REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

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

Faculté *des Sciences et de la technologie*

Département d'Electronique

Mémoire

Présenté pour obtenir

LE DIPLOME DE MASTER

FILIERE : **TELECOMMUNICATION**

Spécialité : Réseaux et Technologie de Télécommunication

Intitulé

Reconnaissance Automatique de Locuteurs en Sciences Forensiques (Criminalistiques)

Rédigé par :

- **Guemmour Sara**
- **Bouzidi Selma**

Encadré par :

Dr N.ASBAI

Soutenu publiquement le 07/09/2019 devant le jury composé de :

Dr AIDEL S	**Président**
Dr MEZAACHE S/E	**Examinateur**

Année Universitaire 2018/2019

Table des matières

Chapitre 3 : Évaluation Expérimentale du Système RALF

Liste des Figures

Chapitre 1

Chapitre 2

Chapitre 3

Liste des Tableaux

Abréviations

FSR	Forensique speech Reconnaissance.
RALF	Reconnaissance automatique de locuteurs en sciences forensiques.
RAL	Reconnaissance Automatique du Locuteur.
LR	Rapport de vraisemblance ou Likelihood ration
GMM	Gaussian Mixtures Models.
MFCC	Mel-Frequencies Cepstral Coefficients.
TPZ	Taux de passage par zéro.
VAD	Détection de l'Activité Vocale.
FFT	Fast Fourrier Transform.
IDCT	Inverse Discrete Cosinus Transform.
IFFT	Inverse Fast Fourrier Transform.
Fe	La fréquence d'échantillonnage.
Fmax	La fréquence maximum.
EM	Expectation Maximization.
MLE	Maximum Likelihood Estimation.
UBM	Universal Background model.
NIST	National Institute of Standards and Technology.
TIMIT	Texas Instruments Massachusetts Institute of Technology.
TIMITF	Texas Instruments Massachusetts Institute of Technology of feminine.
TIMITM	Texas Instruments Massachusetts Institute of Technology of male.
SVM	Machines à support vecteurs.

Résumé

Un système de reconnaissance automatique du locuteur en sciences forensiques (RALF) est mis en œuvre pour identifier correctement un suspect dans le cadre d'une simulation d'enquête policière ou judiciaire, à l'aide d'enregistrements vocaux. En effet, on peut facilement capturer des traces vocales, qui peuvent être analysées au moyen d'un système de reconnaissance automatique du locuteur, et par conséquent, aider le tribunal à prendre une décision. Dans notre travail, nous avons utilisé l'interprétation bayésienne pour calculer le rapport de vraisemblance (LR) qui pondère la preuve en faveur de deux hypothèses contradictoires : 1) le locuteur suspect est la source de l'enregistrement interrogé (trace), 2) le locuteur à l'origine de l'enregistrement interrogé n'est pas le locuteur suspect, avec l'adaptation des modèles de mélange gaussien (GMM) pour les locuteurs utilisant le modèle de fond universel (UBM).

Les expériences réalisées montrent que le système d'identification forensique du locuteur est très intéressant et peut aider énormément à résoudre des problèmes criminalistiques. En effet, nous avons eu des résultats très promoteurs dans différents scénarios de simulation. Nous avons aussi montré que les conditions d'enregistrement et les supports de transmission ont une grande influence sur les performances d'un système de reconnaissance forensique du locuteur.

Mots clés : Reconnaissance automatique du locuteur en sciences forensiques RALF, Identification forensique, interprétation Bayesienne, Modèle de mélange de gaussiennes GMM, rapport de vraisemblance LR, Modèle de fond universel UBM.

Abstract

A Forensic Automatic Speaker Recognition System (FASR) is used to correctly identify a suspect as part of a police or judicial simulated investigation, using voice recordings. Indeed, voice traces can be easily captured, which can be analyzed using an automatic speaker recognition system, and thus help the court to make a decision. In our work, we used Bayesian interpretation to calculate the likelihood ratio (LR), that weights the evidence in favor of two contradictory hypotheses: 1) the suspect speaker is the source of the questioned record (trace), 2) the speaker at the origin of the record being interrogated is not the suspect speaker, with the adaptation of Gaussian Mixture Models (GMM) for speakers using the Universal Background Model (UBM).

The experiments carried out show that the forensic identification system of the speaker is very interesting and can help enormously to solve criminalistic problems. Indeed, we had very promising results in several different scenarios. We have also noticed that recording conditions and transmission media have a great influence on the performance of a forensic speaker identification system.

Keywords: Automatic forensic recognition, Forensic identification, Bayesian interpretation, Gaussian mixture model GMM, likelihood ratio LR, universal background model UBM.

Dédicace

A l'homme de ma vie, mon exemple éternel, mon soutien moral et source de joie et de bonheur, celui qui s'est toujours sacrifié pour me voir réussir, que dieu te garde dans son vaste paradis, à toi mon père.

A la lumière de mes jours, la source de mes efforts, la flamme de mon cœur, ma vie et mon bonheur ; maman que j'adore.

Mes meilleures sœurs Nacima, Amel et Hadjira Et leurs petites filles et Mon beau frère Houssem. Pour leurs soutiens moral et leurs conseils précieux tout au long de mes études.

A mon cher ilyes, qui m'a aidé et supporté dans les moments difficiles.

Aux personnes qui m'ont toujours aidé et encouragé, qui étaient toujours à mes côtés, et qui m'ont accompagnaient durant mon chemin d'études supérieures, mes aimables amis : Abla, Linda, Rania, Mallak, Nadjat, Faiza et Amina, tous mes collègues d'étude.

A mon binôme Sara et toute les familles BOUZIDI et AZZOUG. Et à tous ceux qui ont contribué de près ou de loin pour que ce projet soit possible, Tous ceux que j'aime dans le monde. Je vous dis merci.

Selma

A l'homme de ma vie, mon exemple éternel, mon soutien moral et source de joie et de bonheur, celui qui s'est toujours sacrifié pour me voir réussir, que dieu te garde dans son vaste paradis, à toi mon père.

A la lumière de mes jours, la source de mes efforts, la flamme de mon cœur, ma vie et mon bonheur ; maman que j'adore.

Mes meilleures sœurs Asma et Leila Et leurs petites filles et Mes chers frères Redouane et Yasser.

Je tiens à remercier Mon fiancer Djamel Eddine surtout pour son soutien moral ininterrompu et ses nombreux conseils tout au long de la réalisation de ce travail.

Aux personnes qui m'ont toujours aidé et encouragé, qui étaient toujours à mes côtés, et qui m'ont accompagnaient durant mon chemin d'études supérieures, mes aimables amis : Abla, Linda, Rania, Mallak, Nadjat, et Faiza, tous mes collègues d'étude.

A mon binôme Selma et toute les familles GUEMMOUR et MEROUNE. Et à tous ceux qui ont contribué de près ou de loin pour que ce projet soit possible, Tous ceux que j'aime dans le monde. Je vous dis merci.

Sara

Remerciements

REMERCIEMENTS

Nous remercions en premier lieu le Dieu le tout puissant. C'est grâce à lui que nous avons eu la foi et la force pour accomplir ce travail.

Nous voulons remercier sincèrement Dr. N.ASBAI, Docteur à l'université de BBA, d'abord en tant qu'encadreur de ce mémoire ensuite pour ses précieux conseils, ses incessants encouragements et surtout sa grande disponibilité tout au long de la réalisation de ce travail. Nous le remercions pour toute la confiance accordée à notre égard. Ainsi que pour l'inspiration, l'aide et le temps qu'il a bien voulu nous consacrer sans quoi ce mémoire n'aurait jamais eu autant de succès.

A tous les membres de jury, Vous nous faites le grand honneur en acceptant de juger notre modeste travail, veuillez trouver ici l'expression de nous sincères gratitudes et notre grand respect.

Finalement, nous tenons à remercier tous ceux qui ont contribué de près ou de loin à la finalisation de notre travail.

Introduction Générale

Introduction Générale

Les sciences forensiques constituent l'ensemble des principes scientifiques et des méthodes techniques appliquées à l'investigation criminelle, pour prouver l'existence d'un crime et aider la justice (le juge et / ou le jury) à déterminer l'identité de l'auteur et son mode opératoire à partir d'une mesure quantitative de la valeur des preuves [1] [2]. La reconnaissance automatique forensique du locuteur **RALF,** est un terme générique pour discriminer parmi plusieurs personnes à partir des échantillons de leurs voix [3] [4], afin de trouver les suspects d'être sensés la source de la trace laissée dans une scène de crime. Il convient dans ce domaine de recherche, de reconnaître non pas ce qui a été dit, mais de reconnaître l'identité de la personne qui parle, à partir de ses caractéristiques vocales [11].

L'avantage des systèmes **RAL** est qu'ils sont indépendants du texte, Indépendants de la langue du discours, et la reconnaissance du locuteur est totalement automatisée et ne nécessite aucune intervention humaine [5].

L'objectif principal de ce travail est :

- étudier et évaluer un système d'identification automatique du locuteur en criminalistique [6], en utilisant le modèle **GMM** (Gaussian Mixture Model).
- définir un cadre d'interprétation par l'expert d'un indice que représente l'enregistrement d'une voix peut être envisagé en termes de rapports de vraisemblance de deux hypothèses compétitives, selon une approche Bayésienne [7]. Cette approche nécessite la création de plusieurs bases de données pour construire un modèle **UBM** (Universal Backgroud Model), qui est très intéressant pour la bonne estimation des modèles statistiques des locuteurs [14].
- Examiner les méthodes existantes de prétraitement acoustique, par l'analyse cepstral **MFCC** (Mel Frequency Cepstral Coefficients), avec ces derniers en estimant des modèles **GMM (**Gaussian Mixture Models) [8]. Ce dernier (GMM) utilise la procédure **EM** (Expectation Maximization) pour dériver les modèles probabilistes du locuteur [9] [12] et d'évaluation de la similarité à utiliser pour toute donnée de parole spécifique du locuteur, et indiquer comment ces méthodes sont mises en œuvre pour calculer le rapport de vraisemblance **LR** en tant qu'élément de preuve [10].

Ce mémoire est composé de trois chapitres, organisés comme suit : dans le premier chapitre, nous présentons des généralités d'un système de reconnaissance automatique du

locuteur en sciences forensiques, ainsi l'Interprétation Bayésienne pour évaluer la valeur de la preuve et trouver la valeur de rapport de vraisemblance. Dans le deuxième chapitre, nous avons présenté un type de paramètres acoustiques utilisés dans les systèmes RALF qui sont MFCCs, suivi d'une présentation du modèle statistique le plus utilisé dans les systèmes de reconnaissance automatique du locuteur à savoir, le modèle **GMM** (Gaussian Mixture Model). Le dernier chapitre contient l'ensemble des tests effectués et les résultats que nous avons obtenus.

Chapitre 1

Introduction à la Reconnaissance Forensique du Locuteur

Chapitre 1

Introduction à la reconnaissance forensique du locuteur

1.1. Introduction

La Reconnaissance Forensique du Locuteur (FSR) est le processus permettant de déterminer si un individu spécifique (locuteur suspect) est la source d'un enregistrement vocal mis en cause (trace). Le rôle de l'expert forensique est de témoigner de la valeur de la preuve vocale en utilisant, si possible, une mesure quantitative de cette valeur. Il appartient au juge et / ou au jury d'utiliser ces informations pour faciliter leurs délibérations et leur décision. Ce chapitre a pour objectif de présenter les avancées de la recherche en matière de reconnaissance automatique de locuteurs par la police scientifique, y compris des outils pilotés par les données et une méthodologie qui offrent un moyen cohérent de quantifier et de présenter la voix enregistrée comme preuve biométrique, ainsi que l'évaluation de sa force (rapport de probabilité) dans le cadre d'interprétation bayésien, compatible avec les interprétations dans d'autres disciplines de la criminalistique. Ce chapitre contient des instructions étape par étape pour le calcul de la preuve biométrique et de sa force dans les conditions de fonctionnement du système forensique.

1.2. La reconnaissance automatique de locuteurs en sciences forensiques (RALF)

La Reconnaissance Automatique du Locuteur (RAL) est un moyen permettant de discriminer plusieurs personnes en fonction de leurs voix [11] [12]. Dans ce domaine de recherche, il convient de ne pas reconnaître ce qui a été dit, mais de reconnaître l'identité de la personne qui parle, à partir de ses caractéristiques vocales. L'identification et la vérification du locuteur sont généralement spécifiées [13] [11]. L'identification consiste à reconnaître un locuteur appartenant à une population de plusieurs locuteurs ; on y compare son expression vocale à des références connues. La vérification consiste à accepter ou à refuser une identité proclamée par un locuteur ; à cette fin, nous comparons à un certain seuil la distance entre son expression vocale et sa référence personnelle [11].

La RAL peut regrouper trois catégories principales : applications en contrôle d'accès sur sites sensibles [12], application dans le domaine sécuritaire et juridiques [12] et applications dans les systèmes de communication [12].

La RAL en science forensique consiste en l'utilisation de dispositifs scientifiques pour résoudre les problèmes et les exigences du tribunal en matière d'actes répréhensibles ou de poursuites conjointes [14]. La RAL en forensique (RALF) regroupe l'ensemble des différentes méthodes d'analyse fondées sur les sciences : biologie, informatique, mathématique et statistique, afin de servir au travail d'investigation de manière large. Par exemple, les méthodes de reconnaissance biométrique, tel que l'analyse par l'empreinte vocale. Dans ce mémoire, nous nous intéressons aux méthodes d'identification d'un enregistrement vocal.

L'analyse de la voix est utilisée pour des investigations forensiques, car un suspect peut laisser des enregistrements vocaux sur le téléphone, le message vocal, répondre au courrier, ou dans un enregistreur caché, et à partir de là, il tend à être utilisé comme preuve [14].

La RALF est un système de classification de données de la voix [15], dans lequel, la machine a pour tâche d'extraire l'information du signal vocal qui caractérise les spécificités d'un individu : identité, caractéristiques physiques, L'émotion, particularités régionales,…etc [11].

Dans le système RAL forensique, les modèles statistiques des paramètres audio du signal de parole du locuteur (suspect) sont comparés aux paramètres audio extraits de l'enregistrement sonore en question (trace) [16]. Les similitudes entre les paramètres audio extraits de l'enregistrement (ou trace) concerné et ceux extraits de l'enregistrement du suspect, représentés par son modèle statistique, sont calculés pour évaluer les preuves [16].

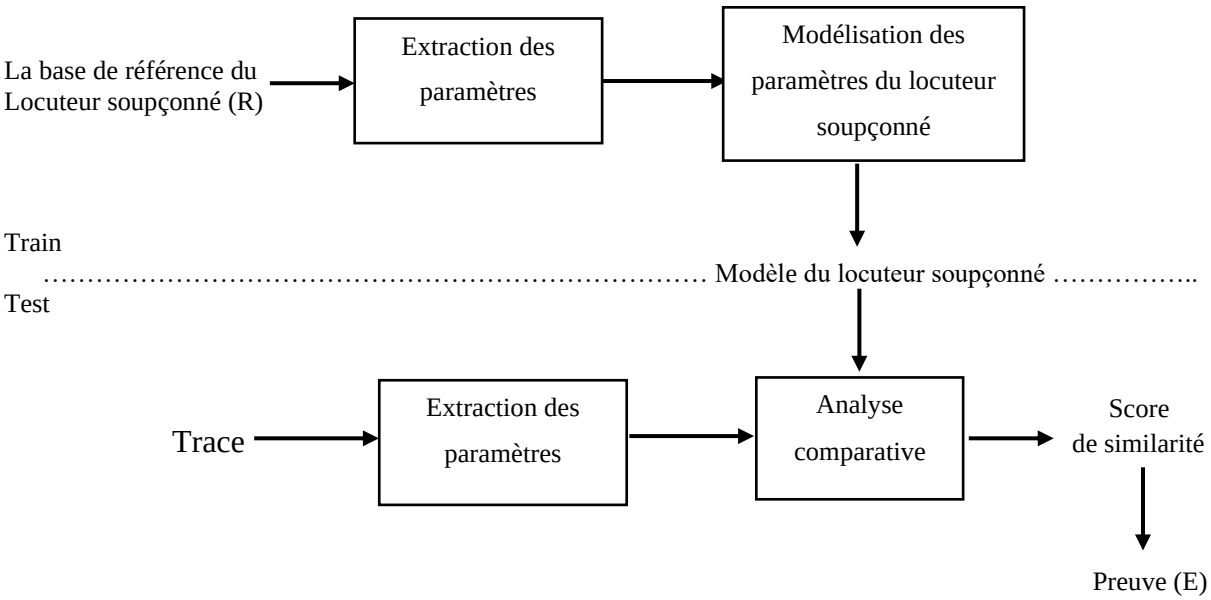


Fig.1.1. chaine de traitement pour le calcul de la preuve E.

1.3. Interprétation Bayésienne pour RALF

Les travaux de recherche préliminaires prouvent qu'un modèle probabiliste (le théorème de Bayes) est un outil adéquat pour aider les scientifiques à évaluer la valeur des preuves scientifiques. Il aide les juristes à interpréter les preuves scientifiques et à clarifier les rôles respectifs des scientifiques et des membres du tribunal [16].

1.3.1. Calcul de la preuve

La preuve E est le résultat de l'analyse comparative des caractéristiques dépendantes du locuteur (x) extraites de l'enregistrement interrogé (X) (trace), et les caractéristiques dépendantes du locuteur (y) étant extraites des signaux de parole du suspect (Y) [16].

1.3.2. Définition de la probabilité

La forme de probabilité du théorème de Bayes montre comment de nouvelles données (trace) peuvent être combinées à des connaissances de base antérieures (probabilités à priori) pour donner des probabilités postérieures à un résultat judiciaire [17]. Cela permet à l'expert légiste de réviser la mesure de probabilité d'incertitude sur la base de nouvelles informations, en calculant le rapport de vraisemblance (**LR**) de la preuve compte tenu de la paire d'hypothèses concurrentes :

H0: le locuteur suspect est la source de l'enregistrement interrogé (trace),

H1: le locuteur à l'origine de l'enregistrement mis en cause n'est pas le locuteur suspect.

1.3.3. Le rapport de vraisemblance (LR)

Dans le cadre du rapport de vraisemblance, la RALF a pour tâche de fournir au tribunal une déclaration relative à la force de la preuve en réponse à la question suivante :
« Quelle est la probabilité que les propriétés observées de la voix sur l'enregistrement du locuteur interrogé (la preuve) aient été produites par le locuteur connu (hypothèse du même locuteur) par rapport à un autre locuteur choisi au hasard parmi les population (l'hypothèse de différents locuteurs) ? » [18].

La réponse à cette question est exprimée quantitativement en tant que rapport de vraisemblance, calculé à l'aide de la formule suivante :

$$\frac{P(H_0|E)}{P(H_1|E)} = \frac{P(E|H_0)}{P(E|H_1)} \times \frac{P(H_0)}{P(H_1)} \quad (1.1)$$

Rapport de probabilité a postérieure	Rapport de vraisemblance	Rapport de probabilité à priori
(Provient de la cour)	(Provient de l'expert)	(Provient de la cour)

Avec les hypothèses $\mathbf{H_0}$ et $\mathbf{H_1}$, le numérateur de **LR**, c'est-à-dire la probabilité $\mathbf{P(E|H_0)}$, correspond à une affirmation numérique sur le degré de similitude de la preuve par rapport au suspect et son dénominateur, c'est-à-dire la probabilité $\mathbf{P(E|H_1)}$, correspond à un nombre indiquant le degré de typicité par rapport à la population concernée. Le **LR** est le rapport entre ces deux déclarations.

La méthodologie utilisant le cadre d'interprétation bayésien concerne le prétraitement de la parole, l'extraction des caractéristiques, la modélisation des caractéristiques (pour créer des modèles du locuteur), l'évaluation de la similarité et le calcul du rapport de vraisemblance (**LR**) [19].

La preuve de parole observée (R) peut être définie au niveau de l'extraction de caractéristiques ou du score de similarité. Sur la base de ce choix, le calcul d'un LR peut être effectué selon la « **méthode directe** » ou la « **méthode de score** » [19].

La valeur d'un rapport de vraisemblance dépend essentiellement des choix que l'on fait pour énoncer la preuve de parole observée (R) et les hypothèses H_0 et H_1, avec les modèles correspondants de variabilité intra-locuteurs et interlocuteurs [14]. Cela dépend également de plusieurs aspects de l'analyse de la parole, y compris les aspects non automatiques de l'étape de préparation des données, le type de caractéristiques, les modèles de caractéristiques et les scores de similarité utilisés, ainsi que les bases de données utilisées dans le calcul de la preuve E.

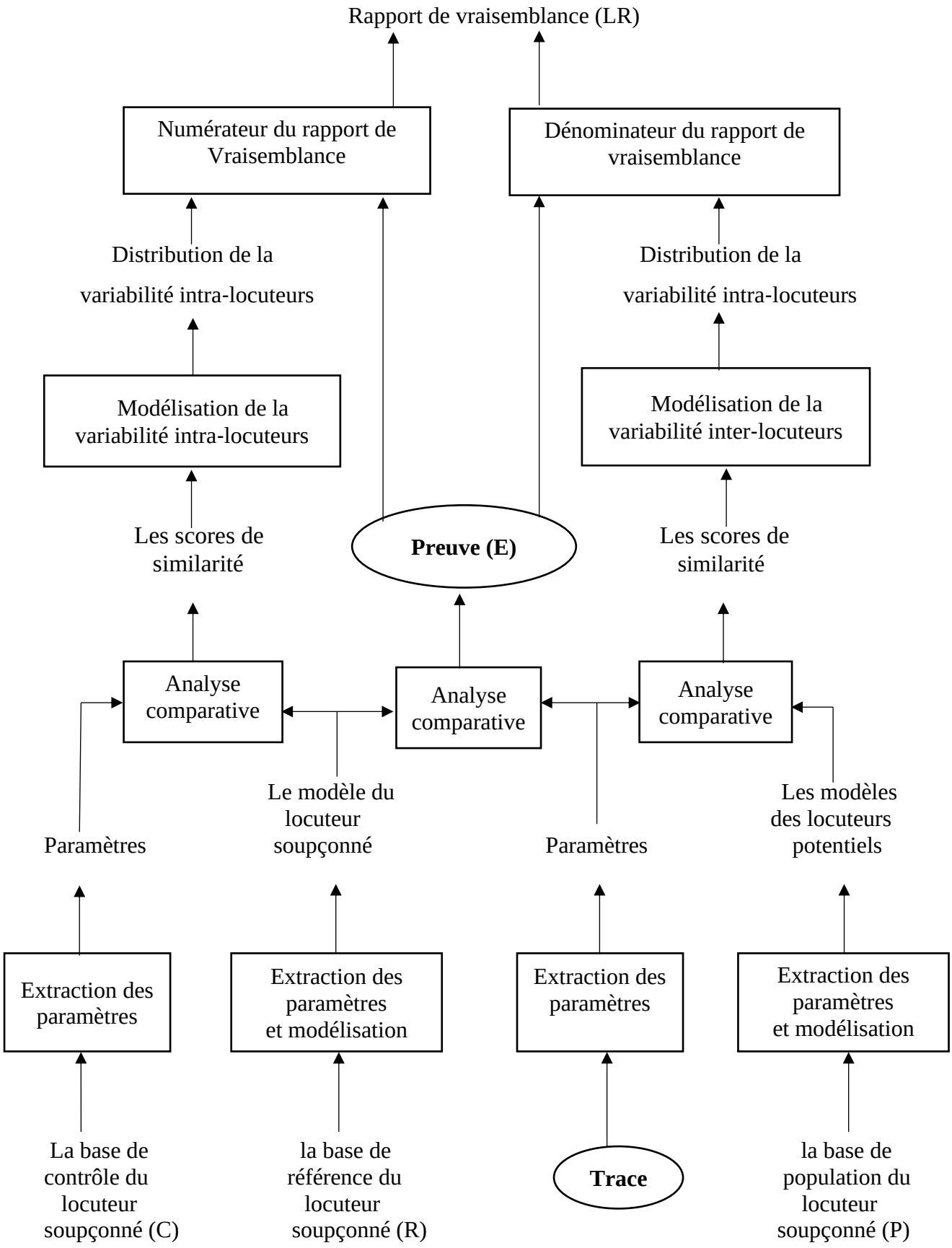


Fig.1.2. Le processus général de calcul et d'interprétation de la preuve E [14].

1.3.4. Les avantages de l'approche Bayésienne en criminalistique

- **Théoriquement Confirmée**

L'approche bayésienne présente de nombreux avantages, par exemple son applicabilité à de nombreux problèmes d'inférence, sa cohérence mathématique et, surtout, Cela fonctionne dans de nombreuses situations pratiques. Il a été démontré que le rapport de vraisemblance est dans la plupart des cas supérieur à 1 si les données proviennent de la même source et inférieur à 1 sinon. Cette hypothèse est valable pour de nombreux domaines d'identification forensique tels que l'ADN, le verre et même pour identifier les locuteurs [14].

- **La combinaison de preuves**

Un autre facteur important dans cette approche est qu'il est plus simple de combiner des preuves provenant de différentes sources. Ainsi, si deux échantillons vocaux ont été comparés avec le respect de deux paramètres vocaux différents et que nous avons constaté qu'ils différaient significativement d'un paramètre à l'autre, mais pas d'un autre. Ce résultat n'est pas facile à expliquer. En revanche, l'approche bayésienne permet de combiner l'annuaire de plusieurs sources en multipliant les rapports de probabilité de leurs rapports de résolution [14].

1.3.5. Les inconvénients de l'approche Bayésienne en criminalistique

- **L'indétermination des probabilités préalables**

Dans l'approche bayésienne, l'estimation des probabilités à priori est très nécessaire pour déterminer la probabilité d'une hypothèse. Dans la plupart des cas réels, les chances sont inconnues. Par conséquent, en effectuant de nombreuses enquêtes différentes, on peut avoir plusieurs estimations de ces probabilités, de sorte que les probabilités de fond peuvent également être différentes. Cependant, les experts légistes ne se soucient pas d'estimer les probabilités, ils sont intéressés par le calcul du rapport de probabilité. Il convient également de noter que la plupart des publications, utilisant l'approche bayésienne pour comparer des échantillons de données, utilisent simplement le rapport de probabilité RL sans aucune possibilité préalable [14].

❖ **la complexité**

Une autre critique est souvent liée à la complexité mathématique et logique du raisonnement bayésien, ce qui rend très difficile l'interprétation de la cour. Cependant, dans le cas réel, le tribunal n'a pas vraiment besoin de comprendre la théorie mathématique de cette approche, c'est pourquoi il est nécessaire de faire appel à des experts légistes. De plus, il n'est pas toujours difficile de comprendre la logique qui sous-tend cette approche, dans tous les cas, si les résultats sont formulés, par exemple, sous la forme suivante :
"Cette preuve est susceptible d'être 10 fois plus élevée sous l'hypothèse H0 que l'hypothèse H1"[14].

1.4. Les bases de données

Les informations fournies par l'analyse d'une trace mènent à spécifier une population de référence initiale. Cette population contient les locuteurs les plus similaires à celui qui a produit l'enregistrement en question. En intégrant les investigations de la police, on peut se baser sur un locuteur de cette population, qui est le locuteur suspect.

La méthode présentée précédemment nécessite trois bases de données pour le calcul et l'interprétation de la preuve : la base de données de la population potentielle (P), la base de données de référence du locuteur suspect (R) et la base de données de contrôle du locuteur suspect (C).

La base de données de la population potentielle (P) permet d'évaluer la variabilité interlocuteurs en utilisant la trace vocale. Cela veut dire, le calcul de la distribution des scores de similarité par la comparaison de la trace avec les modèles des locuteurs (GMMs) de la base de données de la population potentielle.

La base de données de référence du locuteur suspect (R) est enregistrée avec le locuteur suspect pour modéliser ces paramètres acoustiques par un modèle de mélange de gaussiennes GMM [14]. Dans ce cas, les signaux de parole sont produits de la même façon que celles de la base de données (P). Ensuite, le modèle GMM obtenu est utilisé pour calculer la valeur de la preuve (E) en comparant la trace par rapport à ce modèle.

La base de données de contrôle du locuteur suspect (C) est enregistrée avec le locuteur suspect pour évaluer la variabilité intra-locuteur. Le contenu de la base (R) doit être équivalent à la trace en termes de quantité et de type de parole [14].

1.5. Evaluation des systèmes RALF

Pour évaluer les performances des systèmes RALF qui donnent leurs résultats sous formes de valeurs LR, on doit faire plusieurs expériences. Les résultats de ses expériences seront représentés sous forme d'un graphe, dans lequel la distribution de la variabilité interlocuteurs et la distribution de la variabilité intra-locuteurs sont présentées. Ce genre de représentation est utilisé dans toutes les disciplines forensiques [14].

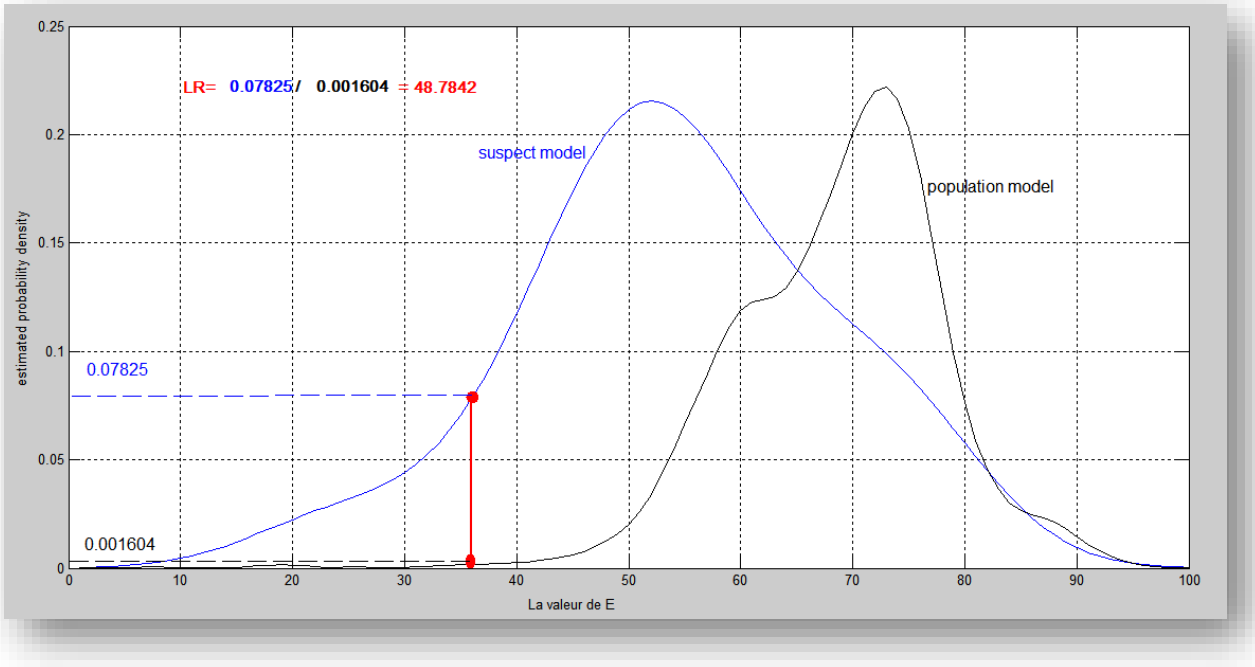


Fig.1.3. Illustration de la méthode des scores.

Dans cet exemple, le score de preuve est **E = 36**. Diviser la valeur de vraisemblance de la distribution H0 (**modèle du suspect**) par la valeur de vraisemblance de la distribution H1 (**modèle de la population**) pour le score de preuve de parole observé E donne **LR = 48.7842.**

Tab 1.1 : quelques descriptions verbales du rapport de vraisemblance [14].

Rapport de vraisemblance	Description verbale équivalente
>10 000	preuve très puissante pour l'hypothèse…
1000 à 10 000	preuve puissante pour l'hypothèse…
100 à 1000	preuve moyennement puissante pour l'hypothèse…
10 à 100	preuve modérée pour l'hypothèse…
1 à 10	preuve limitée pour l'hypothèse…
1 à 0.1	preuve limitée contre l'hypothèse…
0.1 à 0.01	preuve modérée contre l'hypothèse…
0.01 à 0.001	preuve moyennement puissante contre l'hypothèse…
0.001 à 0.0001	preuve puissante contre l'hypothèse…
<0.0001	preuve très puissante contre l'hypothèse…

1.6. Conclusion

Dans ce chapitre, nous avons présenté tous les principes de la reconnaissance automatique du locuteur en sciences forensiques. Par la suite, nous avons détaillé les fondements mathématiques de l'Interprétation Bayésienne pour RALF, qui constitue la base théorique de ce type de système, est la clé du succès des systèmes forensiques utilisant d'autres modalités (visage, iris, empreinte digitale, ADN,… etc). Finalement, nous avons montré dans ce chapitre, l'importance que jouent les probabilités dans la conception des systèmes de reconnaissances d'une manière générale et forensiques d'une manière particulière.

Chapitre 2

Mise en œuvre d'un système forensique du locuteur

Chapitre 2

Mise en œuvre d'un système forensique du locuteur

2.1. Introduction

L'analyse acoustique de la parole est aujourd'hui une composante fondamentale des systèmes de reconnaissance vocale. Cette analyse acoustique du signal (considérée comme étant un signal aléatoire d'une grande variabilité et redondance, continu, d'énergie finie, non stationnaire) a pour but de donner une représentation moins redondante de la parole, tout en permettant une extraction assez précise des paramètres acoustiques qui caractérisent ce signal.

Les principaux paramètres de l'analyse spectrale utilisés en RALF sont les coefficients issus de l'analyse en banc de filtres et leurs différentes transformations (coefficients banc de filtres, MFCC...). Dans les applications de reconnaissance automatique du locuteur en forensique, la modélisation du locuteur (suspect) tient compte de la distribution des paramètres acoustiques. La majorité des systèmes actuels de reconnaissance du locuteur sont basés sur l'utilisation de modèles de mélange de Gaussiennes (GMM) qui constituent l'état de l'art pour la vérification du locuteur. Ces modèles de nature générative sont généralement appris en utilisant les techniques de Maximum de Vraisemblance.

2.2. Extraction des vecteurs acoustiques

Presque toutes informations qui peuvent être extraites d'un signal de paroles se trouvent dans la bande fréquentielle 200Hz-8KHz. Les étapes principales pour extraire les vecteurs acoustiques sont : le prétraitement, l'extraction de paramètres, La figure **(Fig 2.1)** regroupe ces étapes [14].

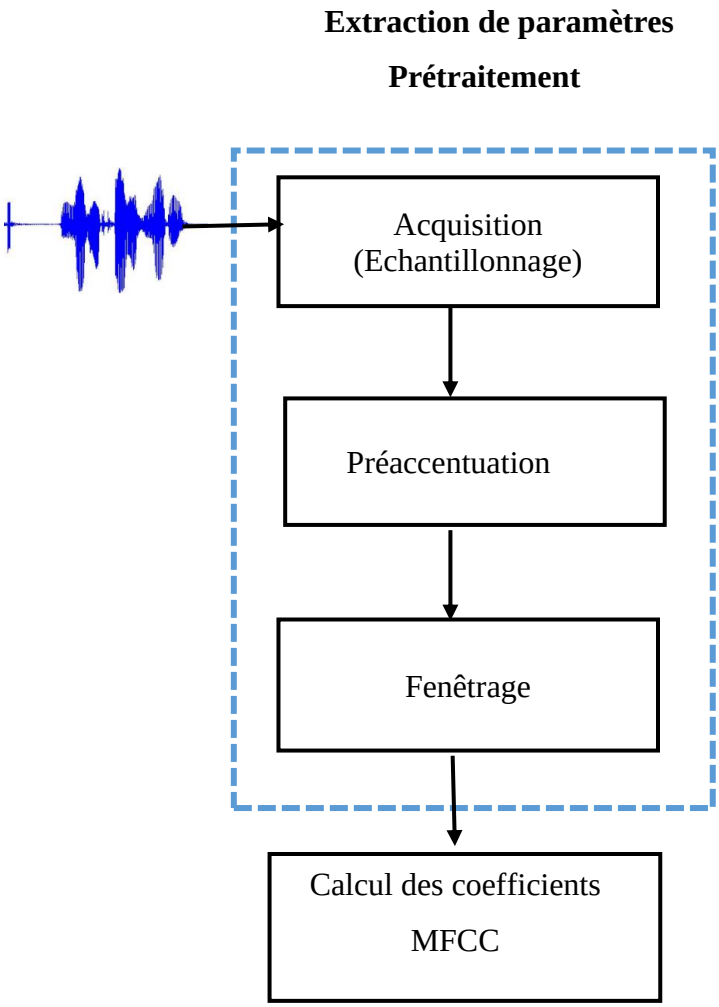


Fig.2.1. les étapes principales pour l'extraction des paramètres [20].

A partir d'un signal vocal échantillonné **x(n),** on peut trouver les vecteurs de paramètre **fx (n; m),** dont *m=0,1,…, M-1* et *n=0,1,…, N-1*, i.e. M vecteurs de taille N. par la suite, les étapes précédentes seront décrites en détail [14].

2.2.1. Prétraitements

2.2.1.1. Acquisition

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

Il s'agit tout simplement de relever à chaque instant « T » le niveau énergétique du signal acoustique tout en respectant le théorème de Shannon [21].

- **Théorème de Shannon**

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

$$f_e \geq 2 f_{max} \qquad \text{avec} \qquad f_e = \frac{1}{T_e} \qquad \textbf{(2.1)}$$

2.2.1.2. Préaccentuation

On remarque qu'au niveau du spectre de la parole, les basses fréquences sont favorisées par rapport aux hautes fréquences, car ce signal se caractérise par une pente globale négative de 6 dB/octave due aux influences de la source d'excitation et du rayonnement des lèvres [22].

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

$$H(z) = 1 - a.z^{-1}, 0.95 \leq n \leq 1 \qquad \textbf{(2.2)}$$

2.2.1.3. Fenêtrage

Il est difficile voire impossible de traiter un signal non stationnaire tel celui de la parole sans le fragmenter en trames. Une analyse à court terme montre que le signal vocal est quasi stationnaire sur des tranches temporelles de durées de 10 à 30 ms [23]. Cette analyse est effectuée à l'aide de fenêtres [24] telles que :

Fenêtre Hamming $\qquad w_n = 0{,}54 - 0{,}46.\cos(2\pi \frac{n}{N}), \quad 0 \leq n \leq N \qquad \textbf{(2. 3)}$

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

N **:** la taille de la fenêtre.

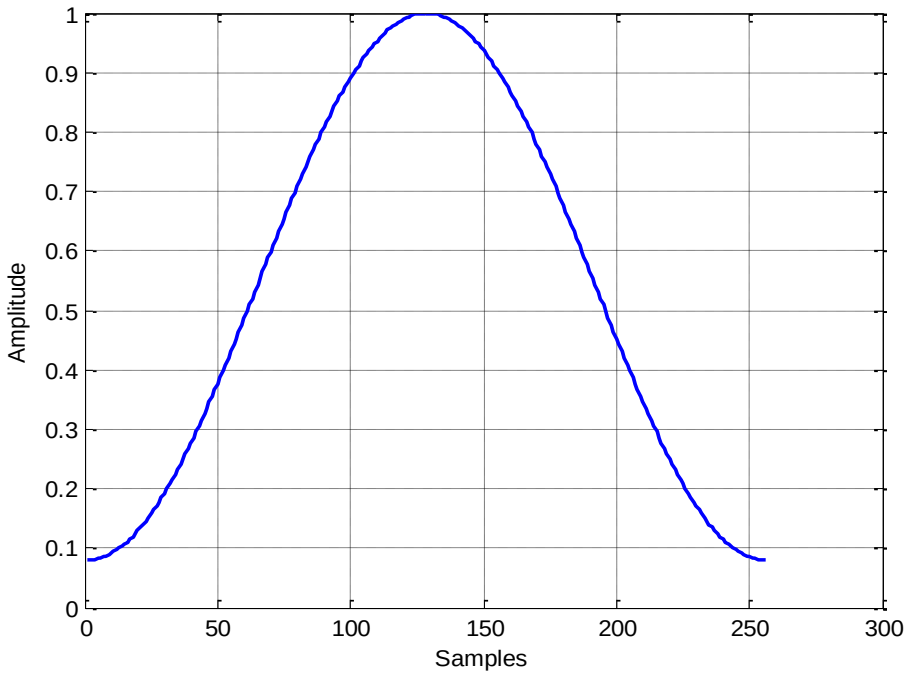


Fig 2.2. Allure temporelle de la fenêtre de Hamming.

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

Fenêtre Hanning $w_n = 0{,}5(1 - \cos(2\pi \frac{n}{N})), \quad 0 \le n \le N$ **(2.4)**

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

N : la taille de la fenêtre.

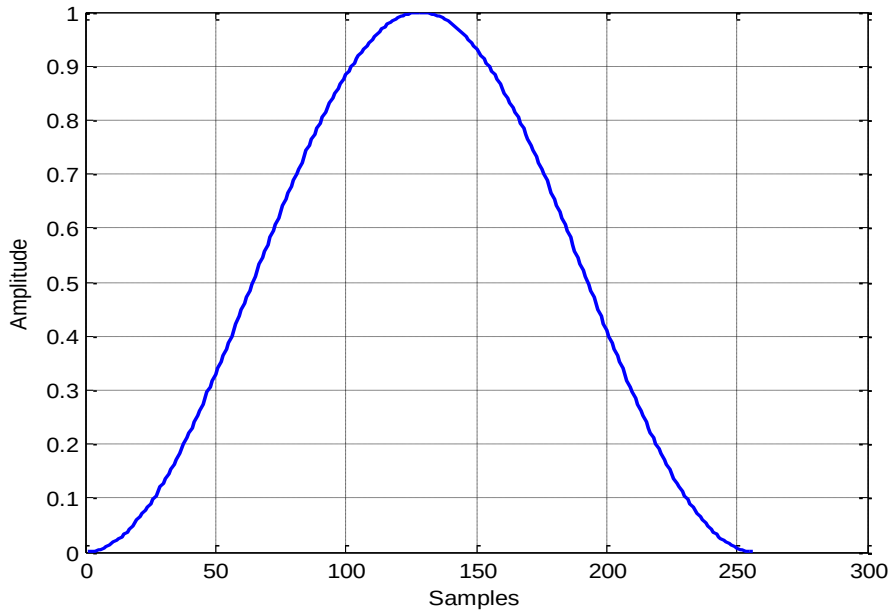


Fig 2.3. Allure temporelle de la fenêtre de Hanning.

Fenêtre Blackman $w_n = 0{,}42 - 0.5\cos(2\pi\frac{n}{N}) + 0.08\cos(4\pi\frac{n}{N}), \quad 0 \le n \le N$ **(2.5)**

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

N : la taille de la fenêtre.

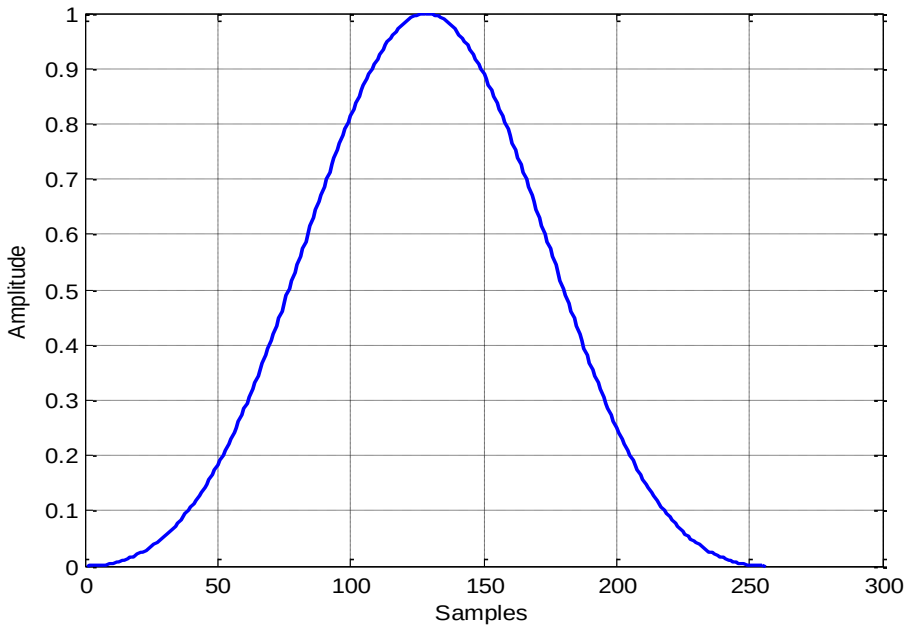


Fig 2.4. Allure temporelle de la fenêtre de Blackman.

2.2.2. Analyse spectrale

L'analyse spectrale présente des avantages au niveau de la perception car l'oreille humaine effectue une discrimination fréquentielle des sons [22]. De plus, cette analyse fait apparaître des propriétés et des paramètres pertinents pour la suite du traitement. Les principaux outils utilisés sont les suivants.

2.2.2.1. La transformée de Fourier discrète

Pour effectuer cette analyse on utilise :

$$X(n)=\sum_{k=0}^{N-1} x(k)\times e^{-j\pi\frac{nk}{N}} \qquad \textbf{(2.6)}$$

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

N : Le nombre d'échantillons de la trame.

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

Ce qui nous donne le spectre fréquentiel du signal analysé.

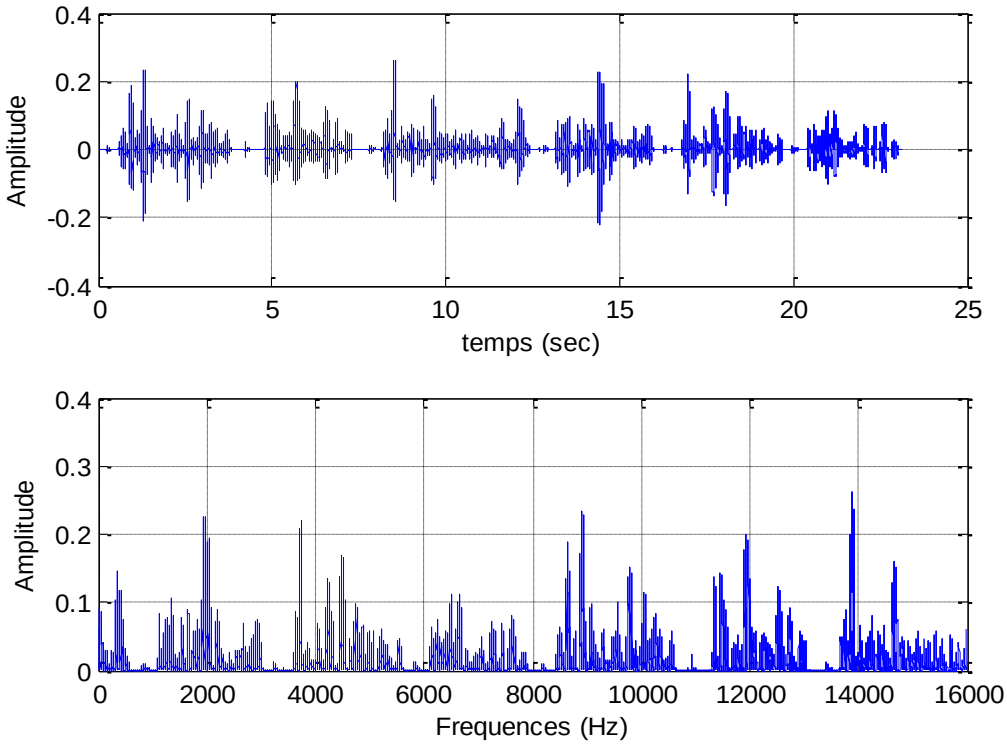


Fig 2.5. un exemple qui montre le signal temporel (en haut) et sa transformée de Fourier (en bas).

2.2.3. Analyse temporelle

2.2.3.1. Energie totale :

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

Elle est calculée à partir de la relation suivante :

$$E=\frac{1}{N}\sum_{k=0}^{N-1}x^2(k)\,. \qquad \textbf{(2.7)}$$

Avec *E :* la valeur à évaluer.

N : la largeur de la fenêtre d'analyse.

x(k) : le signal numérique.

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

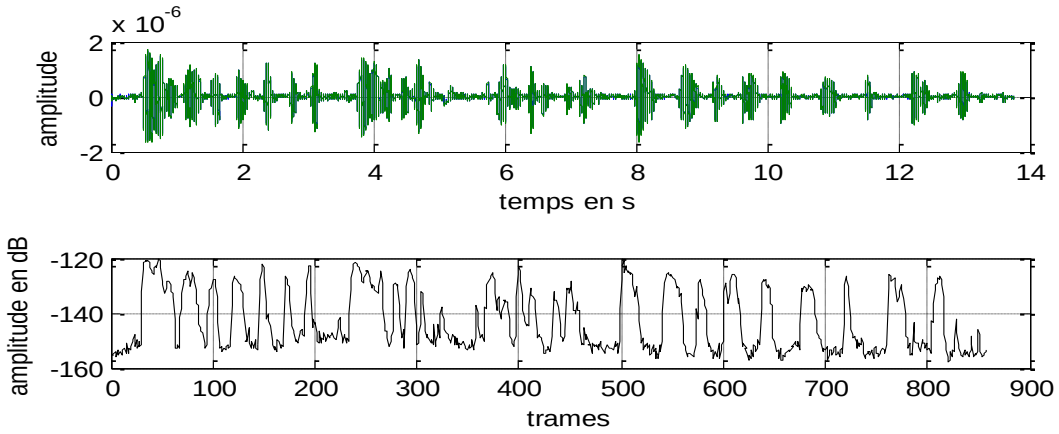


Fig 2.6. Allure temporelle de la phrase ' *L'enseignant explique à ses étudiants ce qu'ils doivent faire'* en haut, son énergie au bas

2.2.3.2. Taux de passage par zéro

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

$$TPZ = \frac{1}{2}\sum_{k=0}^{k-1}\left|sign(x(k+1)) - sign(x(k))\right| \quad \textbf{(2.8)}$$

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

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

2.2.3.3. Détection de l'activité vocale (VAD)

Une façon de sélectionner des trames de parole dans un signal de parole consiste à utiliser l'énergie, en faisant l'hypothèse que les trames les plus énergétiques, correspondant principalement aux zones stables des voyelles et aux zones pour lesquelles le rapport signal à bruit est élevé, sont les plus intéressantes.

Une façon d'obtenir la classification des trames parole non-parole, consiste à utiliser un modèle d'énergie. La distinction énergétique des trames est réalisée par le calcul d'énergie de chaque trame. Les trames de plus faible énergie représentent les trames de non-parole et les trames de plus haute énergie représentent les trames de parole [25]. Une fois ces énergies sont calculées, un seuil est calculé pour attribuer les trames à l'une ou l'autre des classes (c.-à-d. ; parole ou non-parole). Cette méthode est simple à mettre en œuvre et obtient de bons résultats sur des séquences courtes (quelques secondes) en milieux calme.

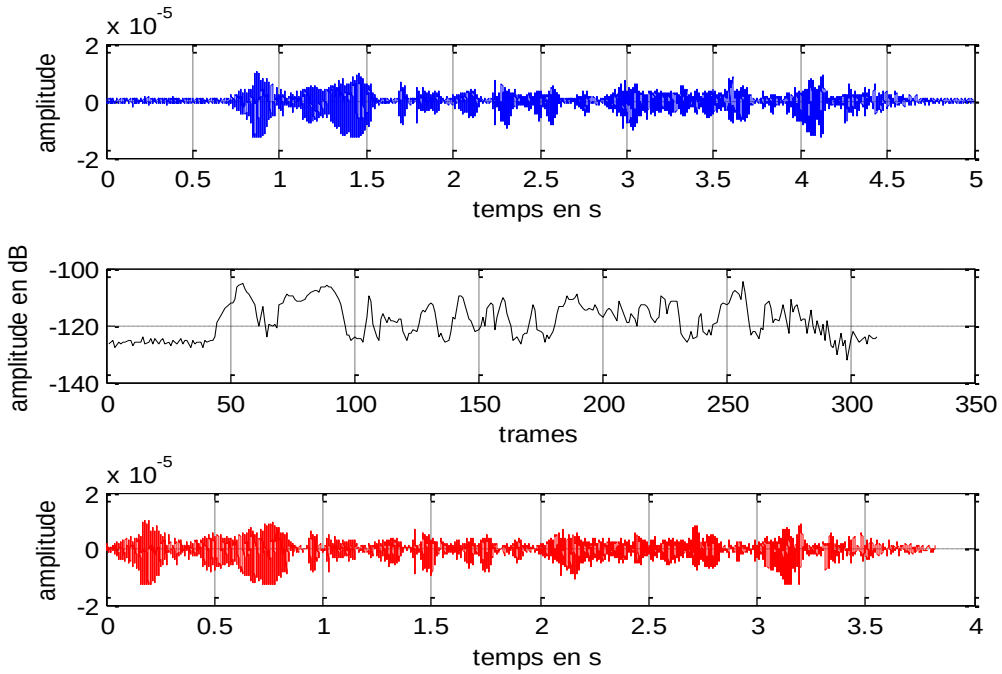


Fig 2.7. Allure temporelle de la phrase ' ***Détection de l'activité vocale*** ' en haut, son énergie au milieu et l'allure temporelle de même phrase après suppression du silence en bas.

2.2.4. Extraction des paramètres

L'extraction de paramètres est une étape très importante. Les performances d'un système de reconnaissance forensique du locuteur dépendent essentiellement de la qualité des paramètres choisis. La plus part des systèmes **RAFL** utilisent les **Mel Frequency Cepstral Coefficients (MFCC)** [14] et ceci pour les raisons suivantes :

- Ces mesures fournissent un bon modèle de signal de paroles, cela est particulièrement vrai dans des régions quasi stationnaires du signal de paroles.
- Ces mesures ont un modèle analytique soluble.
- Des expériences ont montré que ces mesures donnent de bons résultats dans les applications de reconnaissance automatique du locuteur. [14]

2.2.4.1. Étapes de calcul du vecteur caractéristique de type MFCC

Dans nos expériences, une analyse est appliquée toutes les **10 ms** par glissement et recouvrement sur des fenêtres d'analyse de **20 ms**. A chaque trame, un vecteur de représentation acoustique les **MFCC** sont calculés à partir d'un banc de 24 filtres triangulaires répartis dans **l'échelle fréquentielle Mel** [26].

Les **MFCC** d'une trame de parole sont calculés de la façon suivante :

- après le filtrage de préaccentuation, le signal de parole est d'abord découpé en fenêtres de taille fixe réparties uniformément le long du signal.
- la **FFT** (**Fast Fourier Transform**) de la trame est calculée. Ensuite, l'énergie est calculée en élevant au carré la valeur de la **FFT**. L'énergie est passée ensuite à travers chaque filtre Mel. Soit **Sk** l'énergie du signal à la sortie du filtre **K**, nous avons maintenant m_p (**le nombre de filtres**) paramètres **Sk**. (Des études ont montré que les 20 premiers paramètres de chaque trame extraits du filtre Mel représentent très bien le locuteur).
- le logarithme de Sk est calculé.
- finalement les coefficients sont calculés en utilisant la **IDCT** (**inverse Discrete Cosinus Transform**). Avec la **FFT**, nous sommes passées à l'échelle fréquentielle et avec la IDCT nous retournons vers le temporel, nous avons utilisé **IDCT** au lieu de **IFFT** car **IDCT** a l'avantage de la décorrélation (c'est-à-dire. une matrice de covariance diagonale):

$$\boldsymbol{C_i = \sqrt{\frac{2}{m_p}} \sum_{k=1}^{k=m_p} \log(S_k) \cos\left(i(k-1/2)\frac{\pi}{m_p}\right)} \qquad \textbf{(2.9)}$$

i = 1,…., N, où **N** est le nombre des **MFCC** que nous souhaitons obtenir.

Avec : **Ci** = les MFCC ; **Sk** = l'énergie du signal à la sortie du filtre k ; **mp** = nombre de filtres.

Les étapes de cette opération sont illustrées par la figure **(Fig 2.8)**

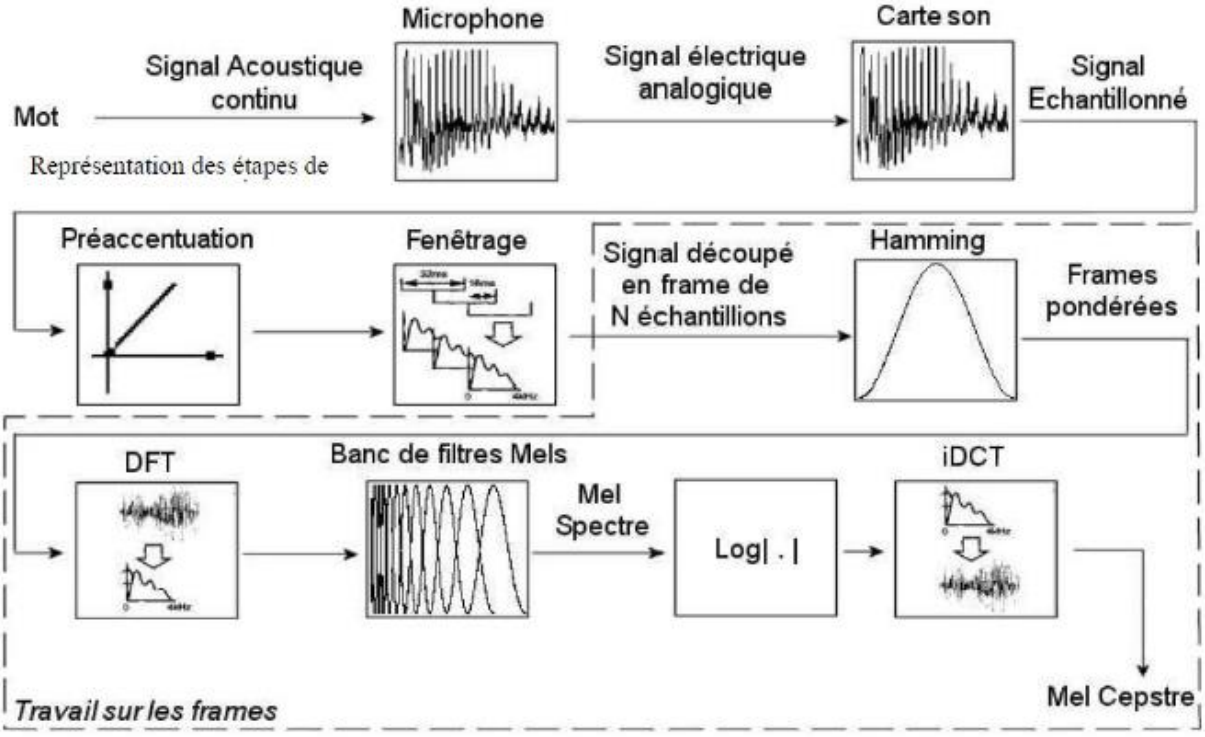


Fig 2.8. Étapes de calcul d'un vecteur caractéristique de type MFCC.

2.3. Filtrage sur l'échelle Mel

On considère que l'oreille humaine perçoit linéairement le son jusqu'à 1000 Hz, mais après, elle perçoit moins d'une octave par doublement de fréquence [27].

La réponse en fréquence magnitude est multipliée par un ensemble de 40 filtres passe-bande triangulaires pour obtenir l'énergie de log de chaque filtre passe-bande triangulaire. Les positions de ces filtres sont également espacées le long de la fréquence Mel. Parmi les fréquences centrales comprises entre 133,33 Hz et 1 kHz, il existe 13 filtres linéaires chevauchants (50%), tandis que pour les fréquences centrales allant de 1 kHz à 8 kHz, il existe 27 filtres chevauchants espacés de manière logarithmique [28].

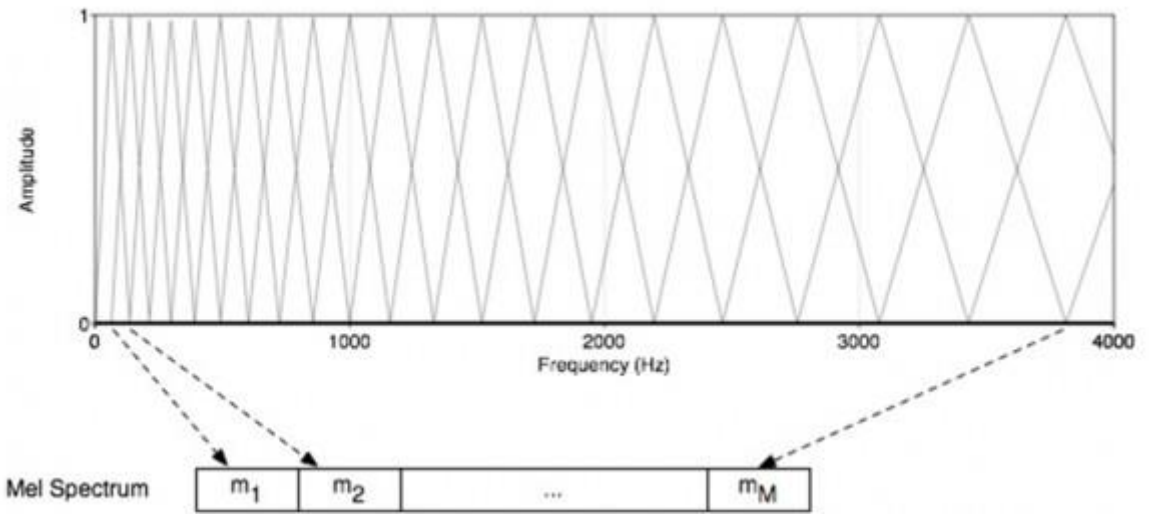


Fig 2.9. Implémentation de bancs de filtres selon l'échelle MEL avec 21 canaux répartis entre 0 et 4000Hz

2.4. Modélisation des MFCC par les GMM

Le modèle de mélange de Gaussiennes est un modèle statistique ou la distribution des données est un mélange de plusieurs lois Gaussiennes. Le GMM est le modèle de référence en reconnaissance du locuteur [29].

La m^{éme} loi gaussienne d'un mélange $\boldsymbol{\lambda}$ à **M** composantes est paramétrée par un vecteur de moyennes $\boldsymbol{\mu_m}$ de dimension **D** (D étant la dimension de l'espace des données), une matrice de covariance $\sum_{\boldsymbol{m}}$ de dimension D × D et un poids $\boldsymbol{\pi_m} \geq 0$.[29] La fonction de densité de probabilité s'écrit sous forme de :

$$P(\boldsymbol{x}|\boldsymbol{\lambda}) = \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \boldsymbol{\pi_m}\, \boldsymbol{b_m}(\boldsymbol{x}) \qquad \textbf{(2.10)}$$

Où :

$b_m(x)$: la densité gaussienne paramétrée par le vecteur moyen μ_m et la matrice de covariance $\sum_m$ cette densité est donnée par :

$$\boldsymbol{b_m}(\boldsymbol{x}) = \frac{\boldsymbol{1}}{(\boldsymbol{2\pi})^{\boldsymbol{D/2}}|\sum_{\boldsymbol{m}}|^{\boldsymbol{1/2}}} \textbf{exp}\,[-\frac{\boldsymbol{1}}{\boldsymbol{2}}(\boldsymbol{x}-\boldsymbol{\mu_m})^{\boldsymbol{T}}(\sum_{\boldsymbol{m}})^{\boldsymbol{-1}}(\boldsymbol{x}-\boldsymbol{\mu_m})] \qquad \textbf{(2.11)}$$

et $$\sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \boldsymbol{\pi_m} = \boldsymbol{1} \qquad \textbf{(2.12)}$$

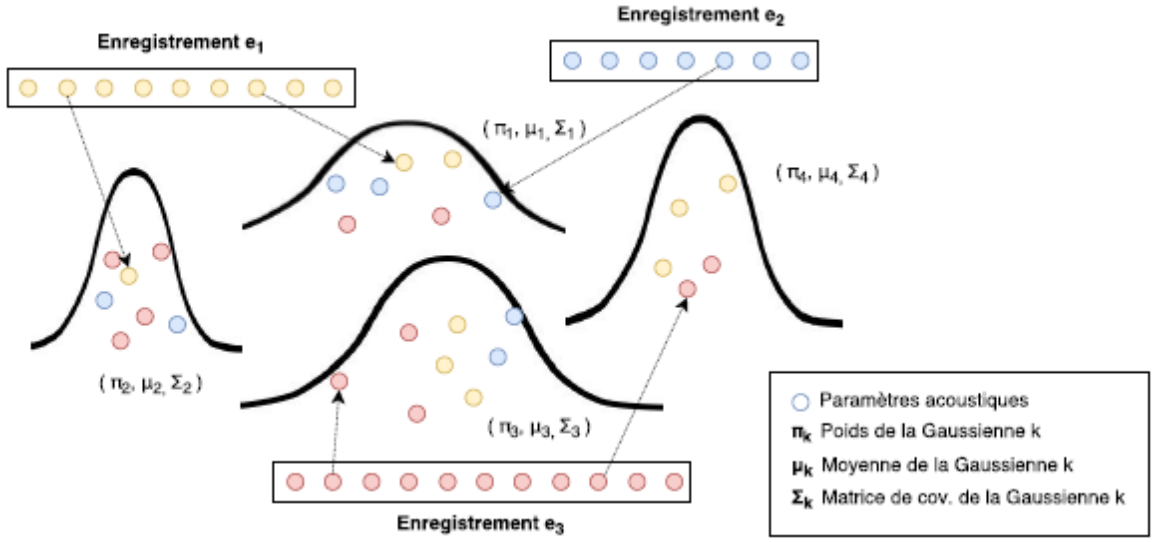


Fig 2.10. Un mélange de Gaussiennes (GMM) construit en utilisant des paramètres acoustiques issus de plusieurs enregistrements [30].

2.4.1. Estimation par maximum de vraisemblance

L'apprentissage d'un modèle GMM consiste en l'estimation de l'ensemble des paramètres $\lambda = \{\pi_m, \mu_m, \sum_m\}$ avec m=1,…, **M**. Cet apprentissage fait souvent appel à la technique d'estimation par **maximum de vraisemblance** (Maximum Likelihood Estimation) **MLE**; on utilise souvent l'algorithme **Espérance-Maximisation** (**Expectation-maximisation**) **EM** [29] pour déterminer les paramètres du modèle qui maximisent la vraisemblance des données d'apprentissage. En utilisant un ensemble de données d'apprentissage $X = \{x_1, x_2, ..., x_N\}$ ($x_N \in R^D$) de N vecteurs d'apprentissage, le maximum de vraisemblance du GMM est donné par :

$$P(X|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda) = \prod_{n=1}^{N} \sum_{m=1}^{M} p(x_n/\pi_m, \mu_m, \sum_m) \qquad (2.13)$$

L'algorithme EM vise ainsi à maximiser la loi de vraisemblance en présence de données incomplètes en maximisant itérativement l'espérance de la log-vraisemblance complète donnée par :

$$V(X, \lambda) = \frac{1}{N} log \prod_{n=1}^{N} p(x_n|\lambda) = \frac{1}{N} \sum_{n=1}^{N} log p(x_n|\lambda) \qquad (2.14)$$

Chacune des itérations de l'algorithme comporte deux étapes [29]:

- L''étape d'espérance (E), où on calcule les probabilités a posteriori que les gaussiennes aient génère les données d'apprentissage :

$$\boldsymbol{P(m|x_n) = \frac{\pi_m\, p(x_n|\mu_m,\sum_m)}{\sum_{k=1}^{M} \pi_m p\left(x_k\middle|\mu_{k,\sum_k}\right)}} \quad \textbf{(2.15)}$$

- L''étape de maximisation (M), où on ré-estime les paramètres du modèle afin de maximiser la vraisemblance :

  - $\boldsymbol{\pi_m = \frac{1}{N}\sum_{n=1}^{N} p(m|x_n)}$ **(2.16)**

  - $\mu_m = \frac{\sum_{n=1}^{N}(p(m|x_n)x_n)}{\sum_{n=1}^{N} p(m|x_n)}$ **(2.17)**

  - $\sum_m = \frac{\sum_{n=1}^{N}(p(m|x_n)x_n}{\sum_{n=1}^{N} p(m|x_n)}$ **(2.18)**

2.4.2. Approche GMM-UBM

L'estimation par maximum de vraisemblance nécessite une grande quantité de données, pour estimer robustement les paramètres d'un modèle GMM. Dans le cas où la quantité de données n'est pas suffisante pour un apprentissage "direct" du GMM, on utilise des méthodes d'adaptation de modèles. En reconnaissance automatique du locuteur, peu de données sont généralement disponibles pour apprendre directement ces modèles [29].

Un modèle GMM du monde ou UBM (Universal Background Model) à matrices diagonales est ainsi appris par l'algorithme EM sur des centaines voire des milliers d'heures d'enregistrements appartenant à plusieurs locuteurs et dans différentes conditions d'enregistrement. Ensuite, le modèle d'un locuteur est appris par adaptation (généralement par la méthode de Maximum A Posteriori de l'UBM aux données de ce locuteur).

2.5. Conclusion

Nous avons présenté dans ce chapitre l'extraction des paramètres acoustiques qui sont utilisés pour estimer le modèle statistique GMM. C'est une étape très importante dans le système forensique du locuteur. Ainsi, les techniques de prétraitement acoustique qui sont insérées dans les chaines d'extraction de ces vecteurs caractéristiques, à savoir la détection de l'activité vocale (**VAD**). Aussi nous avons détaillé l'étude théorique de modèle GMM, qui consiste en l'estimation des probabilités a postériori, afin de modéliser d'une manière assez fidèle toutes les classes phonétiques (voyelles, consonnes, fricatives,…etc) qui existent dans le signale de parole.

Chapitre 3

Evaluation Expérimentale du Système RALF

Chapitre 3

Evaluation Expérimentale du Système RALF

3.1. Introduction

L'évaluation du système RALF débute par la sélection et la constitution de bases de données d'enregistrements de parole. Il est important que la qualité de ces enregistrements soit comparable à celle qui peut être atteinte lors de l'enregistrement vocal ou d'une écoute téléphonique. Les différentes évaluations définies par la comparaison des enregistrements de test par rapport à des enregistrements de locuteurs suspects pour le calcul des rapports de vraisemblance, sont effectuées dans ce chapitre.

3.2. Principe

Les expériences réalisées au chapitre 3 servent à tester le système de reconnaissance automatique de locuteurs dans différentes conditions rencontrées en criminalistique. Le principe d'évaluation de la méthode consiste à estimer les rapports de vraisemblance qui peuvent être obtenus à partir de l'élément de preuve **E**, d'une part lorsque l'hypothèse $\mathbf{H_0}$ est vérifiée, c'est-à-dire lorsque la source du modèle et celle de l'enregistrement de test sont uniques, et d'autre part lorsque l'hypothèse $\mathbf{H_1}$ est vérifiée, c'est-à-dire lorsque la source du modèle et celle de l'enregistrement de test sont différentes.

3.3. Protocole expérimental

Dans notre évaluation, nous avons utilisé le logiciel MATLAB comme outil de simulation, nous avons présenté les différents résultats de l'ensemble de tests d'évaluation effectués sur le système de reconnaissance du locuteur en forensique (criminalistique) RALF [14], sur trois bases de données (NIST, TIMIT, Algerian Speech), Nous avons effectué les expériences sur le corpus NIST (15 locuteurs choisis aléatoirement parmi 199 locuteurs), constitué de la parole téléphonique spontanée échantillonnée à 8 kHz, le corpus TIMIT (15 locuteurs choisis aléatoirement parmi 345 locuteurs), et le corpus Algerian Speech (aussi 15 locuteurs choisis aléatoirement parmi 60 locuteurs). TIMIT et Algerian Speech sont deux bases constituées de la parole, la première spontanée et la deuxième lue, échantillonnées toutes les deux à 16 kHz. Pour l'extraction de paramètres, on trouve un vecteur de 23 coefficients MFCCs extraits à

partir des trames de la parole chevauchée toutes les 10 ms, en utilisant une fenêtre de Hamming de 20 ms [31]. Le modèle d'apprentissage utilisé pour apprendre les données caractéristiques du locuteur (suspect) est le GMM normalisé par UBM à 256 composantes (gaussiennes) [12].

3.3.1. Enregistrement et sélection de bases de données

Cette étape consiste à enregistrer ou à sélectionner les deux bases de données, la première servant à estimer la variabilité interlocuteur à l'intérieur de la population des locuteurs qui sont potentiellement à l'origine de l'enregistrement considéré comme indice, la seconde servant à l'estimation de la variabilité intralocuteur de la ou des personne(s) suspectée(s) d'être la source de l'indice. Elle sert aussi à constituer un ensemble d'enregistrements de test, de manière à simuler des indices matériels qui peuvent être rencontrés en cas d'abus de téléphone ou de mesure de surveillance [32].

3.3.2. Les bases de données

A partir des enregistrements (format Wave) de chaque base de données, nous avons créé deux sous-ensembles de fichiers vocaux [14]. Pour chaque base de données (NIST, TIMIT, et Algerian Speech), nous avons construit un ensemble d'apprentissage et un ensemble de test. La figure **(Fig 3.1)** illustre l'organisation des bases de données.

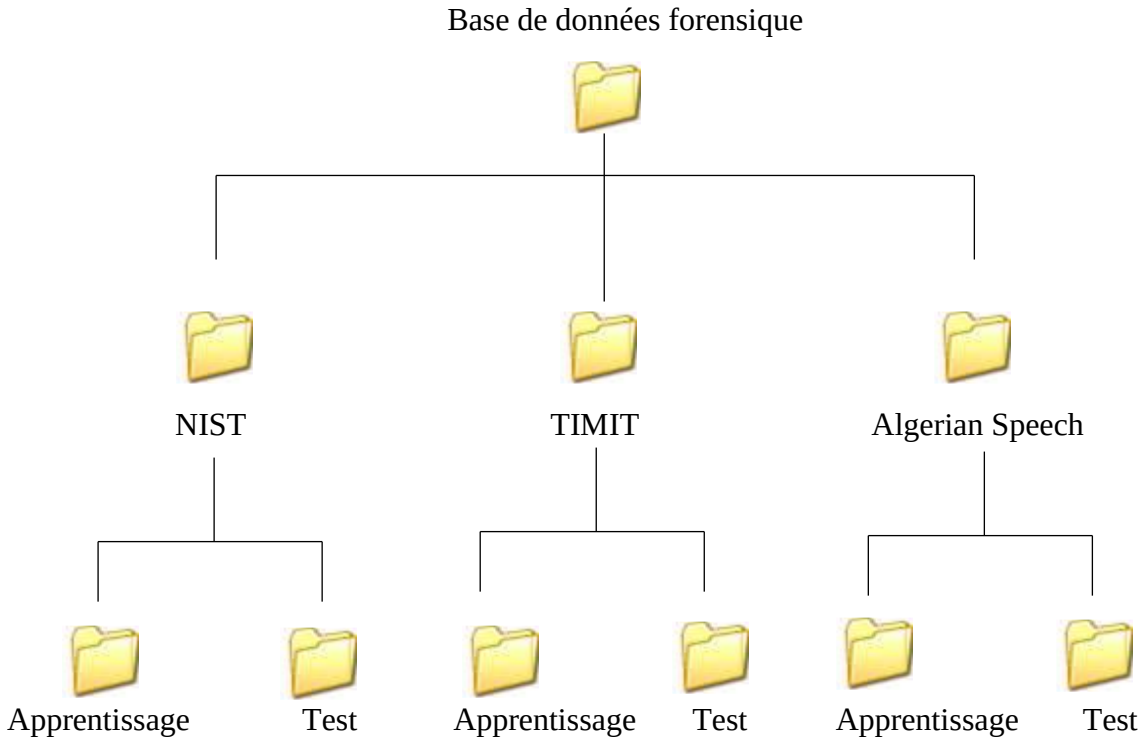


Fig 3.1. L'organisation des bases des données.

Dans la phase d'apprentissage, il y a deux bases de données, la première est la base de données de la population potentielle (P), et la deuxième est la base de données de référence du locuteur suspect (R). Par contre la phase de test contient la base de données de control du locuteur suspect (C) et celle de la trace (T).

3.4. Évaluation de l'influence de l'existence de la trace (T) dans la base (R)

3.4.1. Évaluation sur la base de données « TIMIT »

3.4.1.1. Procédure

Cette évaluation est réalisée sur une sélection de 345 locuteurs (n° 0001 à 0345). L'enregistrement de test est comparé au modèle de la voix des 330 personnes de la population potentielle (n° 0001 à 0330) avec 15 personnes de la référence et de control de locuteurs suspects.

3.4.1.2. Résultats et Discussion

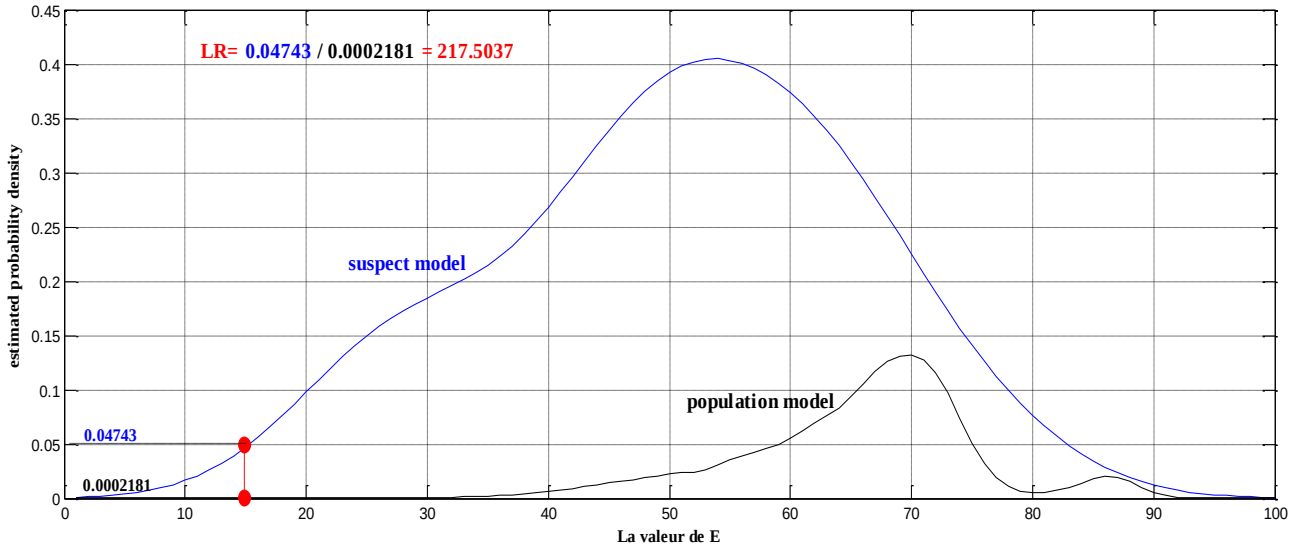
Tab 3.1. Calcul de la vraisemblance de E valant 15, dans les cas :a) la trace existe dans la base de référence. b) la trace n'existe pas dans la base de référence.

	E	**LR**
A	**15**	**217.5037**
B	**36**	**2.3597**

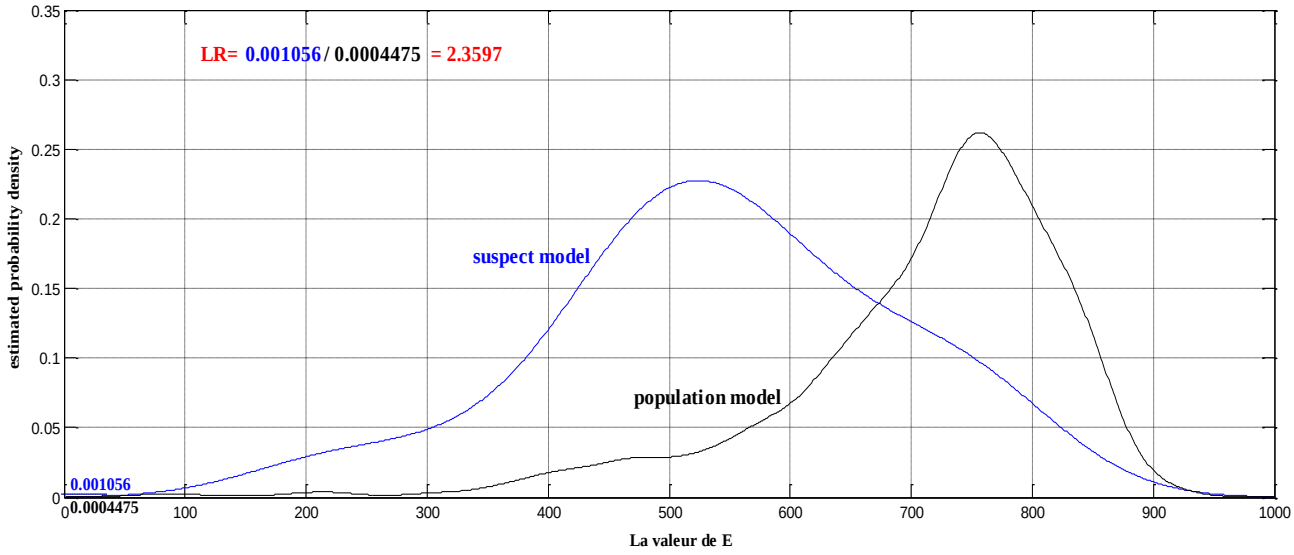
Les rapports de vraisemblance de ces éléments de preuve sont calculés de la manière suivante : le numérateur équivaut à la densité de probabilité de l'élément de preuve E dans la distribution de la variabilité interlocuteur du locuteur dont provient l'enregistrement de test. Le dénominateur du rapport de vraisemblance équivaut à la densité de probabilité de l'élément de preuve E dans la distribution intra-locuteurs de l'enregistrement de test. Voir la figure **(Fig 3.2)**.

Dans la situation ou la valeur de LR est entre 100 à 1000, la preuve moyennement puissante pour l'hypothèse **H0** et quand la valeur de la vraisemblance LR est entre 1et 10, la preuve limitée pour l'hypothèse **H0**. La preuve dans les deux cas (a) et (b) valant 15 et 36, la valeur de Rapport de vraisemblance LR dans le premier cas est supérieure à celle de deuxième cas car l'hypothèse $\mathbf{H_0}$ est vraie, c'est à dire la trace est générée par le suspect. Aussi d'après les résultats de tableau **Tab 3.1**, nous remarquons que dans le cas (b) **LR=2.3597>1 (voir Tab.1.1 de chapitre 1),** et ça malgré que la trace n'existe pas dans la

base de référence. Ceci peut être expliqué par le fait que, la forte corrélation (similitude) existante entre les locuteurs en termes de l'accent, le type de la langue parlée (anglais américain), la même région cohabitée par les suspects, rend la discrimination entre eux (suspects) difficile, d'où le **LR** supérieur légèrement à **1**.



(a)



(b)

Fig 3.2. Allures de l'évaluation de l'influence de l'existence de la trace :a)la trace existe dans la base de référence. b) la trace n'existe pas dans la base de référence

3.5. Évaluation de l'effet du nombre de suspects en fonction du nombre de traces

3.5.1. Évaluation sur les bases de données « TIMIT » et « NIST »

3.5.1.1. Procédure

Cette évaluation est réalisée sur deux bases TIMIT et NIST. Dans la première expérience, nous avons utilisé la base TIMIT sur une sélection de 345 locuteurs (n° 0001 à 0345). L'enregistrement de test (une trace à cinq traces) est comparé au modèle de la voix des 330 jusqu'à 342 personnes de la population potentielle avec 3 à 15 personnes de la référence et de control de locuteurs suspects, à chaque fois on ajoute trois (3) personnes. Il n'existe qu'un seul enregistrement pour chaque personne de la base de données « TIMIT ».

Dans la seconde expérience, l'évaluation est effectuée sur la base NIST pour une sélection de 199 locuteurs (n° 0001 à 0199). L'enregistrement de test (une trace à trois traces) est comparé au modèle de la voix des 193 jusqu'à 197 personnes de la population potentielle avec 2 à 6 personnes de la référence et de control de locuteurs suspects, à chaque fois on ajoute une personne. Il existe six enregistrements pour chaque personne de la base de données « NIST ».

3.5.1.2. Résultats et Discussion

Les rapports de vraisemblance de ces éléments de preuve sont calculés de la manière suivante : le numérateur équivaut à la densité de probabilité de l'élément de preuve E dans la distribution de la variabilité interlocuteurs dont provient l'enregistrement de test. Le dénominateur du rapport de vraisemblance équivaut à la densité de probabilité de l'élément de preuve E dans la distribution intra-locuteurs de l'enregistrement de test.

Dans la globalité des résultats montrés par le tableau (**Tab 3.2**), nous constatons qu'au fur et à mesure le nombre de locuteurs suspects augmente par rapport au nombre de traces, la valeur de LR diminue. Ceci est expliqué par le faite que la tâche d'indentification d'un criminel à travers les échantillons de sa voix, est un problème de classification entre plusieurs classes des échantillons de parole de plusieurs suspects. Et dans le domaine de reconnaissance (classification) des formes, nous savons pertinemment qu'à chaque fois le nombre de classes d'apprentissage est grand, implique automatiquement que le degré de complexité de la classification est grand. D'où les résultats sont moins bons. Nous trouvons dans le tableau (**Tab 3.3**), les mêmes constats que celui du tableau (**Tab 3.2**), à la différence que les résultats obtenus avec la base TIMIT sont plus performants et meilleurs que ceux obtenus en utilisant la base NIST, et ceci est du bien sûr à l'influence de la fréquence d'échantillonnage.

Tab 3.2. Calcul de rapport de vraisemblance pour nombre de suspects et traces différents dans la base de données « TIMIT ».

T_DB / C_DB	1	2	3	4	5
3	E=3 LR= 1.3508e+003	E=6 LR= 1.9866e+003	E=9 LR= 2.7663e+004	X	X
6	E=6 LR= 54.3090	E=12 LR= 4.4102e+003	E=18 LR= 1.1050e+003	E=24 LR= 1.1181e+003	E=30 LR= 164.9025
9	E=9 LR= 1.0682e+003	E=18 LR= 1.5015e+003	E=27 LR= 122.9340	E=36 LR= 75.0454	E=45 LR= 59.2233
12	E=12 LR= 774.0379	E=24 LR= 729.1572	E=36 LR= 103.2866	E=48 LR= 23.3628	E=60 LR= 4.1273
15	E=15 LR= 217.2158	E=30 LR= 185.0663	E=45 LR= 43.6352	E=60 LR= 5.6024	E=75 LR= 3.0278

Tab 3.3. Calcul de rapport de vraisemblance pour nombre de suspects et traces différents dans la base de données « NIST ».

T_db / C_db	1	2	3
2	E=12 LR= 274.3468	E=24 LR= 121.3322	X
3	E=18 LR= 27.0201	E=36 LR= 28.7478	E=54 LR= 1.6061
4	E=24 LR= 32.4630	E=48 LR= 10.2513	E=72 LR= 0.4671
5	E=30 LR= 172.5355	E=60 LR= 2.4669	E=90 LR= 1.2934
6	E=36 LR= 48.7783	E=72 LR= 0.5206	E=108 LR= 18.6355

3.6. Évaluation de l'effet du genre des suspects sur les performances du système RALF

3.6.1. Évaluation sur la base de données « TIMIT »

3.6.1.1. Procédure

Cette expérience est réalisée en utilisant la base TIMIT, nous avons dévisé la base en deux sous bases catégoriques ; féminin (TIMITF) et masculin (TIMITM). L'évaluation est faite avec la base TIMITF sur une sélection de 98 locuteurs (n° 0001 à 0113). Avec la base TIMITM, une sélection sur 232 locuteurs (n° 0001 à 0232) est aussi réalisée. L'enregistrement de test (deux traces) est comparé au modèle de la voix des 98 personnes de la population potentielle avec 15 personnes de la référence et de control de locuteurs suspects pour la base TIMITF et comparé au modèle de la voix des 217 personnes de la population potentielle avec 15 personnes de la référence et de control de locuteurs suspects pour la base TIMITM.

3.6.1.2. Résultats et Discussion

Tab 3.4. Calcul de la vraisemblance de E valant 30, dans chaque base de données TIMIT féminin et masculin.

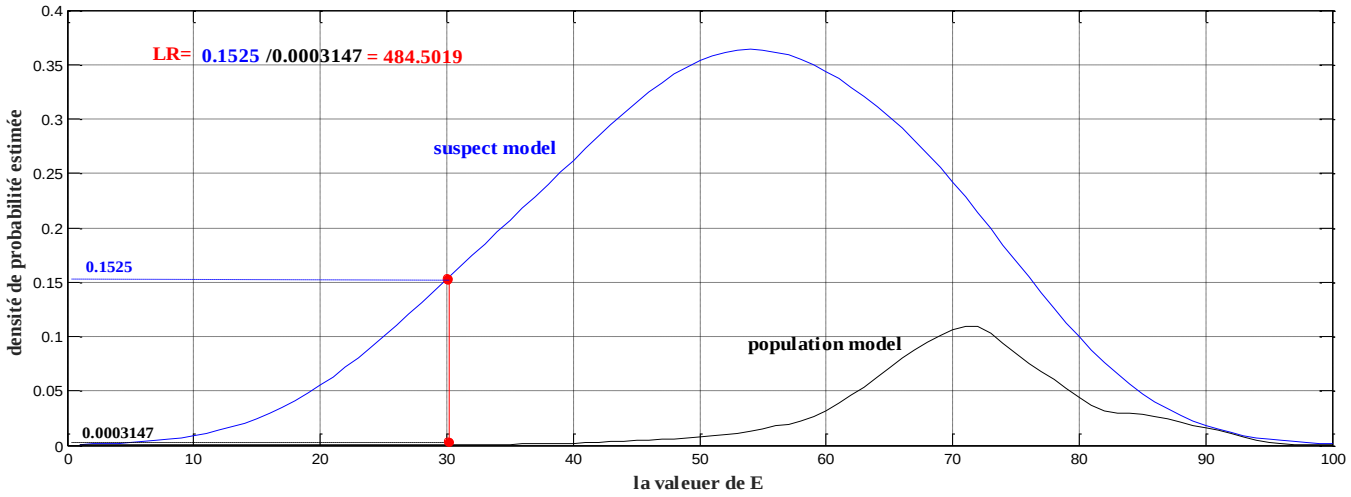
	E	**LR**
TIMITF (féminin)	**30**	**484.5019**
TIMITM (masculin)	**30**	**67.7454**

Les rapports de vraisemblance de ces éléments de preuve sont calculés de la manière suivante : le numérateur équivaut à la densité de probabilité de l'élément de preuve E dans la distribution de la variabilité interlocuteur dont provient l'enregistrement de test. Le dénominateur du rapport de vraisemblance équivaut à la densité de probabilité de l'élément de preuve E dans la distribution intra locuteur de l'enregistrement de test. Voir la figure **(Fig 3.3)**.

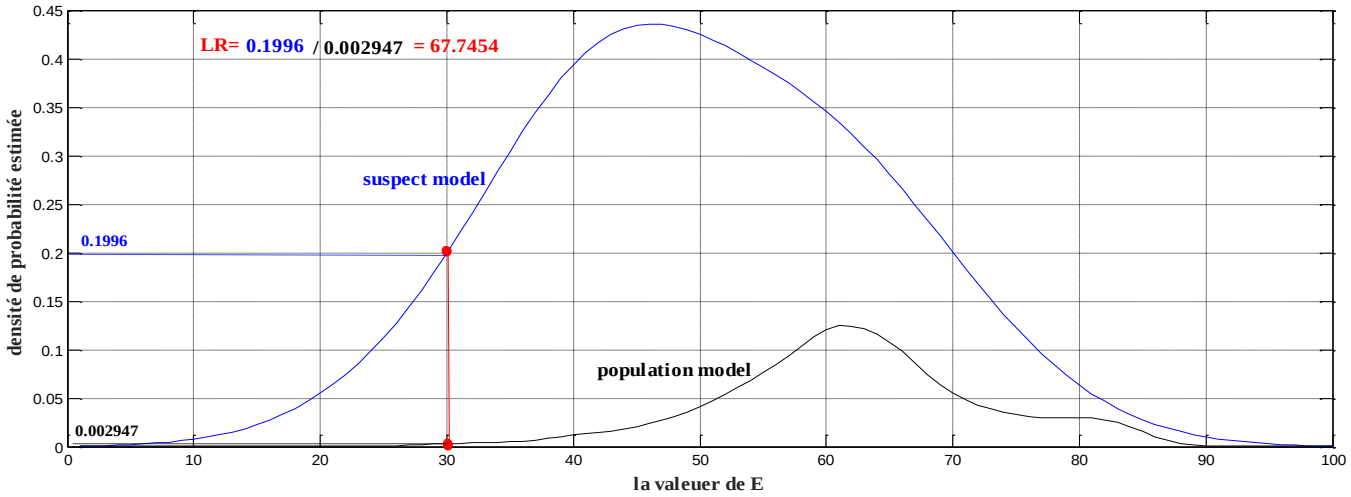
Dans cette évaluation les deux traces (enregistrements) de test existent dans la base de référence de locuteurs suspects.

Dans le cas de féminin la valeur de la vraisemblance LR est entre 100 à 1000 donc la preuve moyennement puissante pour l'hypothèse **H0** et dans le cas de masculin la valeur de la vraisemblance LR est entre 10 à 100 alors la preuve est modérée pour l'hypothèse **H0**.

La valeur de LR dans le cas féminin est supérieure à celui de masculin parce que phonologiquement parlant, la fréquence fondamentale (fréquence avec laquelle les cordes vocales vibrent en produisant les voyelles) des femmes, varie entre 250-400-Hz est supérieure à celle des hommes qui varie entre 80-200-Hz. La plage de variation de la fréquence fondamentale des femmes est plus riche en termes d'informations extralinguistiques que celle des hommes. D'où les résultats sont meilleurs pour le cas des femmes.



(a)



(b)

Fig 3.3. Résultat de l'évaluation de l'effet du genre des suspects sur les performances de système forensique :a)pour la base TIMIT féminin b) pour la base TIMIT.

3.7. Évaluation de l'effet du type de la langue parlée par le suspect sur les performances du système RALF

3.7.1. Évaluation sur les bases de données « Algerian Speech » et « TIMIT »

3.7.1.1. Procédure

Cette expérience est réalisée sur deux bases TIMIT et Algerian Speech. L'évaluation est faite avec la base TIMIT sur une sélection de 345 locuteurs (n° 0001 à 0345). L'enregistrement de test (trois traces) est comparé au modèle de la voix des 330 personnes de la population potentielle avec 15 personnes de la référence et de control de locuteurs suspects.

Par contre avec la base Algerian Speech, l'évaluation est faite sur une sélection de 60 locuteurs (n° 001 à 060). L'enregistrement de test (trois traces) est comparé au modèle de la voix des 45 personnes de la population potentielle avec 15 personnes de la référence et de control de locuteurs suspects.

3.7.1.2. Résultats et Discussion

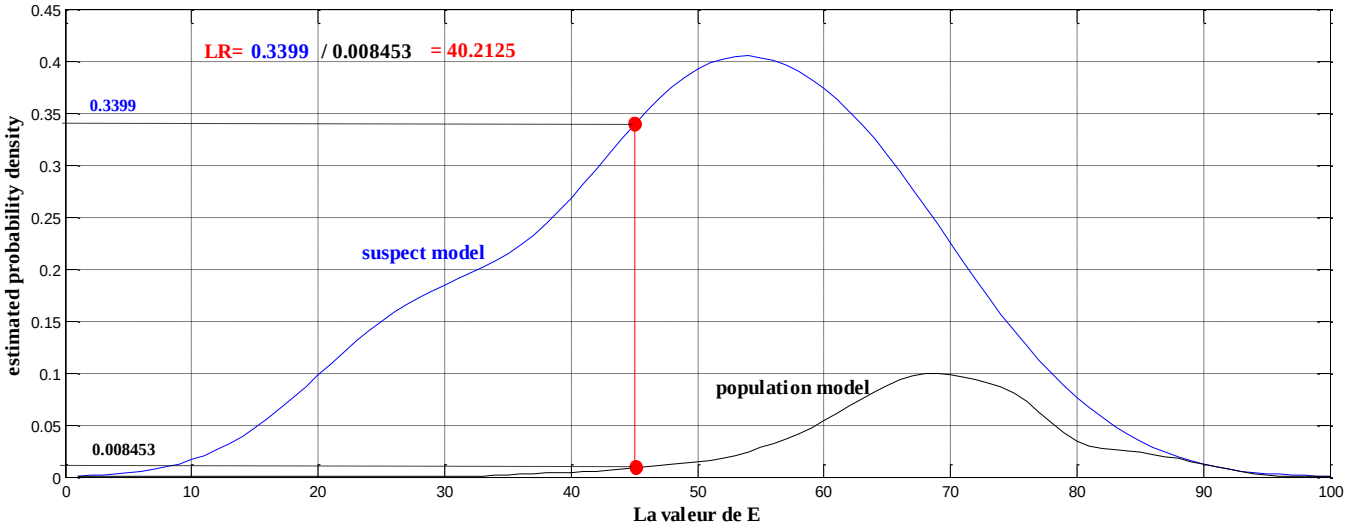
Tab 3.5. Calcul de la vraisemblance de E valant 135 dans la base de données Algerian Speech de la langue « arabe » et de E valant 45 dans la base TIMIT de la langue « anglais ».

La base de données	E	LR
Algerian Speech (arabe)	**135**	**1.5231e+003**
TIMIT (anglais américain)	**45**	**40.2125**

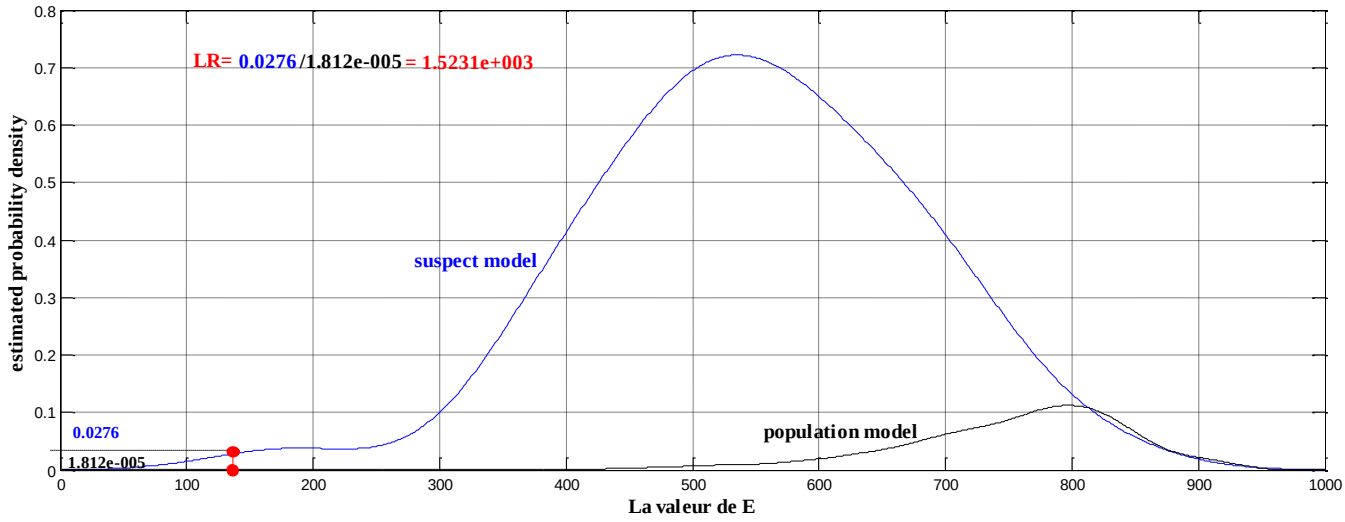
Les rapports de vraisemblance de ces éléments de preuve sont calculés de la manière suivante : le numérateur équivaut à la densité de probabilité de l'élément de preuve E dans la distribution de la variabilité interlocuteur dont provient l'enregistrement de test. Le dénominateur du rapport de vraisemblance équivaut à la densité de probabilité de l'élément de preuve E dans la distribution intra-locuteur de l'enregistrement de test. Voir la figure **(Fig 3.4).**

Dans cette évaluation les trois traces (enregistrements) de test existent dans la base de référence de locuteurs suspects pour les deux bases «Algerian Speech » et « TIMIT ».

Avec la base Algerian speech de la langue « Arabe », la valeur de la vraisemblance LR est entre 1000 à 10 000 donc preuve puissante pour l'hypothèse **H0**, par contre avec la base TIMIT de la langue « Anglais », la valeur de la vraisemblance LR est entre 10 à 100 alors la preuve modérée pour l'hypothèse **H0**. Voir le tableau (**Tab 3.5**). Cette surperformance des résultats obtenus avec la langue Arabe par rapport à la langue anglaise, s'explique par le faite que la langue Arabe est plus riche que la langue anglaise américaine en termes de phonèmes, fricatives, aspect de la coarticulation des phonèmes de la langue,…etc.



(a)



(b)

Fig 3.4. Résultats de l'évaluation de l'effet du type de la langue parlée par le suspect sur les performances du système RALF: a)pour la base TIMIT (anglais). b) pour la base Algerian Speech (arabe).

3.8. Évaluation de l'effet de la fréquence d'échantillonnage sur les performances du système RALF

3.8.1. Évaluation sur les bases de données « TIMIT » et « NIST »

3.8.1.1. Procédure

Cette expérience est réalisée avec deux bases TIMIT et NIST. Elle est appliquée à la base **TIMIT** sur une sélection de 345 locuteurs (n° 0001 à 0345). L'enregistrement de test (deux traces) est comparé au modèle de la voix des 339 personnes de la population potentielle avec 6 personnes de la référence et de control de locuteurs suspects, à une fréquence d'échantillonnage **16000Hz**. Pour la base **NIST,** une sélection sur 199 locuteurs (n° 0001 à 0199) est faite. L'enregistrement de test (deux traces) est comparé au modèle de la voix des 193 personnes de la population potentielle avec 6 personnes de la référence et de control de locuteurs suspect, à une fréquence d'échantillonnage **8000Hz.**

3.8.1.2. Résultats et Discussion

Tab 3.6. Calcul de la vraisemblance de E valant 12 pour la base de données TIMIT de fréquence d'échantillonnage 16000Hz et de E valant 72 pour la base de données NIST de fréquence d'échantillonnage 8000Hz.

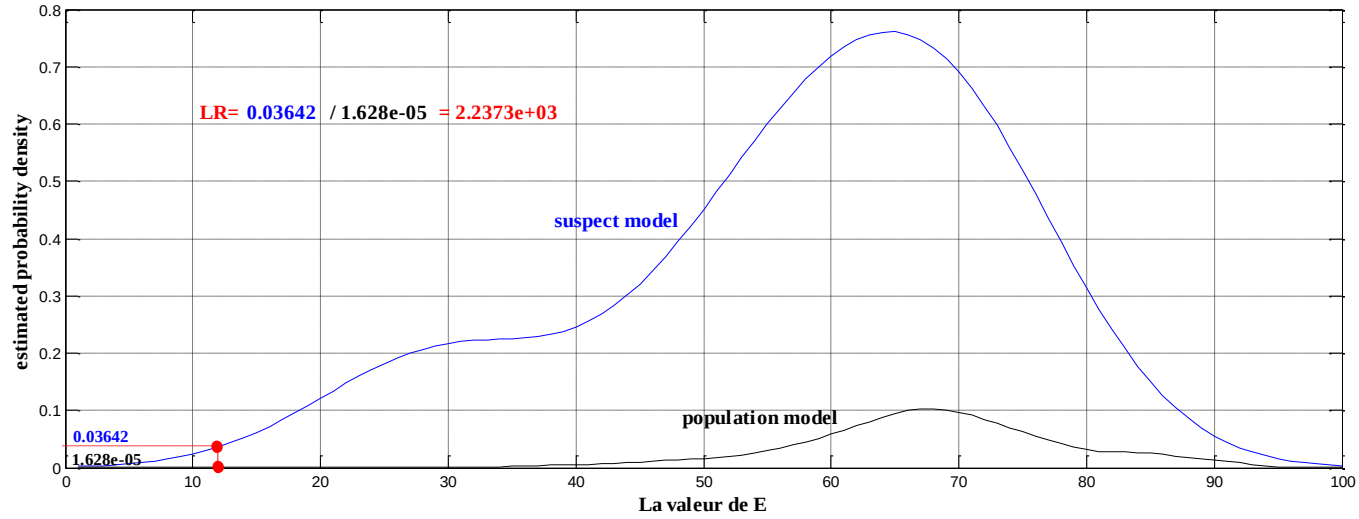
La base de données	E	LR
TIMIT (Fe=16000Hz)	**12**	**2.2373e+003**
NIST (Fe=8000Hz)	**72**	**0.5208**

Les rapports de vraisemblance de ces éléments de preuve sont calculés de la manière suivante : le numérateur équivaut à la densité de probabilité de l'élément de preuve E dans la distribution de la variabilité interlocuteur dont provient l'enregistrement de test. Le dénominateur du rapport de vraisemblance équivaut à la densité de probabilité de l'élément de preuve E dans la distribution intralocuteur de l'enregistrement de test. Voir la figure **(Fig 3.5)**.

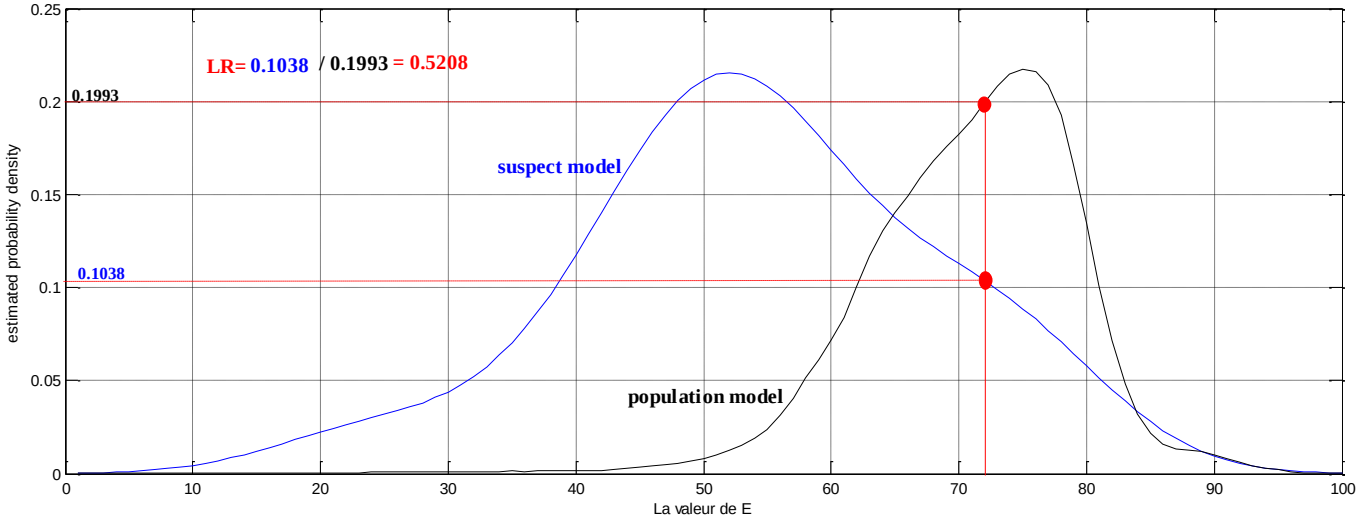
Avec la base TIMIT échantillonnée à 16000Hz, la valeur de la vraisemblance LR est entre 1000 à 10 000, donc preuve puissante pour l'hypothèse **H0.** Mais avec la base NIST

échantillonnée à 8000Hz, la valeur de la vraisemblance LR est entre 1et 10, alors la preuve limitée pour l'hypothèse **H0**.

Le LR dans le cas de la base TIMIT est très supérieur à celui de la base NIST, car la fréquence d'échantillonnage de la base NIST est celle d'une parole téléphonique (**Fe= 8000 Hz, Fe/2= 4000 Hz**). Donc, cette parole est bien affectée par les bruits du canal de transmission (bruit blanc ou gaussien) est aussi par les pertes d'information dues aux opérations de codage/décodage de source et canal. D'où la dégradation des performances en termes de valeur de LR.



(a)



(b)

Fig 3.5. Résultats de l'évaluation de l'effet de la fréquence d'échantillonnage sur les performances du système RALF: a)pour la base TIMIT (16000Hz). b) pour la base NIST (8000Hz)

3.9. Conclusion

Dans ce chapitre expérimental, nous avons développé un système de reconnaissance automatique de locuteurs en sciences forensique RALF basé sur le calcul du rapport de vraisemblance **LR**, qu'exprime une déclaration relative à la force de la preuve **E**. Pour bien évaluer le système RALF en termes de performances, nous avons effectué cinq évaluations à travers lesquelles nous avons étudié diverses influences, y compris l'existence de la trace, le genre du suspect, le type de la langue parlée par le suspect et la fréquence d'échantillonnage de la parole utilisée.

Conclusion Générale et Perspectives

Conclusion Générale et Perspectives

L'objectif principal du sujet abordé dans ce présent mémoire est d'étudier et évaluer le système de reconnaissance automatique du locuteur et principalement dans le domaine criminalistique. Pour pouvoir faire cette étude, nous avons utilisé trois bases de données TIMIT, NIST et Algerian Speech. Ces bases de données nous ont permis de faire plusieurs tests d'évaluation sur les différents scénarios pour le système RALF, à savoir : l'existence de la trace dans les bases (R) et (C), le sexe du suspect, le type de la langue parlée par le suspect et la fréquence d'échantillonnage de la parole utilisée.

L'idée principale sur laquelle est basé ce travail, est d'exprimer le résultat d'identification forensique du locuteur, sous forme d'un rapport de vraisemblance LR, entre deux hypothèses concurrentes. Pour se faire, l'approche Bayésienne est l'élément le plus important et le moyen très puissant qui a permis le calcul de ce rapport LR.

Le rapport LR est combiné avec la probabilité antérieure, calculée à partir d'opérations d'investigations et en analysant les circonstances de la trace en question, pour avoir la probabilité postérieure. Cette dernière sera adressée au juge qui peut, par la suite, faire son jugement.

Malgré que l'approche Bayesienne a été souvent critiquée par le fait de sa complexité d'une part, et de l'absence d'une méthode d'estimation de la probabilité antérieure d'une autre part, elle reste très utilisée dans plusieurs disciplines forensiques, et elle a donné de très bonnes performances dans nos expériences.

A travers les expériences que nous avons effectuées, nous pouvons dire que le modèle GMM-UBM est très puissant et peut représenter des distributions aléatoires très complexes d'une manière très fidèle.

Le bon choix de l'ordre du modèle GMM-UBM est très important. En effet, si nous choisissons un petit ordre, nous pouvons avoir une grande perte de données et par conséquent, une dégradation de performances. Dans le cas inverse, si nous choisissons un grand ordre, nous pouvons avoir le problème de sur-apprentissage du modèle GMM-UBM, c'est à dire présenter des données qui n'existent pas dans l'espace de paramètres acoustiques du locuteur en question. Par exemple, l'ordre d'un modèle UBM doit être suffisamment grand pour représenter l'ensemble des vecteurs acoustiques d'une population donnée, dans la pratique il peut aller jusqu'à 2048 gaussiennes et il est déterminé en fonction du nombre des locuteurs ainsi que la durée des enregistrements utilisés lors de la phase d'apprentissage du modèle.

Cependant, nos expériences ont montré qu'un modèle GMM-UBM composé de 256 gaussiennes est largement suffisant pour représenter la distribution des vecteurs acoustiques des suspects que nous avons simulés.

Comme perspectives de ce travail, nous envisagerons dans le futur :

- d'utiliser d'autres modèles d'apprentissage à savoir ; machines à support vecteurs (SVMs) et GMM-SVM.
- Aussi, l'investigation d'autres types de paramètres acoustiques (LPC, LPCC, LSF,…etc) extraits du signal de parole.
- d'étudier et d'évaluer la force de la preuve E en fonction des valeurs du rapport de vraisemblance LR.

Références

Références

[1] Laurent Besacier, (17 Avril 1998), « Un modèle parallèle pour la reconnaissance automatique du locuteur», Thèse de doctorat de l'Université d'Avignon et des Pays Vaucluse, École doctorale Mathématiques et informatique. page (1-134)

[2] Frank CRISPINO, (2006), « Le principe de Locard est-il scientifique ? Ou analyse de la scientificité des principes fondamentaux de la criminalistique », thèse de Doctorat, l'Institut de Police Scientifique de l'Ecole des Sciences Criminelles Université de Lausanne. page (1-166).

[3] Houda KADI, 18/06/2014, « La reconnaissance automatique du locuteur par la voix IP », master sciences et techniques, systèmes intelligents et réseaux. page (1-72).

[4] Alexandre PRETI, (décembre, 2008), Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur », thèse de Doctorat, École Doctorale 166 I2S «Mathématiques et Informatique», Laboratoire d'Informatique d'Avignon (EA 4128), page (1-176).

[5] Abdelghani HARRAG, (26/06/2011), « Extraction des données d'une base: Application à l'extraction des traits du locuteur », thèse de Doctorat, la faculté de technologie, Département Electronique, page (1-144).

[6] Akrouf, Samir, (19-oct-2014), « Une approche multimodale pour l'identification du locuteur », Thèse de doctorat, faculté des sciences, Département informatique, page (1-25)

[7] Frédéric DELSUC et Emmanuel J. P. DOUZERY, (2004), **«** les méthodes probabilistes en phylogénie moléculaire (2) L'approche bayésienne», biosystèmes, Avenir et pertinence des méthodes d'analyse en phylogénie moléculaire, Page (75 à 86).

[8] Othman LACHHAB, (15/04/2017), « Reconnaissance Statistique de la Parole Continue pour Voix Laryngée et Alaryngée», thèse de Doctorat, Formation doctorale: Informatique Structure de recherche: Équipe de recherche en Informatique et Télécommunications, page (1-110).

[9] Loïc BARRAULT, (Juillet 2008), « Diagnostic pour la combinaison de systèmes de reconnaissance automatique de la parole », thèse de Doctorat, École Doctorale 166 I2S « Mathématiques et Informatique », Laboratoire d'Informatique (EA 931), page (30-148).

[10] Simon Baechler, (2015), « From false identity documents to forensic intelligence Development of a systematic and transversal approach to process forensic data in support of crime intelligence», Thèse de doctorat ès sciences en science forensique, page (88-322).

[11] Sayoud Halim, (2003), « Reconnaissance Automatique du Locuteur-Approche connexionniste », thèse de Doctorat, Faculté d'Electronique et d'Informatique, page (1-229).

[12] ASBAI Nassim, (29/06/2015), « Identification et Authentification de Locuteurs, par les Techniques De Fusion des Paramètres et des Modèles dans un Environnement Rée», thèse de Doctorat ES-science, Faculté d'Electronique et d'Informatique, page (1-175).

[13] Yassine Mami, 2003, « Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence, Télécom ParisTech », thèse de Doctorat, Ecole nationale superieure des télécommunications, Spécialité signal et image, page (1-156).

[14] Tounsi bilal, (2008), « Inférence d'identité dans le domaine forensique en utilisant un système de reconnaissance automatique du locuteur adapté au dialecte Algérien», Institut National de Formation en Informatique », Mémoire de Magister Spécialité Informatique Industrielle (II), page (1-110).

[15] G. Pouchoulin, C. Fredouille, J.-F. Bonastre1, A. Ghio, A. Giovanni, « Analyse Phonétique dans le Domaine Fréquentiel pour la Classiffication des Voix Dysphoniques », 1'Université d'Avignon, Laboratoire Informatique d'Avignon (EA331), F-84018 Avignon (France) CNRS-LPL, Aix en Provence (France), page(1-4).

[16] Didier Meuwly, Andrzej Drygajlo, (2015), « Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM) », Institut de Police Scientifique et Criminologie, University of Lausanne, Switzerland, Signal Processing Laboratory, Swiss Federal Institute of Technology, Lausanne, page (1-7).

[17] Andrzej Drygajlo, Didier Meuwly and Anil Alexander, (2003), « Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognition », Speech Processing Group, EPFL, Lausanne, Switzerland, The Forensic Science Service, Birmingham, U.K, page (690.691).

[18] Geoffrey Stewart Morrison BSc, MTS, MA, PhD.Ewald Enzinger MPhil, PhD Cuiling Zhang BSc, MSc, PhD, (2017), « Forensic Speech Science », page (1-140).

[19] Andrzej Drygajlo, Michael Jessen, Stefan Gfroerer, Isolde Wagner, Jos Vermeulen and Tuija Niemi, (2015), « Methodological Guidelines for Best Practice in Forensic Semiautomatic And Automatic Speaker Recognition », European Network of Forensic Science Institutes, page (1-90).

[20] Hadjer Saboune, Mohamed DEBYECHE (novembre 2012), « Approche HMM multibandes Pour la reconnaissance du Locuteur de type GSM », Speech Communication and Signal Processing laboratory(LPCTS), Faculty of electronics and computer Sciences, USTHB P.O.Box 32, bab ezzouar, Algiers, Algeria.page (1-6).

[21] Shannon, C. E., (2001). «A mathematical theory of communication», ACM SIGMOBILE Mobile Computing and Communications Review, 5(1), page (3-55).

[22] Kim, D.S., Lee, S.Y., Kil, R.M., (1999). «Auditory processing of speech signal for robust speech recognition in real-world noisy environments», IEEE Transactions on Speech Audio Processing, vol.7, no.1, page (55–69).

[23] Noll, A., (1964). «Short-time spectrum and 'cepstrum' techniques for vocal-pitch detection. Journal of the Acoustical Society of America 36(2), page (430–451).

[24] Harris, F., (1978). «On the use of windows for harmonic analysis with the discrete Fourier transform. Proceedings of the IEEE, Vol. 66, No. 1, page (51-84).

[25] Kinnunen, T., Rajan, P., (2013). «A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data», Proc. Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2013, page (7229-7233).

[26] Ouassila Kenai, Mhania Guerti « Application Forensique à la Reconnaissance Vocale du Locuteur », Laboratoire Signal et Communications, Ecole Nationale Polytechnique, Alger, Algérie, page (54-61).

[27] K. Dash, A Novel Bpnn, (2012). « Approach for Speaker Identification Using Mfcc », Page (1-90)

[28] Koustav Chakraborty, Asmita Talele, Prof. Savitha Upadhya, (novembre 2014), « Voice Recognition Using MFCC Algorithm », International Journal of Innovative Research in Advanced Engineering (IJIRAE), page (158-161).

[29] Reda JOURANI, (6 septembre 2012), « Reconnaissance automatique du locuteur par des GMM à grande marge », thèse de Doctorat École doctorale et discipline ou spécialité ED MITT : Image, Information, Hypermedia, page (1-158).

[30] Waad BEN KHEDER, (juillet 2017), « Reconnaissance du locuteur en milieux difficiles », thèse de Doctorat École Doctorale 380 «Sciences et Agronomie» Laboratoire d'Informatique (EA4128), page (18-214).

[31] Djeghiour, S, Asbai, N, Kenai, O, & Guerti, M, (2018). « Forensic Automatic Speaker Recognition under Noisy Environments». 1st International Conference on Electronics and Electrical Engineering (IC3E'2018). University of Bouira, page (1-5).

[32] Didier Meuwly, (2001), « reconnaissance de locuteurs en sciences forensiques : l'apport d'une approche automatique », Thèse de doctorat l'Institut de Police Scientifique et de Criminologie de l'Université de Lausanne. page (1-245).