permet une tolérance plus élevée aux erreurs de classification et conduit à une marge de séparation plus étroite.

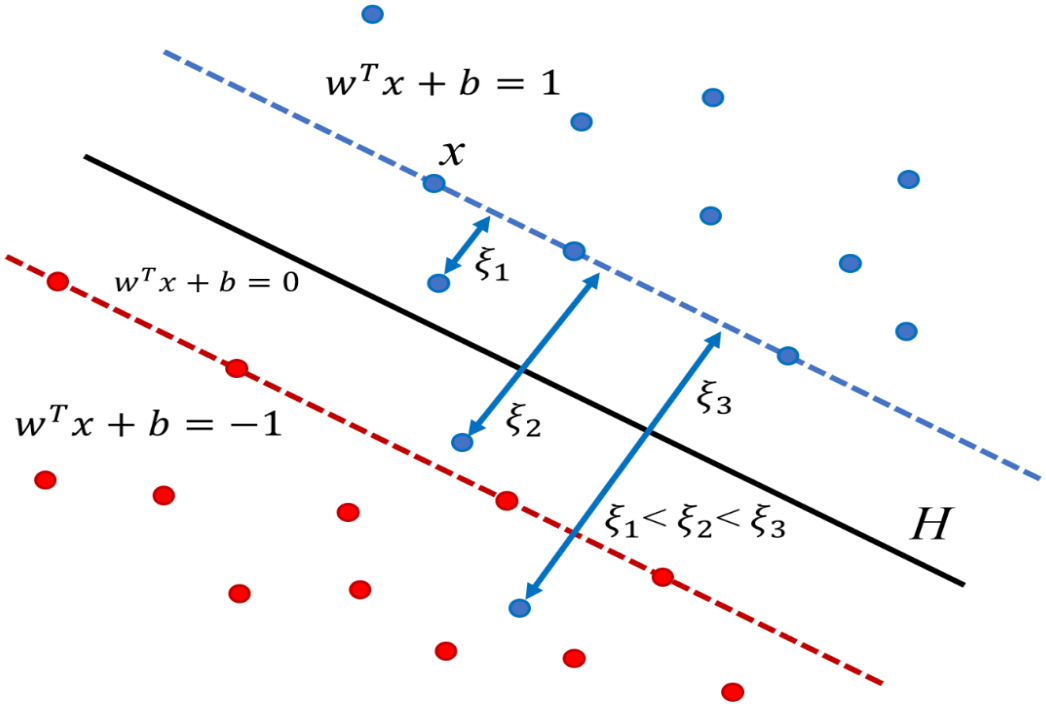

Figure 2.2 Données non linéairement séparables

La nouvelle formulation d'optimisation est alors :

$$\begin{cases} \min_{w,b,\varepsilon}\left(\frac{1}{2}w^T w\right) + C\sum_{i=1}^{l}\varepsilon_i \ , & C \geq 0 \\ y_i(\langle w, x_i\rangle + b) \geq +1 - \varepsilon_i \ , & \forall \varepsilon_i \geq 0 pour i = 1, \ldots, l \end{cases} \quad (2.22)$$

En introduisant les multiplicateurs de Lagrange, le Lagrangien associé au nouveau problème d'optimisation devient :

$$L(w,b,\varepsilon_i,\mu) = \frac{1}{2}w^T w + C\sum_{i=1}^{l}\varepsilon_i - \sum_{i=1}^{l}\alpha_i[y_i(w^T x_i + b) + \varepsilon_i - 1] - \sum_{i=1}^{l}\varepsilon_i \mu_i \quad (2.23)$$

$$= \frac{1}{2}w^T w + \sum_{i=1}^{l}(C - \alpha_i - \mu_i)\varepsilon_i - (\sum_{i=1}^{l}\alpha_i y_i x_i)w - (\sum_{i=1}^{l}\alpha_i y_i)b + \sum_{i=1}^{l}\alpha_i \quad (2.24)$$

Le Lagrangien doit être minimisé par rapport à w,b,ε_i et maximisé par rapport $\alpha et\ \mu$.

$$\frac{\partial L}{\partial w} = 0 \quad (*)$$

$$\frac{\partial L}{\partial b} = 0 \quad (**)$$

$$\frac{\partial L}{\partial \varepsilon_i} = 0 \quad (***)$$

De ces dernières relations, on peut tirer les égalités suivantes :

$w = \sum_{i=1}^{l} \alpha_i y_i x_i$; $\sum_{i=1}^{l} \alpha_i y_i = 0$ et$\alpha_i = C - \mu_i$ (2.25)

Ce qui conduit à un problème dual légèrement différent de celui du cas séparable :

$L_D = \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j x_i x_j$ (2.26)

À maximiser sous les contraintes

$$\begin{cases} 0 \leq \alpha_i \leq \mathrm{C} \quad \forall i = 1, \ldots\ldots, l \\ \quad et \sum_{i=1}^{l} \alpha_i y_i = 0 \end{cases} \quad (2.27)$$

La seule différence avec le cas linéairement séparable est donc l'introduction d'une borne supérieure pour les paramètresα_i

2.6.Les machines à vecteurs de support multi-classe

Sont souvent utilisées pour des tâches de classification binaire, mais elles peuvent également être étendues pour effectuer des classifications multi-classes. L'adaptation des SVM binaires au cas multi-classe peut être réalisée de trois manières différentes, en fonction de la taille du problème.

- L'approche "un contre tous" consiste à entraîner un SVM binaire en utilisant les éléments d'une classe par rapport à tous les autres. Cela implique de résoudre environ c problèmes de SVM, chacun ayant une taille n.
- L'approche "un contre un" consiste à entraîner $c(c-1)/2$ SVMs pour chaque paire de classes, puis à décider de la classe gagnante par un vote majoritaire ou en post-traitant les résultats à l'aide d'estimations de probabilités a posteriori. Le nombre de classifieurs SVM à entraîner peut être réduit en utilisant un codage astucieux des classes, comme un code correcteur d'erreur ou un graphe acyclique direct (DAGSVM).
- L'approche globale consiste à traiter le problème en une seule fois. Cela peut être réalisé en posant formellement le problème, par exemple en notant $f'(x) - b'$ la fonction de discrimination associée à la classe '.

2.7.Les approches basées sur le modèle hybride GMM-SVM

Combinent deux techniques d'apprentissage automatique : le modèle de mélange de Gaussiennes (GMM) et les machines à vecteurs de support (SVM). Cette approche hybride est souvent utilisée dans des tâches telles que la classification et la reconnaissance de motifs.

L'approche hybride GMM-SVM combine ces deux techniques de la manière

suivante :

- **Entraînement du modèle GMM**

  Les données d'entraînement sont utilisées pour estimer les paramètres du GMM, c'est-à-dire les moyennes, les covariances et les pondérations des distributions gaussiennes [27].

- **Calcul des caractéristiques**

  Les caractéristiques sont extraites à partir des données d'entraînement à l'aide du modèle GMM. Ces caractéristiques sont généralement les probabilités d'appartenance aux différentes distributions gaussiennes.[22]

- **Entraînement du SVM**

  Les caractéristiques extraites sont utilisées comme entrées pour l'entraînement du SVM. Le SVM cherche à trouver l'hyperplan optimal qui sépare les données en fonction de leur classe.[22]

- **Classification**

  Une fois que le modèle hybride GMM-SVM est entraîné, il peut être utilisé pour la classification de nouvelles données. Les caractéristiques sont extraites à partir du modèle GMM et utilisées comme entrées pour le SVM, qui prédit la classe de la nouvelle donnée.[28]

2.8.Conclusion

En conclusion, l'apprentissage et la classification des données du locuteur à l'aide de la combinaison de modèles de mélange gaussien (GMM) et de machines à vecteurs de support (SVM) sont des approches prometteuses pour résoudre les problèmes liés à l'identification et à la vérification des locuteurs.

Le modèle de mélange gaussien permet de modéliser la distribution des caractéristiques acoustiques extraites des données vocales, en considérant que chaque locuteur peut être représenté par plusieurs composantes gaussiennes. Cela permet de capturer les

variations et les nuances de la voix d'un locuteur. Tandis que les SVM avec leur pouvoir de discrimination, ils séparent mieux les distributions gaussiennes de chaque classe (locuteur), l'une par rapport à l'autre

Chapitre 3

Vérification automatique du locuteur par GMM/SVM

3.1.Introduction

La vérification automatique du locuteur (VAL) est une technique utilisée pour vérifier l'identité d'un locuteur à partir de sa voix. Elle trouve de nombreuses applications dans les domaines de la sécurité, de l'authentification et de la reconnaissance vocale.

L'une des approches couramment utilisées pour la VAL est basée sur le modèle de mélange gaussien (GMM) combiné avec une machine à vecteurs de support (SVM). Le GMM est un modèle statistique qui représente la distribution de probabilité des caractéristiques acoustiques extraites de la voix d'un locuteur. Il est utilisé pour modéliser la voix du locuteur réel et construire un modèle de référence.

L'utilisation de GMM/SVM pour la VAL combine les avantages de la modélisation probabiliste (GMM) pour la représentation des caractéristiques acoustiques et de la classification binaire (SVM) pour la décision finale. Cette approche permet d'obtenir de bons résultats en termes de précision et de robustesse dans la vérification automatique du locuteur.

3.2..Architecture générale d'un système de vérification automatique du locuteur (VAL)

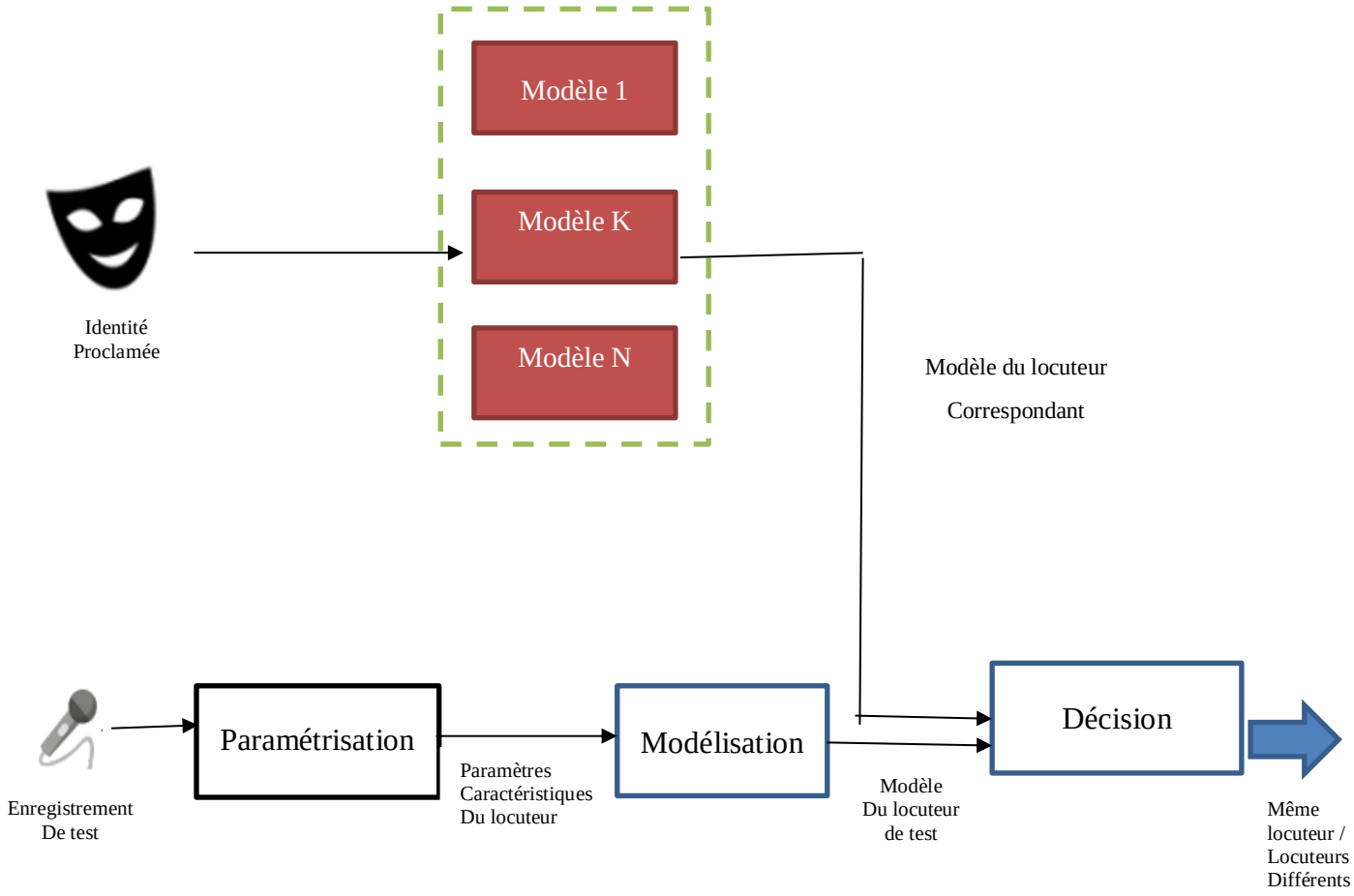

Figure 3.1 Structure d'un système de vérification du locuteur.

- **Paramétrisation (extraction des paramètres)**: Cette étape vise à capturer des paramètres caractéristiques de la parole d'une personne donnée. Suite à de nombreux travaux de recherche [29], il s'est avéré que les paramètres basés sur la représentation spectrale de la parole sont les plus pertinents pour la VAL. Ces paramètres sont corrélés à la forme du conduit vocal et sont les plus utilisés dans les systèmes de VAL modernes. Cependant, les paramètres prosodiques qui décrivent le style de parole du locuteur sont aussi utilisés en pratique.
- **Modélisation**: Les paramètres acoustiques extraits d'un enregistrement donné sont utilisés pour construire un modèle qui résume l'information acoustique correspondante.
- **Décision:** La phase de décision désigne l'identité du locuteur reconnu. Dans le cas de la vérification, cette décision est binaire et consiste à confirmer ou infirmer la correspondance de la session de test à une identité proclamée. Vu qu'il est impossible d'avoir une similarité de 100% entre le signal du locuteur de test et celui des locuteurs clients, les modèles sont conçus de telle sorte qu'une telle comparaison fournisse un score (une valeur scalaire) indiquant si les deux énoncés correspondent au même locuteur. Si ce score est supérieur (inférieur) à un seuil prédéfini, le système accepte (ou rejette) le locuteur de test.

3.3.Métriques d'évaluation des performances en VAL

L'évaluation des performances des systèmes de VAL est un processus délicat qui dépend de la tâche ciblée (vérification) et d'un nombre de paramètres (erreurs) qui peuvent influencer sa qualité ou fiabilité. Ces erreurs sont détaillées ci-dessous.

3.3.1 Types d'erreurs

Dans le cas d'un système de vérification du locuteur, deux types d'erreurs peuvent être observées :

- **Fausses acceptations (FA False acceptance)**: le cas où le système accepte le locuteur alors que celui-ci n'est pas la personne qu'il prétend être.
- **Faux rejets (FR False rejects)** : le cas où le système refuse l'accès à un locuteur alors qu'il correspond bien à l'identité proclamée.

Les taux d'erreurs correspondants ; FAR (taux de fausses acceptations) et FRR (taux de faux rejets) sont définis comme suit:

$$FAR = \frac{\#FA}{\#\ comparaisonimposteurs} \quad (3.1)$$

$$FRR = \frac{\#FR}{\#\ comparaisonclients} \quad (3.2)$$

- **Taux d'égale erreur (EER) :** Le taux d'égale erreur (EER ;EqualError Rate) est l'une des mesures les plus populaires en vérification de locuteurs vu qu'elle permet de comparer deux systèmes en se basant sur une seule mesure. Elle est définie comme le point opératoire où les valeurs FAR et FRR deviennent presque égales. Cette configuration est atteinte en faisant varier le seuil de décision τ jusqu'à ce que les deux zones correspondant aux fausses acceptations et aux faux rejets (voir la figure 3.2) deviennent égales. Il convient de noter que l'EER ne constitue pas nécessairement un point opératoire optimal dans les applications réelles qui peuvent nécessiter des niveaux de sécurité élevés.
- **Courbe DET (DetectionErrorTradeoff):**Lorsqu'il existe un compromis entre différents types d'erreurs (FA/FR), l'utilisation d'une seule mesure de performances peut s'avérer insuffisante pour représenter les capacités d'un système. En effet, les performances d'un système de vérification de locuteur peuvent être étudiées sur plusieurs points opératoires (operating points) et seraient mieux représentées par une courbe de performances.

  Traditionnellement, la courbe DET (DetectionErrorTradeoff ou courbe de détection de compromis) serait plus adaptée aux applications de vérification de locuteur affichant deux types d'erreurs sur les deux axes. Dans ces courbes, le taux de fausses acceptations (False positive rate) est tracé sur l'axe horizontal tandis que le taux de faux rejets (miss detection date) est tracé sur l'axe vertical.

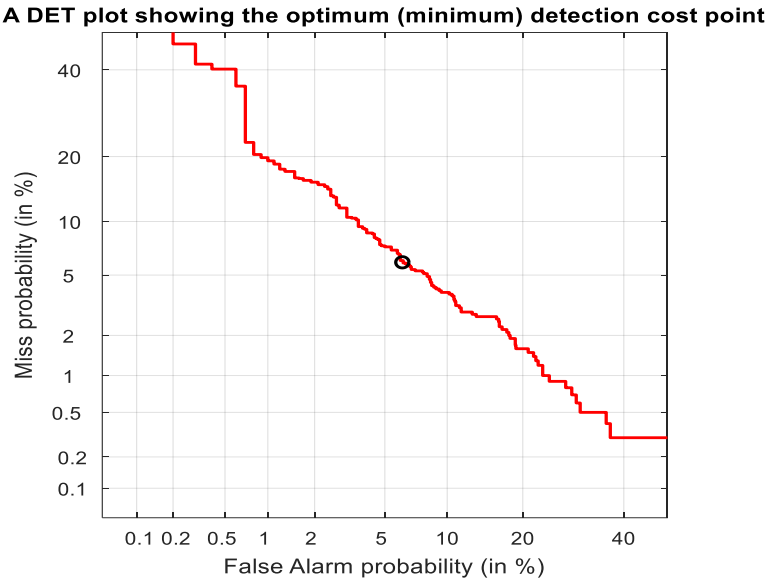

Figure 3.2 Courbe DET pour la VAL avec EER

3.4.Protocole expérimental

Dans cette section, la vérification du locuteur a été évaluée sur un corpus TIMIT composé de 168 locuteurs[30]. Le corpus TIMIT est une base de données contenant de la parole lue, échantillonnée à une fréquence de 16 kHz. Pour extraire les caractéristiques du locuteur, un modèle d'apprentissage basé sur un GMM (Mélange de Modèles de Gaussiennes) normalisé par UBM (Universal Background Model) a été utilisé. L'UBM utilisé dans cette étude est un GMM composé de 256 composantes, et il a été entraîné sur 64 minutes de parole. Le deuxième modèle SVM est conçu en utilisant un noyau RBF, donné par

$$\begin{cases} k(x, x_i) = e^{-\gamma \|x - x_i\|^2} \\ \gamma = 2 \end{cases} \quad (3.3)$$

Dans la phase de prétraitement, nous utilisons la méthode VAD pour détecter la présence ou l'absence des segments vocaux dans un signal de parole. En phase de paramétrage, nous spécifions l'espace des caractéristiques. En effet, comme le signal de parole est dynamique et variable, nous représentons les séquences d'observation de différentes tailles par des vecteurs de taille fixe. Chaque vecteur est donné par les différents types de caractéristiques extraites toutes les 10 ms, en utilisant une fenêtre de 25 ms. Les paramètres utilisés dans ce travail sont les coefficients : MFCC (de 12 à 22 coefficients) et leurs dérivées premières et secondes plus le paramètre de l'énergie.

3.5.Résultats expérimentaux

3.4.1 étude de l influence de paramètre MFCC sur les performance de la vérification de locuteur

Dans cette expérience on va faire une simulation des plusieurs valeurs MFCC de (12 à 22)
Pour obtenir la meilleur valeur pour continue notre simulation

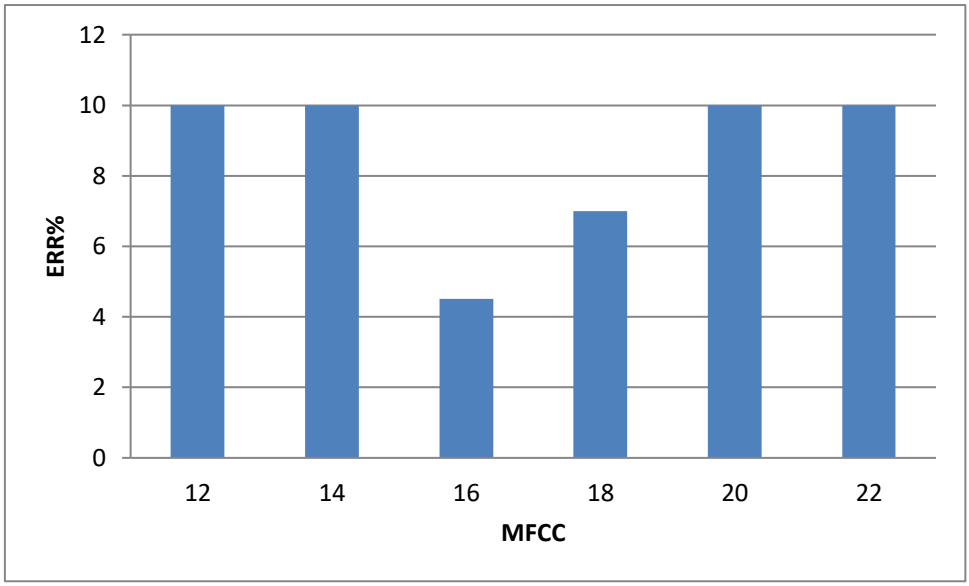

Figure 3.3 histogramme d'EER en fonction du nombre deMFCCs

❖ **Discussion**

L'histogramme d'EER en fonction du nombre de MFCCs (Mel FrequencyCepstral Coefficients) permet de visualiser comment l'EER varie à mesure que le nombre de MFCCs utilisés dans le système de vérification automatique du locuteur change.

En analysant cet histogramme, nous pouvons tirer les conclusions suivantes :

- Avec 12 MFCCs, l'EER est de 10%, ce qui signifie que le taux d'erreur est assez élevé. Cela suggère que la représentation des caractéristiques vocales avec seulement 12 MFCCs n'est pas suffisante pour obtenir une performance de vérification précise.
- Avec 14 MFCCs, l'EER reste à 10%, ce qui indique que l'ajout de deux MFCCs supplémentaires ne contribue pas à améliorer la performance de vérification.

- Lorsque le nombre de MFCCs est augmenté à 16, l'EER diminue à 4,5%. Cela suggère que l'ajout de MFCCs supplémentaires améliore la précision de la vérification.
- Avec 18 MFCCs, l'EER augmente légèrement à 7%, indiquant une légère dégradation de la performance par rapport à la configuration précédente.
- Lorsque le nombre de MFCCs est augmenté à 20 et 22, l'EER revient à 10%, ce qui suggère que l'ajout de MFCCs supplémentaires au-delà de 18 n'apporte pas d'amélioration significative à la performance.

cet histogramme met en évidence l'importance de choisir un nombre approprié de MFCCs pour obtenir une performance de vérification optimale. Dans cette expérience, une configuration de 16 MFCCs semble fournir les meilleurs résultats, avec un EER de 4,5%. Cependant, il est important de noter que ces résultats sont spécifiques à l'ensemble de données et au système utilisés, et peuvent varier dans d'autres contextes.

3.4.2 étude de l influence des paramètres prosodiques

Dans cette expérience on va faire étude de l influence des paramètres prosodiques avec la valeur de MFCC 16 qu'on a l obtenu de l'expérience précédente

Voici l'histogramme ci dessue

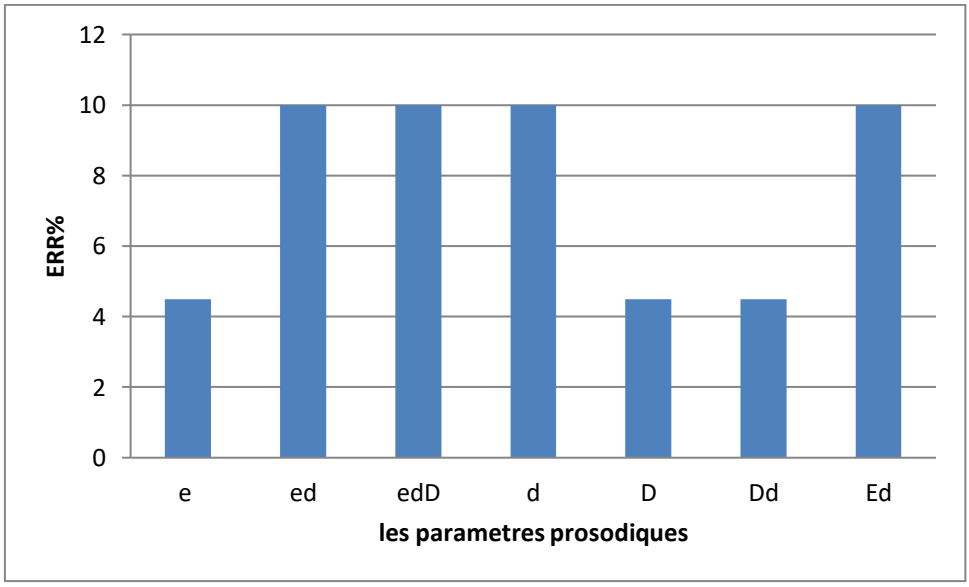

Figure 3.4 histogramme des taux d'erreur en fonction des paramètres prosodiques fixe sur MFCC16

❖ **Discussion**

Le taux d'erreur en fonction des paramètres prosodiques dans le contexte de GMM/SVM est une mesure de la performance d'un système de vérification automatique du locuteur utilisant la modélisation par mélange gaussien (GMM) et la machine à vecteurs de support (SVM) avec différents paramètres prosodiques.

En analysant cet histogramme, nous pouvons tirer les conclusions suivantes :

Les paramètres prosodiques "ed", "edD", "d" et "eD" ont tous un EER de 10%. Cela indique que ces paramètres ne contribuent pas à améliorer la performance de la vérification automatique du locuteur par rapport à la configuration de base de 16 MFCCs.

Les paramètres prosodiques "e", "D" et "dD" ont un EER de 4,5%. Ces paramètres semblent apporter une amélioration significative à la performance de vérification par rapport à la configuration de base.

Cet histogramme met en évidence l'impact des paramètres prosodiques sur l'EER dans la configuration de 16 MFCCs. Les paramètres "e", "D" et "dD" présentent des taux d'erreur plus bas, suggérant qu'ils peuvent être des paramètres prosodiques plus discriminants pour la vérification du locuteur. On a choisi la valeur "e" comme meilleure valeur pour continue notre simulation

3.4.3 étude de l influence des types des SVM

Dans cette série d'expériences, nous avons fixé le nombre de coefficients MFCC à 16 et avons utilisé le type "e" des paramètres prosodiques. L'objectif était de calculer le pourcentage d'erreur en utilisant différents types de SVM.

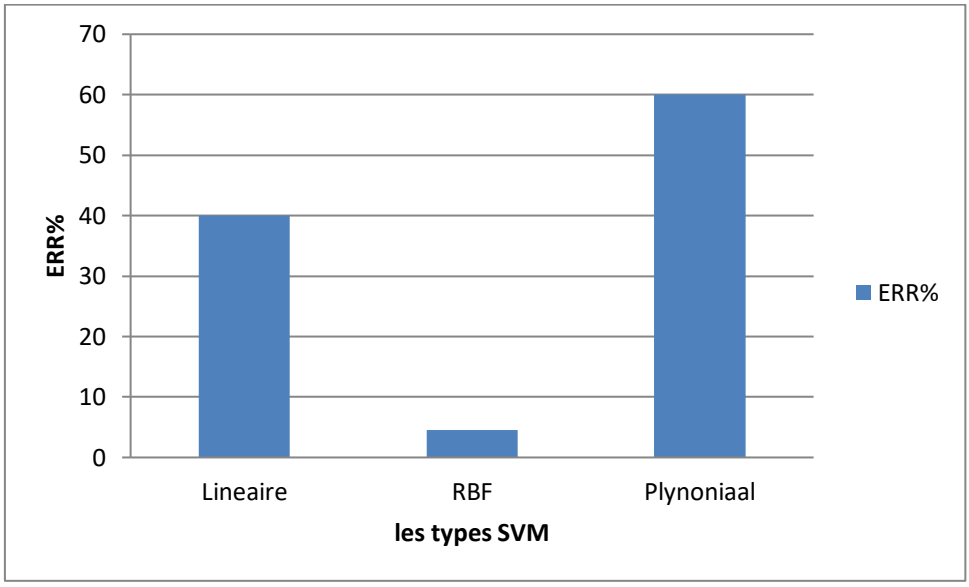

Figure 3.5 histogramme d'EER en fonction de type de noyau SVM

❖ **Discussion**

L'histogramme fournit des informations sur les performances de différents types de SVM (Support Vector Machine) en termes de taux d'erreur de classement (EER) pour différentes configurations.

- Pour le type de SVM linéaire, le taux d'erreur de classement (EER) est de 40%.
- Pour le type de SVM -RBF (Radial Basis Function), le taux d'erreur de classement (EER) est de 4.5%.
- Pour le type de SVM polynomial, taux d'erreur de classement (EER) est de 80%.

D'après les informations fournies, le noyau RBF semble offrir de meilleures performances avec un taux d'erreur de classement (EER) de 4.5%

3.4.4 L'étude des préférences de système de vérification (ERR%) en fonction de gamma et C

A la base des donnes obtenus de dernier expériences, la valeur MFCC a 16 et le paramètre « e » des paramètres prosodiques et le type RBF de noyau. Dans cette simulation on va calculer le pourcentage d'err on fonction de gamma et c comme le tableau suivant

Gamma / C	1	2	3	4	5	6	7	8	9	10
10	1.2%	1.2%	0.6%	0.9%	1.3%	1.7%	1.9%	2%	3%	3%
40	1.2%	4%	0.4%	0.5%	0.5%	1%	1.9%	2%	2.5%	10%
80	1.2%	3.5%	0.4%	0.5%	0.7%	0.8%	1.9%	3%	4%	10%
100	1.2%	3.5%	0.4%	0.4%	0.7%	1.5%	2.2%	3%	4%	10%
200	2.5%	4.5%	0.3%	0.5%	1.3%	2.2%	4.5%	6%	7%	5.5%
400	4%	3.5%	0.4%	0.9%	3%	4.5%	5.2%	9%	5.5%	8%
600	4%	2.5%	0.4%	1.6%	4%	4.5%	5.2%	10%	5.5%	10%
800	2.5%	4%	0.5%	2%	3.5%	6%	8%	9%	9%	10%
1000	4%	2.5%	0.7%	2.2%	3.5%	6%	9%	7%	12%	13%

Tableau 3.1représente les performances du modèle GMM/SVM en termes d'ERR en fonction des valeurs de gamma et de C

- **Discussion**

Le tableau présenté montre les performances du modèle GMM/SVM en fonction des valeurs de gamma (1 à 10)et de C (de 10 ,40,80,100,200,400,600,800,1000) pour calculer l'erreur EER. La valeur la plus basse du tableau, qui est 0.3%, semble être la meilleure performance obtenue pour notre teste.

Il est important de noter que la valeur de 0.3%, se trouve à l'intersection de la ligne correspondant à gamma = 3 et de la colonne correspondant à C = 200. Cette combinaison spécifique semble produire les meilleures performances selon les métriques mesurées. Alors on va prendre cette valeur pour continuer les prochaines expériences.

3.4.5 étude de composante du modèle du mode UBM

D'âpres les expériences présidents on a , la valeur MFCC a 16 et le paramètre « e » des paramètres prosodiques et le type RBF de noyau, et gamma=3 et c =200 dans cette simulation on va voir le pourcentage d'EER en fonction de nombre de composante du modèle du monde UBM

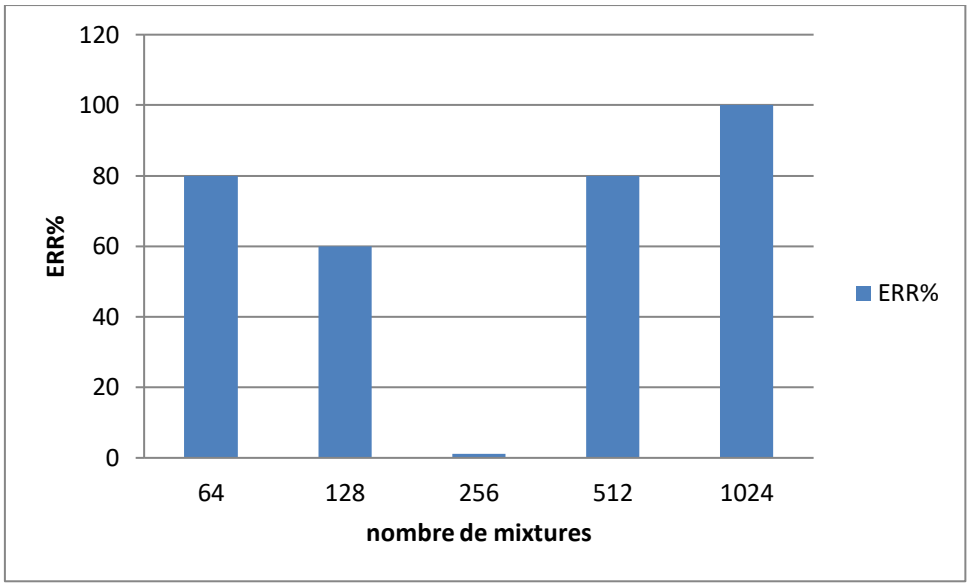

Figure 3.6 histogramme d'EER en fonction de nombre de composante du modèle du mode UBM

❖ **Discussion**

L'histogramme fournit des informations sur les performances d'un modèle de mélange de gaussiennes (GMM-UBM) combiné avec SVM, sous différents nombres de mixtures, mesurées en termes de taux d'erreur (ERR %).

- Pour un modèle GMM-UBM avec 64 mixtures, le taux d'erreur est de 80%.
- Pour un modèle GMM-UBM avec 128 mixtures, le taux d'erreur est de 60%.
- Pour un modèle GMM-UBM avec 256 mixtures, le taux d'erreur est de 0.3%.
- Pour un modèle GMM-UBM avec 512 mixtures, le taux d'erreur est de 80%.
- Pour un modèle GMM-UBM avec 1024 mixtures, le taux d'erreur est de 100%.

D'après les informations fournies, le modèle GMM-UBM avec 256 mixtures semble offrir les meilleures performances avec un taux d'erreur de 0.3%

3.4.6 étude d'EER en fonction du nombre de locuteurs

On a la valeur MFCC a 16 et le paramètre « e » des paramètres prosodiques et le type RBF de noyau, et gamma=3 et c =200 et le modèle GMM-UBM avec 256 mixtures, ces valeurs sont les meilleures obtenu de les expériences présidents pour faire notre dernière simulation de calculé l'EER en fonction du nombre de locuteurs .

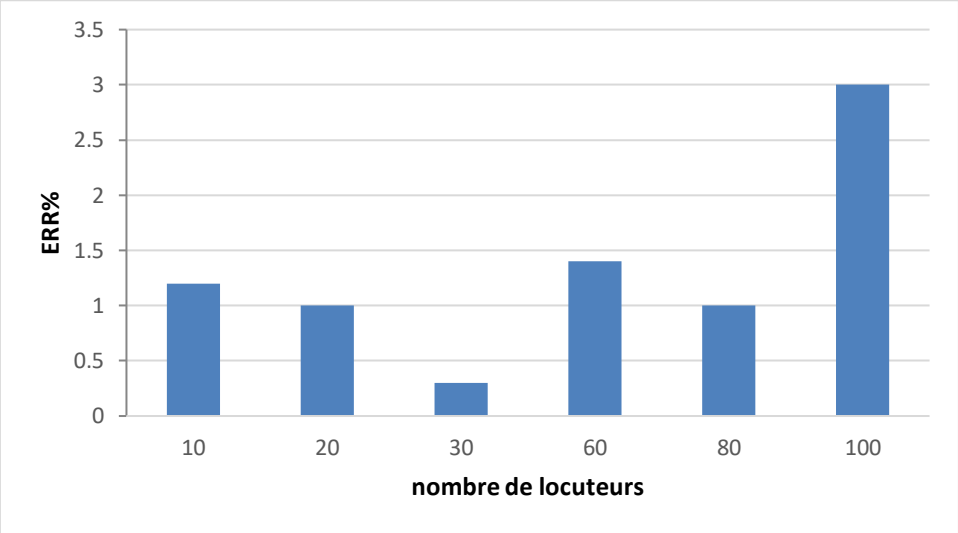

Figure 3.7.histogramme d'EER en fonction du nombre de locuteurs

❖ **Discussion**

L'histogramme fournit des informations sur les performances d'un système de vérification du locuteur basé sur GMM/SVM, en fonction du nombre de locuteurs dans les données et mesure le taux d'erreur (EER%).

Dans ce cas, on peut observer que le système de vérification de locuteurs fonctionne relativement bien pour la plupart des configurations de nombre de locuteurs, avec des taux d'erreur allant de 0.12% à 3%.

Le nombre de locuteur 30 est le meilleur, car le système de vérification de locuteurs (GMM/SVM) semble trouver toutes les informations nécessaires pour bien discriminer entre les locuteurs.

3.6. Conclusion

Cette étude a amélioré la vérification du locuteur en optimisant les paramètres tels que le nombre de MFCCs, les paramètres prosodiques, le type de SVM et les composantes du modèle UBM. Les résultats les plus significatifs ont été obtenus avec une configuration de 16 MFCCs, les paramètres prosodiques "e" et le noyau RBF du SVM. Grâce à ces paramètres optimaux, le taux d'erreur du système de vérification du locuteur a été réduit à 0,3%. Ces résultats démontrent l'efficacité du système dans la discrimination précise des locuteurs. Ces expériences sont basées sur les conditions spécifiques de l'étude et peuvent varier dans d'autres contextes. Les améliorations apportées aux paramètres ouvrent la voie à de futures améliorations et applications pratiques dans ce domaine de recherche.

Conclusion Générale

Conclusion générale

Le traitement et l'analyse de la parole pour la vérification du locuteur sont des domaines de recherche et d'application qui visent à identifier et à vérifier l'identité d'un locuteur à partir des caractéristiques acoustiques de sa voix, L'objectif de notre travail était d'étudier la performance de Vérification automatique du locuteur par GMM/SVM, en utilisant des techniques basées sur les modèles de mélange gaussien (GMM) et les machines à vecteurs de support (SVM). Notre étude visait à évaluer la précision et la fiabilité de ce système de vérification automatique du locuteur dans des scénarios réels.

Nous avons réalisé des expériences en utilisant des enregistrements audio de locuteurs multiples, en extrayant les caractéristiques acoustiques pertinentes telles que les MFCC (Mel-frequency cepstral coefficients), et en entraînant le modèle GMM-SVM sur ces données. Nous avons ensuite évalué les performances du système en mesurant des métriques telles que le taux de réussite d'identification et de vérification du locuteur.

Les résultats de notre étude ont démontré que le système de vérification automatique du locuteur basé sur GMM/SVM est efficace et précis dans la tâche d'identification et de vérification des locuteurs. Nous avons obtenu des meilleurs résultats avec plus bas pourcentage d'erreur, ce qui indique que le système est capable de distinguer de manière fiable les différentes voix et de vérifier l'identité des locuteurs.

Références

Références

[1] Prince, S. J., & Elder, J. H. (2007). Probabilistic linear discriminant analysis for inferences about identity. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1-8).

[2] Burget, L., Plchot, O., Cumani, S., Glembek, O., Matějka, P., & Brümmer, N. (2011, May). Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4832-4835). IEEE

[3] Kent, R. D. (1997). The biology of phonation. Journal of Speech, Language, and HearingResearch, 40(3), 493-509

[4] Levelt, W. J. (1999). Producingspokenlanguage:ablueprint of the speaker. In Cognition, Vol. 6 (pp. 201-259). Elsevier.

[5] Ladefoged, P., & Johnson, K. (2014). A course in phonetics. Cengage Learning

[6]Hickok, G., &Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience, 8(5), 393-402.

[7] Marslen-Wilson, W. D., & Tyler, L. K. (2007). Morphology, language and the brain: the decompositionalsubstrate for languagecomprehension. Philosophical Transactions of the Royal Society B:Biological Sciences, 362(1481), 823-836

[8] Lamel, L. F., &Roach, P. J. (2011). The Sounds of Language: An Introduction to Phonetics. Pearson Education Limited.

[9] Rabiner, L. R., &Juang, B. H. (1993). Fundamentals of Speech Recognition. Prentice Hall.

[10] Huang, X., Acero, A., & Hon, H. W. (2001). Spokenlanguageprocessing: A guide to theory, algorithm, and system development. Prentice Hall

[11] Lyons, R. G. (2011). Understanding digital signal processing (3rd ed.). Prentice Hall..

[12] Ladefoged, P., & Johnson, K. (2011). A course in phonetics. Cengage Learning

[13] "Fourier Analysis: An Introduction

[14]Rabiner, L. R., &Schafer, R. W. (2011). Theory and applications of digital speech processing. Prentice Hall.

[15] "Digital Processing of Speech Signals" de L.R. Rabiner et R.W. Schafer

[16] Stevens, K. N. (1998). AcousticPhonetics. MIT Press.

[17] Ladefoged, P., & Johnson, K. (2015). A course in phonetics (7th ed.). WadsworthCengage Learning

[18] Deller, J. R., Jr., Hansen, J. H. L., &Proakis, J. G. (2013). Discrete-time processing of speech signals. Springe

[19]Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition : fromfeatures to supervectors. Speech communication, 52(12), 12-40.

[20] Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1, I-729.

[21] Thèse de Doctorat préparée par Mr SayoudHalim, sous la direction de Mme Malika Boudraa,

[22] Reynolds, D. A., & Rose, R. C. (1995). Robusttext-independent speaker identificationUsingGaussian mixture speaker models. IEEE Transactions on Speech and AudioProcessing, UsingGaussian mixture models. IEEE Transactions on Speech and Audio Processing,

[23] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning:

Data Mining, Inference, and Prediction, Second Edition. Springer-Verlag.

[24] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verificationusingAdaptedGaussian mixture models. Digital signal processing,

[25] Chen, Y., Chu, M., Chang, E., Liu, J., & Liu, R. (2003, September). VoiceConversionwithsmoothed GMM and MAPadaptation.Ininterspeech.

[26]Abe, S. (2005). *Support vector machines for pattern classification* (Vol. 2, p. 44). London: Springer.

[27] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. Chapter 9 of this book provides a detailedexplanation of Gaussian Mixture Models,Includingtheir training using the Expectation-Maximization (EM) algorithm

[28] Campbell, W. M., & Sturim, D. E. (1996). Support vector machines using GMM supervectors for speaker verification. In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

[29] J. Bonastre& H. Meloni, 1992. A study of spectral variability for speaker characterisation. 19èmes Journées d'Etudes sur la Parole 555.

[30] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., &Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/ReconTechnical Report N*, *93*, 27403.