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 : **TELECOMUNICATION**

Spécialité : Systèmes de Télécommunication

Par

- **DERRARDJA INES**
- **BEN MALEK NOUR El HOUDA**

Intitulé

Identification du locuteur par GMM

Soutenu le : 19/09/2022

Devant le Jury composé de :

Nom & Prénom	*Grade*	*Qualité*	*Etablissement*
M. S BENDIB	*MCB*	*Président*	*Univ-BBA*
Mr. NASSIM ASBAI	*MCA*	*Encadreur*	*Univ-BBA*
M. HADJER BOUNAZOU	*Doctorant(e)*	*Co-Encadreur*	*Univ-BBA*
Mr. S AIDEL	*Professeur*	*Examinateur*	*Univ-BBA*

Année Universitaire 2021/2022

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.

Ma meilleure sœur Katreelnada et et Mon cher frère Mohamed Bachir.

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 , tous mes collègues d'étude.

A mon binôme Ines et toute la famille DERRARDJA. Et à tous ce ux 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.

Nour elhouda

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 ait pitié de toi mon père, 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.

Mon chére frére Didine,Mes très chères sœurs Zina , yasmine , Ahlem et ses enfants Aridj et wassim et son mari Abdelnasser .
Je tiens à remercier tout particulièrement mon chére ami Rabah Tairi Un soutien moral constant et ses nombreux conseils tout le temps accomplir 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 , tous mes collègues d'étude Mon chere ami Meriem Ben Meaamar .

A mon binôme NourElHouda et toute la famille BENMALEK. 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.

INES

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.

Nous remerciements vont aussi Co-Encadreur Doctarnte, HADJER BOUNAZOU, pour nous encourager et nous adier cette modeste étude

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.

Résumé

Ce travail est le résultat d’une évaluation d’un système d'identification automatique du locuteur, que nous avons développé et qui est basé sur la méthode de modélisation du locuteur GMM, Dans ce système, la tâche d’identification est dévolue au GMM-UBM. Pour cela, Nous avons effectué plusieurs expériences d’identification automatique du locuteur dans un milieu fermé et milieu ouvert.

Les expériences réalisées dans ce travail ont montré que l'utilisation de l’adaptation MAP par le modèle de mélange gaussien donne un meilleur taux d’identification. En revanche, une dégradation des performances est observée lorsque l'environnement où le système d'identification est opérationnel devient bruité.

Abstract

This work is the result of an evaluation of an automatic speaker identification system, which we have developed and which is based on the GMM speaker modeling method, In this system, the identification task is assigned to the GMM-UBM.For this purpose, we have carried out several experiments of automatic speaker identification in a closed environment and open environment.

The experiments carried out in this work showed that the use of the MAP adaptation by the Gaussian mixture model gives a better identification rate. On the other hand, a performance degradation is observed when the environment where the identification system is operational becomes noisy.

الملخص

هذا العمل هو نتيجة تقييم نظام التعرف التلقائي على السماعات ، والذي قمنا بتطويره والذي يعتمد على طريقة لهذا ، قمنا .GMM. UBMفي هذا النظام ، يتم تعيين مهمة تحديد الهوية إلى .GMMنمذجة مكبرات الصوت .بتنفيذ العديد من تجارب التعرف التلقائي على السماعات في بيئة مغلقة ومفتوحة

يعطي MAPأظهرت التجارب التي تم إجراؤها في هذا العمل أن استخدام نموذج الخليط الغاوسي للتكيف مع معدل تعريف أفضل. من ناحية أخرى ، يتم ملاحظة تدهور الأداء عندما تصبح البيئة التي يعمل فيها نظام تحديد الهوية صاخبة

Table des matières

Liste des figures

Liste des tableaux

Abréviations

RAL	Reconnaissance automatique de locuteurs.
LR	Rapport de vraisemblance ou Likelihood ration
GMM	Gaussian Mixtures Models.
MFCC	Mel-Frequencies Cepstral Coefficients.
LPC	Linear Predictive Coding
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.
EM	Expectation Maximization.
UBM	Universal Background model.
TIMIT	Texas Instruments Massachusetts Institute of Technology.

Introduction Générale

Introduction Générale

La parole est un moyen de communication universel, que la technologie permet aujourd'hui de diffuser, stocker et restituer à une échelle planétaire. Le traitement du signal de parole est l'enjeu d'un grand nombre d'applications, dans des domaines aussi variés que la sécurité, le pilotage de machines ou l'indexation de documents électroniques [1].Le signal de parole est porteur de plusieurs types d'informations comme le message, la langue, les émotions, ou même l'environnement. Cet avantage a donné naissance à plusieurs travaux de recherche dont l'objectif est la conception des systèmes de reconnaissances [2].

L'utilisation de la parole est considérée de nos jours comme une des formes les plus simples de la technologie biométrique, car elle n'est pas intrusive, et ne demande aucun contact physique avec le locuteur à identifier. Certes, la voix n'est pas aussi efficace et fiable que peuvent l'être les empreintes digitales, l'iris, ou encore l'ADN. Cependant, elle est dans certains cas le seul moyen d'identification d'un individu [3].

La reconnaissance (identification) automatique du locuteur (RAL ou IAL)) désigne l'ensemble des procédures automatiques visant à discriminer des locuteurs à partir de leurs énoncés de voix. Elle s'appuie sur la théorie du signal et sur des techniques d'apprentissage automatique. Ces dernières privilégient à la définition de règles la validation statistique.

Des représentations numériques du signal de parole, ainsi que des méthodes de détection, sont élaborées afin d'évaluer des hypothèses sur l'identité d'un locuteur présumé. Une majorité des approches mathématiques mises en œuvre en RAL (IAL) reposent sur des approches probabilistes [1].

L'avantage des systèmes IAL est qu'ils sont indépendants du texte, Indépendants de la langue du discours, et l'identification du locuteur est totalement automatisée et ne nécessite aucune intervention humaine.

Dans ce contexte, nous nous intéressons à l'identification du locuteur basée sur les modèles de probabilités à savoir Modèle de Mélange de Gaussien (GMM) , qui fait l'objet des recherches les plus récentes en termes d'amélioration des performances. En pratique, de nombreux facteurs liés aux conditions d'enregistrement peuvent dégrader significativement les performances de ces systèmes. Ces facteurs peuvent

être liés à l'environnement (bruit supplémentaire, écho, etc.), au dispositif d'enregistrement (changement de canal) ou au locuteur lui-même (état psychologique, tension vocale, changement de voix, etc.). Souvent, ces facteurs ne peuvent pas être connus à l'avance, ce qui présente un défi pour les applications réelles [4].

L'objectif de ce travail, Validation des résultats obtenus dans l'article.

Ce mémoire est composé de trois chapitres, organisés comme suit :

Dans le premier chapitre de ce travail, nous avons introduit des généralités sur le signal de parole pour comprendre les différents concepts qui tournent autour de lui. Dans le deuxième chapitre nous nous sommes intéressées au modèle GMM (Gaussian Mixture Model) le plus utilisé dans les systèmes d'identification automatique du locuteur et détaillé ses différentes étapes. Alors que le troisième chapitre contient l'ensemble des tests effectués et les résultats que nous avons obtenus.

Enfin nous terminerons ce mémoire par une conclusion générale qui résume les résultats obtenus au cours de notre travail.

Chapitre 1

Analyse numérique du signal de parole et identification du locuteur

1.1. Introduction

La parole est le principal moyen de communication dans toute société humaine. L'importance de la parole fait que toute interaction homme-machine devrait plus ou moins passer par elle [5].Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur, Le but essentiel du traitement du signal vocal est de donner une représentation moins redondante de la parole, tout en permettant une extraction assez précise des paramètres pertinents qui caractérisent le signal de la parole. Dans ce chapitre, Nous définissons brièvement certaines caractéristiques de ce signal ainsi que les différentes approches de traitement numérique utilisées pour l'analyser et les différentes étapes qui permettent l'extraction des paramètres cepstraux MFCCs. Ces derniers (MFCCs), sont utilisés par la suite comme étant des vecteurs caractérisant les locuteurs dans la tache de l'identification (IAL). D'où, une définition et explication de l'identification automatique du locuteur (IAL) sont détaillées dans ce chapitre.

1.2. La parole

La parole est un signal réel continu, aléatoire, redondant à énergie finie, non stationnaire et variable dans le temps [6].

1.2.1. Caractéristique du Signal de Parole

Pour en revenir sur les caractéristiques principales du signal vocal il faut bien comprendre la nature de la parole en temps qu'un phénomène vibratoire produit par un système articulatoire de l'appareil phonatoire humain. On peut résumer les principales caractéristiques du signal vocale dans ce que suit :

- **Redondance**

Le signal de parole présente un caractère redondant et il renferme plusieurs types d'informations ; les sons, la syntaxe. Si cette redondance lui confère une bonne résistance au bruit, elle oblige à extraire du signal les informations pertinentes, en essayant de ne pas trop les dégrader. [7]

- **La continuité**

Il est continu dans le temps, ce qui nécessite une discrétisation préalable du signal.

- **La variabilité du signal**

Le signal de parole possède une très grande variabilité. Les origines de cette variabilité sont diverses. Beaucoup sont intrinsèques à la nature même du langage parlé [6].

- **La non stationnarité de signal**

Le concept de processus stationnaire n'est qu'un modèle simplifié, car un phénomène physique n'est jamais rigoureusement stationnaire (il est souvent influencé par l'évolution du système physique auquel il est associé). Ce modèle est toutefois commode d'emploi et se révèle très largement utilisable en pratique lorsque l'on peut se contenter d'observer les durées limitées pendant lesquelles le phénomène présente un caractère permanent. Le signal de parole est un signal non stationnaire ce qui explique sa complexité ; afin de simplifier son étude on doit le considérer étant localement stationnaire (quasi-stationnaire). [8]

1.3. Prétraitement acoustiques

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

1.3.1. 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 [9].

• 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 « fe », est supérieure ou égale à la fréquence maximum du spectre du signal notée ' fmax'.

$$fe \geq 2fmax \tag{1.1}$$

Avec $$fe = \frac{1}{Te}$$

1.3.2. Préaccentuation et Fenêtrage

1.3.2.1. 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 [10].

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

$$H(z) = 1 - a.z^{-1}, 0.95 \leq n \leq 1 \quad (1.2)$$

1.3.2.2. 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 [11]. Cette analyse est effectuée à l'aide de fenêtres [12] telles que :

- Fenêtre Hamming

$$w_n = 0{,}54 - 0{,}46.\cos(2\pi \frac{n}{N-1}), \quad 0 \leq n \leq N-1 \quad (1.3)$$

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

N **:** la taille de la fenêtre.

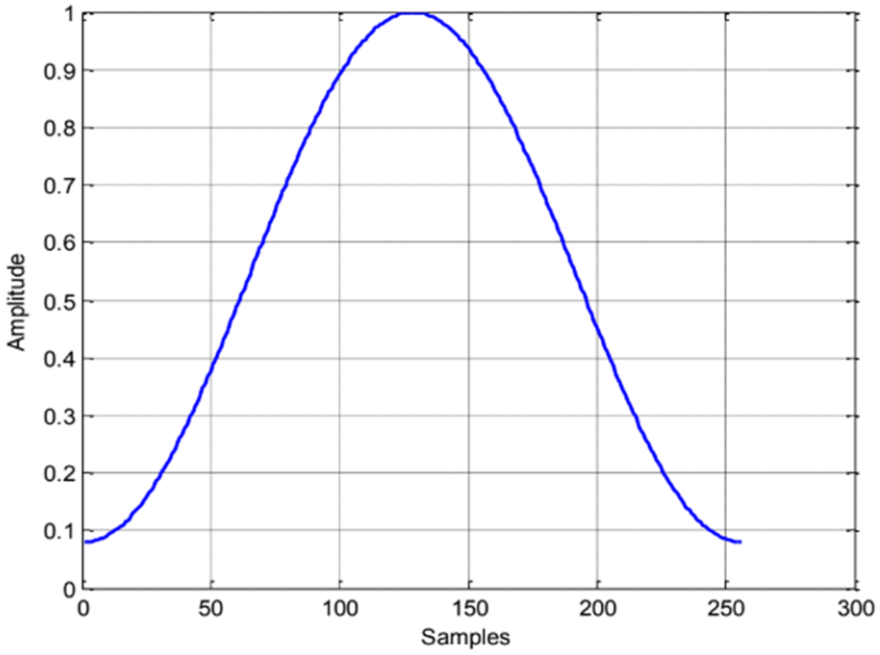

Fig.1. 1. Allure temporelle de la fenêtre Hamming

Cette fenêtre de Hamming est souvent utilisée, vu que son spectre n'introduit pas trop de distorsion sur le signal vocal. [13]

1.4. Méthodes de traitement

1.4.1. Analyse temporelle

- **Energie**

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 :

$$E = \frac{1}{N}\sum_{k}^{N-1} x^2(k) \tag{1.4}$$

Avec E : la valeur à évaluer.

N : la largeur de la fenêtre d'analyse.

x(k) : le signal numérique.

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

- **Taux de passage par zéro (TPZ)**

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

$$TPZ = \frac{1}{2}\sum_{k=0}^{k-1}\left|sign(x(k+1)) - sign(x(k))\right| \tag{1.5}$$

x(k) : le signal numérique.

k : l'échantillon de la trame.

- **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. 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. [14]

1.4.2. Analyse fréquentielle

L'analyse fréquentielle est la plus couramment utilisée. Elle est avantageuse pour deux points de vue :

- L'étude de l'audition montre que l'oreille effectue une sorte d'analyse fréquentielle.
- Cette forme d'analyse permet une représentation plus fidèle du signal vocal. [15].

1.4.2.1. La Transformée de Fourier Discrète

Pour effectuer cette analyse on utilise :

$$X(n) = \sum_{k=0}^{N-1} x(k) \times e^{-j\pi \frac{nk}{N}} \qquad (1.6)$$

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.

1.5. Variabilité de la parole

À contenu phonétique égal, le signal vocal est très variable pour un même locuteur (variabilité intra locuteur) ou pour des locuteurs différents (variabilité interlocuteur)

1.5.1. Variabilité intra-locuteur

Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution en détermine la durée. Toute affection de l'appareil phonatoire peut altérer la qualité de la production, et l'intensité de l'onde sonore fléchit, l'articulation perd de sa clarté. La diction évolue dans le temps: l'enfance, l'adolescence, l'âge mûr, puis la vieillesse, autant d'âges qui marquent la voix de leurs sceaux.

1.5.2. Variabilité interlocuteurs

La grande variabilité entre les locuteurs est due, d'une part, à l'héritage linguistique et au milieu socioculturel de l'individu, et d'autre part aux différences physiologiques

des organes responsables de la production vocale. L'expression acoustique de ces différences peut être traduite par une variation de la fréquence fondamentale, dans l'échelle des formants. [13]

1.6. Extraction des paramètres

Les coefficients cepstraux sur l'échelle Mel (MFCC, Mel-Frequency Cepstral coefficients) ont été intensivement utilisés comme vecteur de traits caractéristiques dans les systèmes de reconnaissance de la parole et du locuteur. [16]

L'extraction de coefficients MFCC est basée sur l'analyse par banc de filtres qui consiste à filtrer le signal par un ensemble de filtres passe-bande. L'énergie en sortie de chaque filtre est attribuée à sa fréquence centrale. Pour simuler le fonctionnement du système auditif humain, les fréquences centrales sont réparties uniformément sur une échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'information utile dans le signal de parole [8] [16].

1.6.1. Paramètres MFCC

Dans ce qui suit, nous décrivons chacune des étapes nécessaires pour l'obtention d'un vecteur caractéristique tiré des coefficients MFCC, tel qu'illustré par la Figure 3

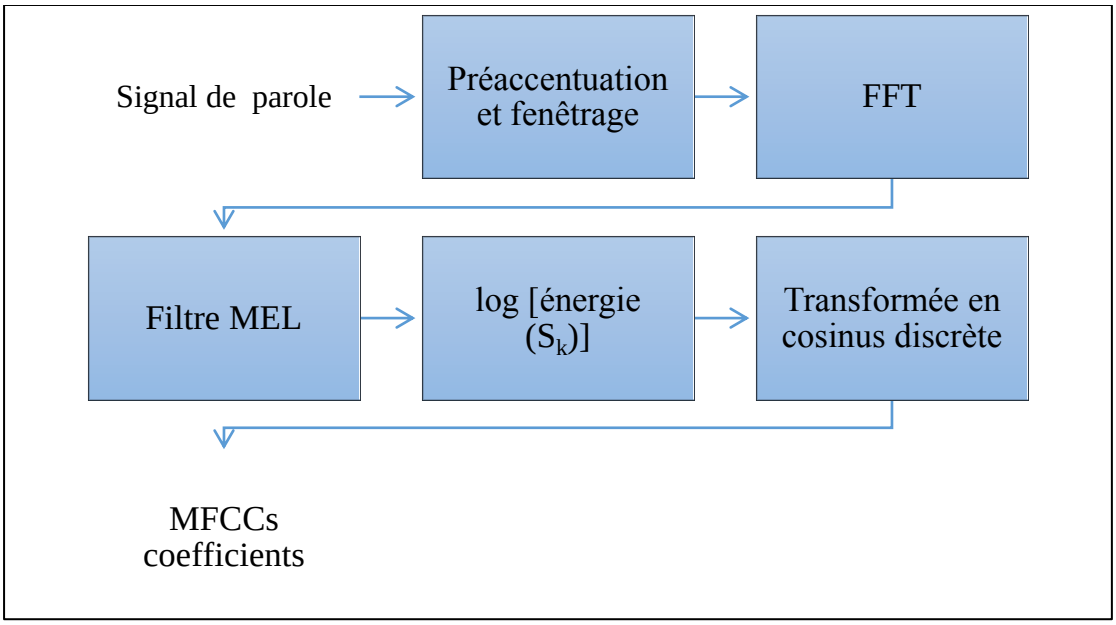

Fig.1.2. Étapes d'extraction des vecteurs caractéristiques MFCCs

Les MFCCs 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 S_k l'énergie du signal à la sortie du filtre K, nous avons maintenant m_p (le nombre de filtres) paramètres S_k. (Des études ont montré que les 20 premiers paramètres de chaque trame extraits du filtre Mel représentent très bien le locuteur).
- le logarithme de S_k 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).
- 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.

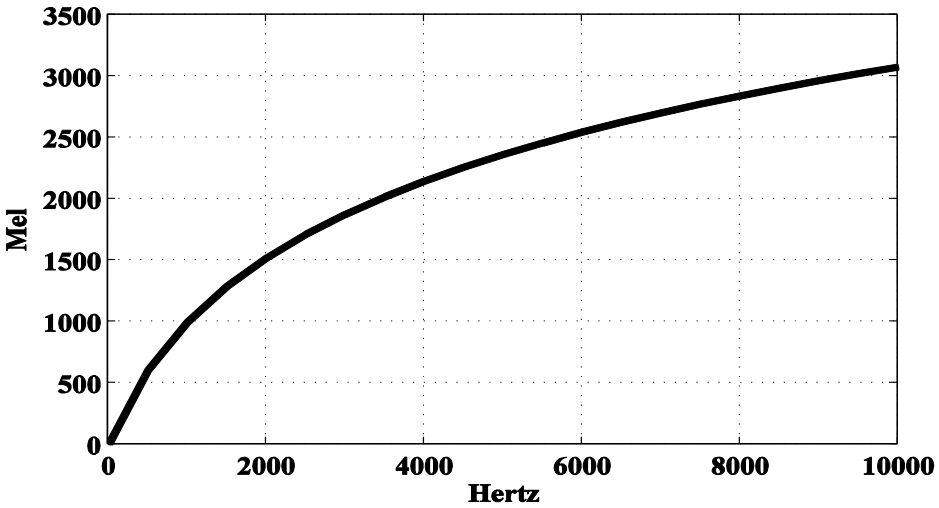

Fig.1.3. Echelle Mel

1.7. Identification Automatique du Locuteur (IAL)

L'Identification Automatique du Locuteur (IAL) consiste à déterminer, à partir d'un ensemble de locuteurs référencés dans le système, l'identité du locuteur présent dans

un signal vocal (signal de test) [17], [18]. Pour cela, le système calcule des mesures de similarités entre ce signal et tous les modèles des locuteurs de la base. Deux conditions d'identification sont connues : milieu fermé et milieu ouvert. Dans le cas où le système doit fournir un ensemble d'au moins un locuteur, on parle d'une identification en milieu fermé. Mais dans certaines applications, le système peut être amené à fournir un ensemble vide : c'est l'identification en milieu ouvert. En milieu fermé, chaque accès de test est comparé à tous les modèles des locuteurs référencés dans le système. L'identité du locuteur possédant la référence la plus proche est émise en sortie du système.

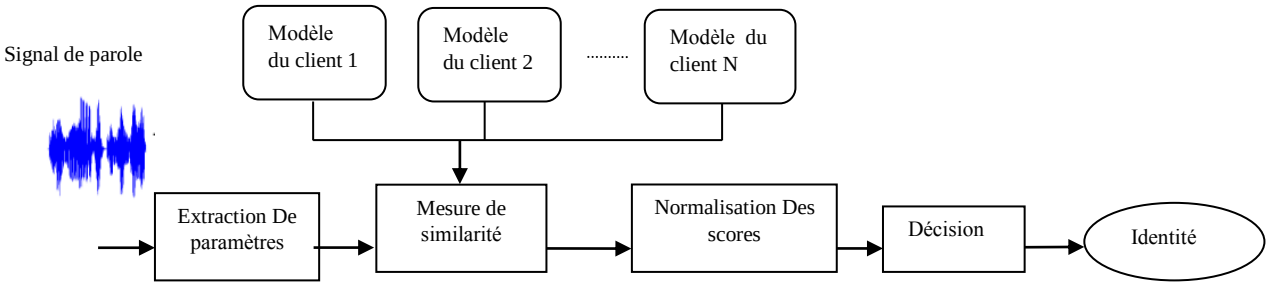

Fig.1.4 Schéma modulaire d'un système d'IAL.

1.7.1. Mode dépendant et indépendant du texte

Les modes d'IAL sont données par le mode dépendant du texte et le mode indépendant du texte.

1.7.1.1. Mode dépendant du texte

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

1.7.1.2. Mode indépendant du texte

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

1.8. Conclusion

Nous avons présenté dans ce chapitre l'analyse acoustique du signal audio puis les étapes nécessaires à l'obtention des coefficients cepstraux, aussi appelés MFCC. Ces étapes sont le fenêtrage, la transformée rapide de Fourier, un passage dans le banc de filtres à l'échelle Mel de type triangulaire, puis transformée en cosinus (DCT). Ces coefficients MFCCs sont très pertinents pour la tâche de l'identification du locuteur, dont des explications et les définitions sont bien accordées au lecteur dans ce chapitre.

Chapitre 2

Modélisation GMM des données du locuteur

2.1. Introduction

Le modèle de mélange gaussien ou mélange de gaussiens, comme on l'appelle parfois, n'est pas tant un modèle qu'une distribution de probabilité. Il s'agit d'un modèle universellement utilisé pour l'apprentissage génératif non supervisé ou le clustering. Il est également appelé Expectation-Maximization Clustering ou EM Clustering et est basé sur la stratégie d'optimisation [19]. Les modèles de mélange gaussien sont utilisés pour représenter des sous-populations (classes) normalement distribuées au sein d'une population globale (classe globale)). L'avantage des modèles de mélange est qu'ils ne nécessitent pas à quelle sous-population appartient un point de données. Il permet au modèle d'apprendre automatiquement les sous-populations. Cela constitue une forme d'apprentissage non supervisé (apprentissage non étiqueté (ou indexé) par l'utilisateur).

2.2. Modèle de mélange de gaussiennes (GMM)

Le modèle de mélange de gaussiennes fait partis des méthodes de classification paramétrique globale. Il consiste à supposer que la distribution des données peut être décrite comme une somme pondérée de densités gaussiennes. Chaque gaussienne du modèle est caractérisée par son poids, son vecteur moyenne et sa matrice de covariance [20].

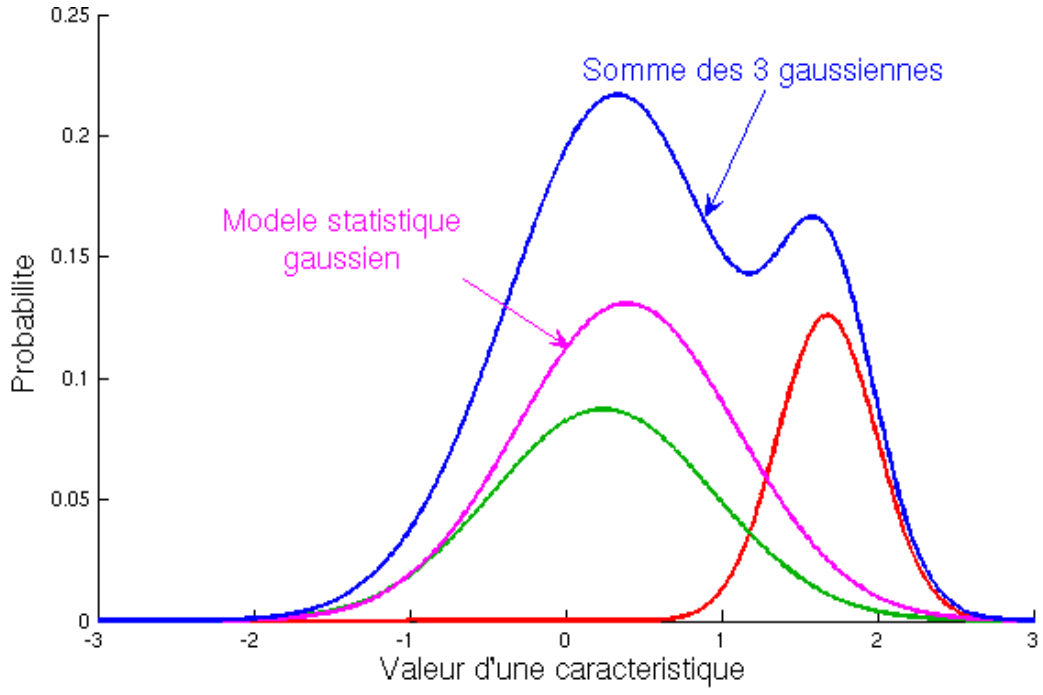

Fig.2.1 Exemple de modèle de mélange de 3 gaussiennes[recueillies à partir d'internet]

Soit une cible *s* et un vecteur acoustique *x* de dimension d, le mélange de gaussienne est défini comme suit [21] :

$$p(x|\lambda) = \sum_{m=1}^{M} \pi_m^s \; b_m^s(x) \tag{2.1}$$

où :

- $b_m^s(x)$: la densité gaussienne paramétrée par le vecteur moyen μ_m^s et la matrice de covariance Σ_M^s. Cette densité est donnée par:

$$b_m^s(x) = \frac{1}{(2\pi)^{d/2} |\sum_M^s|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu_m^s)(\sum_m^s)^{-1}(x - \mu_m^s)\right] \tag{2.2}$$

- π_m^s : le poids de mélange, avec la contrainte: $\Sigma_{m=1}^{M} \pi_m^s = 1$

Une cible est complètement spécifiée par l'ensemble des paramètres noté λs [18] [19] [20].

$$\lambda_s = (\pi_m^s, \mu_m^s, \Sigma_M^s) \tag{2.3}$$

Pour la modélisation des émotions, chaque émotion est représentée par son vecteur de paramètres dans le modèle λ_s. Ce modèle peut prendre plusieurs formes, dépend du choix de la matrice de covariance. On peut assigner une matrice de covariance à chaque gaussienne, une matrice de covariance pour chaque modèle ou bien utiliser une matrice de covariance global pour tous les modèles.

2.3. L'intérêt de la modélisation GMM

L'utilisation des GMM pour la modélisation des signaux audio a été initiée par les travaux de Douglas Reynolds [20]. Cette approche à donner, depuis plus de vingt ans maintenant, les meilleures performances pour les systèmes de reconnaissances du locuteur basés sur l'approche probabiliste. La plupart des systèmes actuels qui traitent la reconnaissance des signaux audio utilisent une modélisation de GMM. Ceci n'est pas un hasard, deux raisons principales font des densités de mélanges de gaussiennes une modélisation incontournable pour la présentation de l'identité de locuteur.

La première raison est notion intuitive, étant donné que les densités multimodales, telles que GMM, permettent de modéliser les ensembles fondamentaux des paramètres acoustiques.

Puisque rappelons que ces vecteurs acoustiques issus de la phase de para métrisation du signal de parole ont une distribution complexe, les modèles simples mono-gaussiens sont insuffisants pour prendre en compte tous les détails de cette distribution. Aussi, l'espace acoustique peut être caractérisé par un groupe de classes acoustiques (nuages) représentant de larges événements phonétiques, ces classes acoustiques sont le reflet des configurations générales du conduit vocal (déformations physiques) qui caractérisent l'identité de la personne. Le positionnement du ième nuage acoustique (zone de concentration de vecteurs acoustiques) peut être représenté par la moyenne de la ieme composante de la densité multimodale, tandis que sa forme (dispersion : sphérique, ovale...) et son orientation (horizontale, verticale...) sont définies par la matrice de covariance.

L'autre raison d'utilisation des densités de mélanges de gaussiennes, est qu'une combinaison linéaire de mono gaussiennes est capable de représenter une large gamme de distribution. La puissance des GMMs est son habilité à approximer fidèlement des distributions aléatoires que celles des paramètres acoustiques [25].

2.4. Apprentissage

L'apprentissage a pour but d'estimer, à partir des données extraites des segments de paroles, les paramètres du GMM qui donnent la meilleure distribution possible des vecteurs acoustiques.

L'apprentissage des différents paramètres d'un GMM est classiquement réalisé par un algorithme de type EM (Expectation-Maximisation) pour déterminer les paramètres du modèle qui maximisent la vraisemblance des données d'apprentissage [20]. En effet pour une séquence de N vecteurs d'apprentissage $x = \{x_1, x_2, \dots, x_N\}$ suffisamment indépendants, le maximum de vraisemblance du GMM est donné par :

$$p(X|\lambda_S) = \Pi_{n=1}^{N} p(x_n|\lambda_S) = \Pi_{n=1}^{N} \Sigma_{m=1}^{M} p(x_n/\pi_m^s, \mu_m^s, \Sigma_M^s) \quad (2.4)$$

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 :

$$V(X, \lambda_s) = \frac{1}{N} \log \Pi_{n=1}^{N} p(x_n|\lambda_s) = \frac{1}{N} \Sigma_{n=1}^{N} log p(x_n|\lambda_s) \quad (2.5)$$

Donc on a une expression de la vraisemblance complexe contenant le logarithme d'une somme et une fraction non linéaire des paramètres du modèle λ ce qui rend la maximisation directe très difficile.

Cependant, la variable indicatrice « m » est une donnée constitutive du problème qui présente l'inconvénient de ne pouvoir être observée en pratique. En effet on observe des réalisations du vecteur aléatoire x_n sans savoir de manière certaine quelle est la classe du mélange associée à chaque observation. Au sens de l'algorithme EM (Expectation- Maximisation), la variable « m » constitue une donnée manquante ou non- observée.

2.4.1. Apprentissage par Maximum de Vraisemblance

L'algorithme EM (Expectation-Maximisation) fait intervenir à la fois des observations *X* et des variables manquantes (l'indice de la gaussienne $m = 1, \ldots, M$). Cet algorithme maximise, de façon itérative, la fonction de la vraisemblance. Cette maximisation n'est pas directe, elle fait intervenir la fonction auxiliaire $(\theta,\theta^{(t)})$ qui est définie comme étant l'espérance mathématique du logarithme de la vraisemblance jointe (incluant les variables observées et les variables cachées) sur l'ensemble complet des variables d'entraînement, calculée sur base des paramètres courants [20], cette fonction auxiliaire s'exprime de la manière suivante:

$$Q\left(\theta,\theta^{(t)}\right) = \Sigma_{m=1}^{M}\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\log p(x_n,m|\theta) \qquad (2.6)$$

Où θ désigne l'ensemble des paramètres à estimer (π_m, μ_m, Σ_m) et $\theta^{(t)}$ l'ensemble des paramètres estimés à l'itération *t.* Ce qui donne :

$$Q\left(\theta,\theta^{(t)}\right) =$$
$$\Sigma_{m=1}^{M}\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\log p(m|\theta) + \Sigma_{m=1}^{M}\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\log(x_n|m,\theta) \qquad (2.7)$$

On remplace $p(x_n|m, \theta)$ par sa valeur ,on trouve :

$$Q\left(\theta,\theta^{(t)}\right) = \Sigma_{m=1}^{M}\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\log p(x_n,m|\theta) + \Sigma_{m=1}^{M}\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\log(\pi_m \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_m|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(x-\mu_m)^t(\Sigma_m)^{-1}(x-\mu_m)^t\right] \quad (2.8)$$

$$Q\left(\theta,\theta^{(t)}\right) = \Sigma_{m=1}^{M}\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\left[log\pi_m - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_m|\right] - \Sigma_{m=1}^{M}\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\left[\frac{1}{2}(x_n-\mu_m)^t\Sigma_m^{-1}(x_n-\mu_{m)}\right] \quad (2.9)$$

En supposant que p($x_n|\theta$) sont des densités gaussiennes à matrice de covariance diagonale, l'expression (3.6) devient :

$$Q\left(\theta,\theta^{(t)}\right) = \Sigma_{m=1}^{M}\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right) log\pi_m - \frac{1}{2}\Sigma_{m=1}^{M}\ \Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\left[Cste + log{\sigma_m}^2 + \frac{(X_n-\mu_m)^2}{{\sigma_m}^2}\right] \quad (2.10)$$

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

Les paramètres sont estimés en annulant la dérivée partielle de la fonction auxiliaire Q par rapport à chacun de ceux-ci. Dans le cas de la moyenne μ_m, nous avons donc:

$$\frac{\partial Q}{\partial \mu_m} = \Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\left[\frac{(x_n-\mu_m)}{{\sigma_m}^2}\right] = 0 \quad (2.11)$$

En multipliant le tout par σ_m^2

$$\Sigma_{n=1}^{N} x_n p\left(m\middle|x_n,\theta^{(t)}\right) = \Sigma_{n=1}^{N}\mu_m p\left(m\middle|x_n,\theta^{(t)}\right) \quad (2.12)$$

Le nouvel estimateur de la moyenne devient donc :

$${\mu_m}^{(t+1)} = \frac{\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)x_n}{\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)} \quad (2.13)$$

En ce qui concerne la variance, nous avons donc :

$$\frac{\partial Q}{\partial \sigma_m} = \Sigma_{n=1}^{N} p\left(m\middle|x_n, \theta^{(t)}\right)\left[\frac{(x_n-\mu_m)^2}{{\sigma_m}^3} - \frac{1}{\sigma_m}\right] = 0 \quad (2.14)$$

En multipliant le tout par σ_m :

$$\frac{\partial Q}{\partial \sigma_m} = \Sigma_{n=1}^{N} p\left(m\middle|x_n, \theta^{(t)}\right)\left[\frac{(x_n-\mu_m)^2}{{\sigma_m}^2} - 1\right] = 0 \quad (2.15)$$

L'estimateur de la variance devient :

$${\sigma_m}^{2(t+1)} = \frac{\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\left(x_n-{\mu_m}^{(t)}\right)^2}{\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)} \quad (2.16)$$

Dans le cas de matrice de covariance pleines, un raisonnement similaire conduit à :

$${\sigma_m}^{2(t+1)} = \frac{\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)\left(x_n-{\mu_m}^{(t)}\right)^2}{\Sigma_{n=1}^{N} p\left(m\middle|x_n,\theta^{(t)}\right)} \quad (2.17)$$

L'estimation des poids des composantes de mélange π_m est assez simple puisqu'il s'agit de paramètres scalaires. Ceci dit, il faut cependant tenir compte de la contrainte existe sur ces paramètres$\Sigma_{m=1}^{M}\pi_m = 1$. La maximisation sous contrainte se résout simplement en introduisant un multiplicateur de Lagrange associé à cette contrainte [24]. La fonction à maximiser devient alors :

$$Q^*\left(\theta, \theta^{(t)}\right) = Q\left(\theta, \theta^{(t)}\right) + \tau(\Sigma_{m=1}^{M}\pi_m - 1) \quad (2.18)$$

Où τ représente le multiplicateur de Lagrange. En annulant la dérivée partielle de Q^* par rapport π_m (ce qui fait disparaître les termes contenant les moyennes et les variances), nous obtenons :

$$\frac{\partial Q^*}{\partial \pi_m} = \frac{1}{\pi_m}\Sigma_{n=1}^{N} p\left(m\middle|x_n, \theta^{(t)}\right) + \tau = 0 \quad (2.19)$$

En sommant cette expression sur toutes les composantes m, nous obtenons que $\tau = -N$, ce qui nous donne alors :

$$\pi_m{}^{(t+1)} = \frac{1}{N}\Sigma_{n=1}^{N} p\left(m\middle|x_n, \theta^{(t)}\right) \tag{2.20}$$

Dans laquelle les valeurs de $p\left(m\middle|x_n, \theta^{(t)}\right)$ peuvent simplement être obtenues par la loi de Bayes :

$$p\left(m\middle|x_n, \theta^{(t)}\right) = \frac{p\left(m\middle|\theta^{(t)}\right)p\left(x_n\middle|\theta^{(t)}\right)}{\Sigma_{n=1}^{N} p\left(K\middle|\theta^{(t)}\right)p\left(x_n\middle|\theta^{(t)}\right)} \tag{2.21}$$

Avec $p\left(m\middle|\theta^{(t)}\right) = \pi_m{}^{t}$ donc :

$$\gamma(n,m)^{(t)} = p\left(m\middle|x_n, \theta^{(t)}\right) = \frac{\pi_m{}^{(t)} p\left(X_n\middle|\mu_m{}^{(t)}, \Sigma_m{}^{(t)}\right)}{\Sigma_{K=1}^{M} \pi_K{}^{(t)} p\left(X_n\middle|\mu_K{}^{(t)}, \theta^{(t)}\right)} \tag{2.22}$$

2.5. Maximum *a Posteriori* (MAP) Adaptation

Pour surmonter le manque de données d'entraînement (apprentissage) dans l'estimation des locuteurs par GMM, l'adaptation *Maximum a Postériori* (MAP) [25] du modèle de monde (UBM) est utilisée. Cette approche représente l'état de l'art de la reconnaissance du locuteur en mode indépendant du texte. L'adaptation d'un modèle GMM, consiste à modifier les paramètres initiaux du modèle vers un autre modèle plus spécifique, relatif aux données d'adaptation. Plus précisément, étant donné un énoncé d'apprentissage avec une séquence de vecteurs acoustiques $X = \{x_1, x_2, ..., x_N\}$, les formules suivantes sont appliquées aux vecteurs moyens μ_i de l'UBM pour obtenir les vecteurs moyennes adaptées $\bar{\mu}_i$:

$$\bar{\mu}_i = \alpha_i E_i(X) + (1+\alpha_i)\mu_i, i = 1,...,M \tag{2.23}$$

$$\alpha_i = \frac{n_i(X)}{n_i(X) + r} \tag{2.24}$$

$$n_i(X) = \sum_{j=1}^{N} P(i / x_j) \tag{2.25}$$

$$E_i(X) = \frac{1}{n_i} \sum_{j=1}^{N} P(i / x_j) x_j \tag{2.26}$$

$$P(i / x_j) = \frac{\pi_i p_i(x_j)}{\sum_{k=1}^{M} \pi_k p_k(x_j)} \tag{2.27}$$

Où π_i et $p_i(x)$ sont le poids du mélange et la fonction de densité du $i^{ième}$ mélange (gaussienne), respectivement, et r est un facteur de pertinence contrôlant le degré d'adaptation.

2.6. Conclusion

Dans ce chapitre, nous avons présenté une technique de clustering basée sur un modèle, qui est la maximisation de vraisemblance (EM). Nous l'avons appliquée à l'aide de modèles de mélange gaussien (GMM). Avec le clustering EM, nous avons pu aller plus loin et décrire chaque cluster par son centre de gravité (moyenne), sa covariance (afin que nous puissions avoir des clusters (classes) elliptiques) et son poids (la taille du cluster). La probabilité qu'un point appartienne à un cluster est maintenant donnée par une distribution de probabilité gaussienne multivariée (multivariée - dépendant de plusieurs variables). Cela signifie également que nous avons pu calculer la probabilité qu'un point soit sous une "cloche" gaussienne, c'est-à-dire la probabilité qu'un point appartienne à un cluster.

En résumé, ce chapitre, a été consacré à rappeler qu'est-ce qu'une distribution Gaussienne, puis introduire le Mélange de Gaussiennes et terminer par une explication de l'algorithme Espérance-Maximisation (EM).

Chapitre 3

Résultats Expérimentaux et Discussions

3.1. Introduction

Dans ce chapitre nous présentons les résultats obtenus avec un système d'identification automatique du locuteur, que nous avons élaborés et qui est basé sur la méthode de modélisation du locuteur GMM, décrite dans le chapitre précédent. Dans ce système, la tâche d'identification est dévolue au GMM-UBM. Les protocoles de développement et d'évaluation des différentes expériences réalisés pour l'identification du locuteur, sont décrits dans ce volet. Ils mettent en jeu le module d'extraction de paramètres acoustiques MFCCs ainsi que le milieu dans lequel le système est opérationnel.

3.2. Base de Données

La base de données (BD) que nous avons utilisée dans nos tests, est un sous ensemble extrait d'une BD réelle (TIMIT) [26].

3.3. Protocole expérimental

L'identification du locuteur en mode indépendant du texte, a été évaluée dans cette section sur des séquences de parole de corpus TIMIT décrit précédemment. Ce data-set est constitué de parole de 100 Locuteurs de taille de **2 min** chacun pour la phase d'apprentissage et de 100 Locuteurs de taille de **1 min** chacun pour la phase de test et de 346 locuteurs de taille de **2 min** chacun pour UBM, échantillonnée à **16kHz**. Des vecteurs caractéristiques de 23 coefficients MFCCs sont extraits, en utilisant une fenêtre Hamming de 30 ms avec un chevauchement de 40-50% (10-15 ms). L'adaptation *MAP* a été réalisée, en utilisant 256 gaussiennes pour le modèle UBM, avec un coefficient d'adaptation de r =16 et un nombre d'itérations de 10 dans la phase d'adaptation. Les expériences dans des conditions perturbées sont effectuées en ajoutant synthétiquement un bruit blanc aux séquences de parole de test.

- L'amplitude du segment de bruit varie en fonction de SNR souhaité (0 dB, 5 dB, 10 dB et 15 dB).

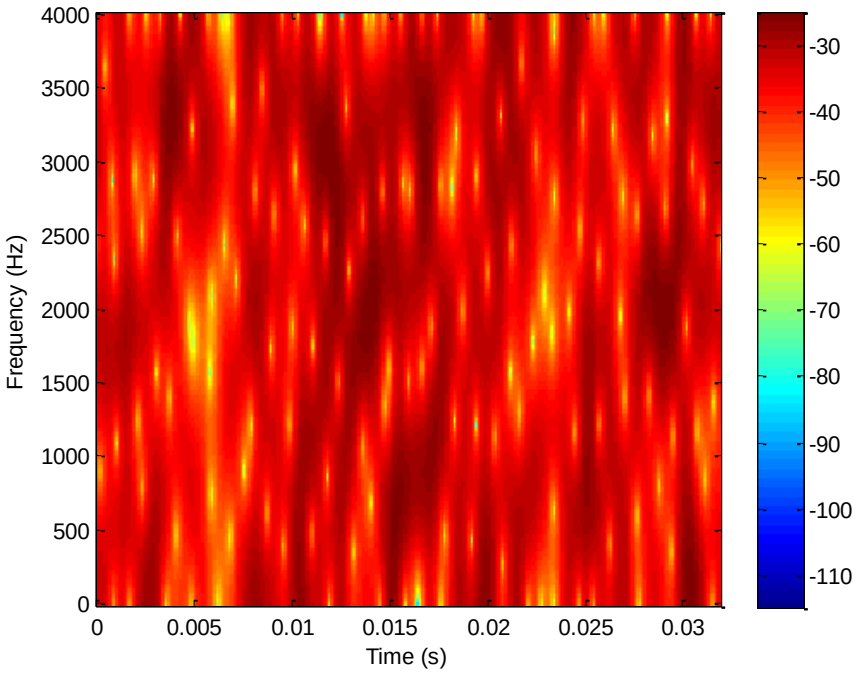

Fig.3.1. Spectrogramme d'un segment de 30 ms extrait d'un bruit Blanc

La métrique d'évaluation utilisée dans ce travail est bien le taux d'identification, qui est définie comme ceci,

$$Taux\ d'identfication\ (\%) = \frac{nombre\ de\ locuteurs\ correctement\ identifiés}{nombre\ total\ de\ locuteurs\ de\ test}$$

3.4. Résultats expérimentaux

3.4.1. L'évolution de rapport de vraisemblance en fonction du nombre de gaussiennes.

Dans cette section, nous étudions les performances de GMM lorsque le nombre de gaussiennes utilisées pour l'identification du locuteur varie, et aussi lorsque MFCCs sont considérés comme vecteurs d'entrées pour le GMM. Les résultats obtenus sont présentés par les figures ci-dessous.

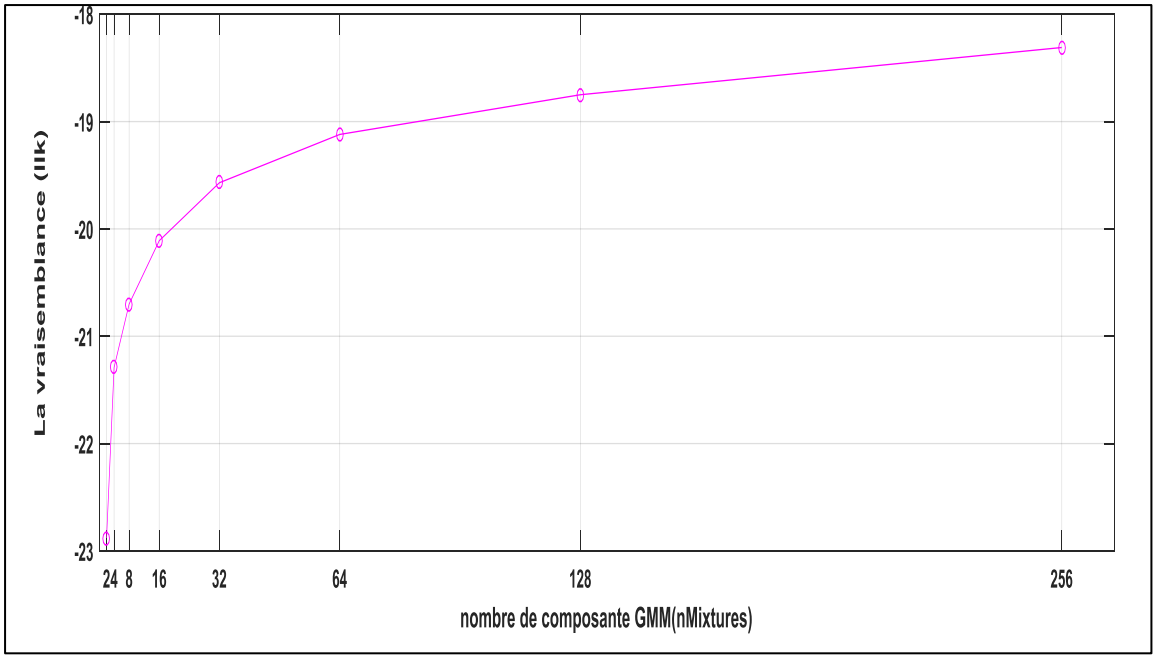

Fig.3.2. L'effet du nombre de gaussiennes sur la qualité de performance d'apprentissage GMM.

❖ **Discussion**

D'après la figure (3.1), nous constatons que la courbe est croissante de 2 à 256, puis après se stabilise, donc la valeur de la vraisemblance augmente en fonction du nombre de gaussiennes et la valeur optimale avec laquelle le graphe est stabilisé est 256 (le meilleur choix).

3.4.2. L'évolution de rapport de vraisemblance en fonction du nombre d'itération pour la gaussienne 256.

Dans cette section, nous étudions les performances de GMM lorsque le nombre d'itération utilisé pour l'identification du locuteur varie, le nombre gaussien reste 256 seulement. Les résultats obtenus sont présentés par la figure ci-dessous.

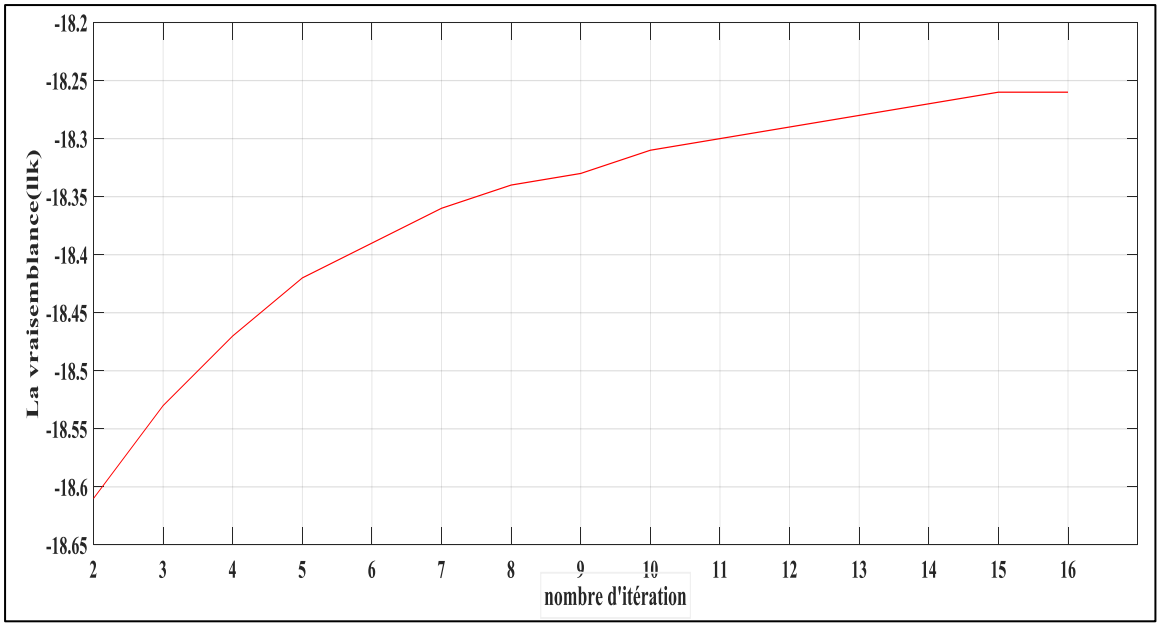

Fig.3.3. L'effet du nombre d'itération sur la qualité de performance d'apprentissage GMM.

❖ **Discussion**

Les résultats montrés par la figure (3.2) indiquent que la courbe de performances est croissante et au bout de la 10ème itération, les valeurs de vraisemblance presque changent lentement. Cela nous mène à dire que, les meilleurs nombres d'itérations pour une bonne estimation (apprentissage) de Maximum de vraisemblance (ML) de GMM sont donc, à partir de 10ème itération au plus.

3.4.3. L'effet du nombre de paramètres MFCCs sur le taux d'identification pour la gaussienne 256.

Dans cette section, nous étudions l'effet du nombre de paramètre MFCCs sur le taux d'identification avec la stabilité du gaussienne (256 gaussienne). Les résultats obtenus sont présentés par le tableau ci-dessous.

Tableau 1. Taux d'identification en fonction du nombre de paramètres MFCC.

Nombre de coefficients MFCCs	08	12	16	20	23
Le taux d'identification	100%	100%	100%	100%	100%

❖ **Discussion**

Les résultats présentés dans le tableau (1) montrent que le nombre de coefficients MFCCs n'a pas une grande influence sur le taux d'identification, car dans le milieu calme (sans bruit) qui est notre cas, l'information caractéristique (extralinguistique) du locuteur est bien présentée par un nombre limitée de MFCCs (exp 08), du coup, le taux d'identification est le même que ce soit pour le nombre MFCCs=08 ou MFCCs=23.

3.4.4. L'évolution de rapport de vraisemblance en fonction du nombre des MFCCs.

Dans cette section, nous étudions les performances d'apprentissage de GMM (256 gaussiennes) lorsque le nombre de coefficients MFCCs est varié. Les résultats obtenus sont présentés par la figure ci-dessous.

Tableau 2. La vraisemblance en fonction du nombre de paramètres MFCC.

Nombre de coefficients MFCCs	08	12	16	20	23
La vraisemblance (llk)	-12.20	-15.13	-17.08	-18.12	-18.31

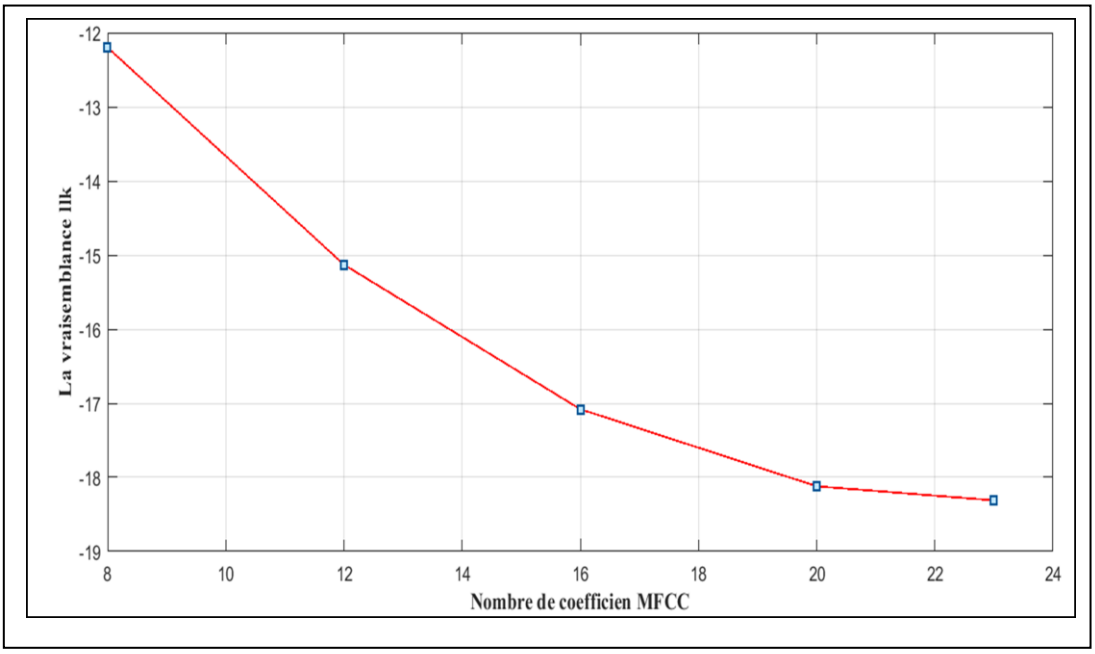

Fig3.4..L'effet de nombre de coefficients MFCCs sur la qualité de performance d'apprentissage

❖ **Discussion**

D'après la figure (3.3), nous constatons que la courbe est décroissante, donc la valeur de la vraisemblance diminue en fonction du nombre de coefficients MFCCs. Cela veut dire que pour une bonne estimation (modélisation ou apprentissage) des données du locuteur par une GMM de 256 gaussiennes dans un milieu calme sans bruit, il suffit seulement présenter le locuteur par des vecteurs caractéristiques MFCCs de taille 08. Donc, le temps de calcul et le degré de complexité sont diminués avec une telle taille.

3.4.5. L'effet de coefficient d'adaptation sur le taux d'identification pour la gaussienne 256.

Dans cette section, nous étudions les performances de GMM, On change la valeur de coefficient d'adaptation [8 : 16] et fixer le nombre gaussienne 256. Les résultats obtenus sont présentés par le tableau ci-dessous

Tableau 3 : le taux d'identification en fonction de coefficient d'adaptation.

Le coefficient d'adaptation	08	09	10	11	12	13	14	15	16
Taux d'identification	100%	100%	100%	100%	100%	100%	100%	100%	100%

❖ **Discussion**

Les résultats obtenus montrent que lorsque nous modifions la valeur de coefficient d'adaptation le taux d'identification n'est pas affecté, car en réalité les données des locuteurs que nous avons utilisées sont pratiquement complètes (pas de manque de données) et en plus l'adaptation est faite en milieu calme sans bruit (donc pas de perte d'information).

3.4.6. Le taux d'identification (%) en fonction de nombre de gaussiennes.

Dans cette section, nous étudions l'influence de nombres de gaussiennes, on prend les nombres 2, 4, 8, 16, 32, 64, 128 et 256 et on fait l'étude de variation du taux d'identification en fonction de nombres de gaussiennes. Les résultats obtenus sont donnés dans le tableau ci-dessous.

Tableau 4 : le taux d'identification en fonction de nombre de gaussien.

Nombre de gaussien	02	04	08	16	32	64	128	256
Taux d'identification	100%	100%	100%	100%	100%	100%	100%	100%

- **Discussion**

A partir de tableau 4, le meilleur taux d'identification est 100% obtenu pour toutes les gaussiennes, alors que dans l'étude de performance d'apprentissage de GMM en fonction du nombre de gaussienne (expérience 3.4.1), il est constaté que la meilleure gaussienne est 256. Ceci est expliqué par le faite que les gaussiennes d'ordre 2,4,8,16,32 et 64 ont provoqués une surestimation (mauvaise estimation) de GMM et comme par hasard le milieu est calme, les résultats d'identification correspondants sont bons. Sinon, dans le cas réel (milieu bruité) les taux d'identification correspondants à ces ordre vont être médiocres.

3.4.7. L'effet du bruit blanc sur le taux d'identification

Dans cette section, nous évaluons les performances de taux d'identification lorsque le SNR du bruit blanc (bruit des télécommunications) varie de 0dB jusqu'à 15dB et le nombre de coefficient MFCC varie de 8 jusqu'à 23. Les résultats obtenus sont présentés par le tableau ci-dessous

Tableau 5 : le taux d'identification en fonction de rapport SNR et de nombre de coefficient MFCC.

Le rapport SNR / **Nombre de coefficient MFCC**	0	5	10	15
8	1%	1%	1%	7%
12	1%	1%	3%	1%
16	2%	2%	3%	17%
20	0%	3%	2%	19%
23	1%	2%	7%	23%

❖ **Discussion**

A partir du tableau 5, nous constatons l'augmentation du taux d'identification en fonction du **SNR**, alors que le meilleur taux d'identification est de 23% obtenue par le bruit blanc (SNR=15dB) avec un nombre de MFCCs=23. Sinon quand SNR=0 et 5 dB, le taux ne donnent pas de bons résultats, car le milieu dans ces deux cas est fortement bruité, donc entraine ce qu'on appelle perte d'information. Dans le cas du milieu bruité qui est le cas réel, nous avons bien montré que le nombre de MFCCs=23, est celui qui donne de meilleurs taux d'indentification pour tous les niveaux de SNR par rapport aux autres nombres et aussi par rapport au cas calme (idéal). Ceci est expliqué par le faite que MFCCs =23, veut dire que le nombre de MFCCs qui sont affectés par le bruit est moins élevés par rapport à d'autres ordres (8,12,16 et 20), du coup, l'information est bien préservée.

3.5. Conclusion

Dans ce chapitre nous avons effectué diverses expériences de l'identification automatique du locuteur en environnement calme et bruité, en utilisant les différentes approches proposées dans le chapitre précédent. Nous avons examiné les différents facteurs affectant les connaissances du locuteur : Nombre de gaussiennes, le taux d'identification, nombre d'itérations…etc.

L'utilisation des coefficients MFCC par le modèle de mélange de gaussienne donne un meilleur taux. En ce qui concerne l'ordre des modèles, l'augmentation du nombre de coefficient d'adaptation n'apporte pas une amélioration sensible du taux d'identification en milieu sans bruit (calme). Nous concluons aussi qu'une dégradation des performances est observée, quand le milieu où notre système d'identification est opérationnel devient bruité.

Conclusion générale

Conclusion générale

Ce travail s'inscrit dans la tâche d'identification automatique du locuteur, notre étude a été focalisée sur le mode indépendant du texte. Pour cela, nous avons simulé et évalué un système d'identification du locuteur, basé sur la méthode de modélisation du locuteur GMM.

Nous avons effectué diverses expériences de l'identification automatique du locuteur en environnement calme et bruité, ce dernier est effectué en ajoutant synthétiquement un bruit blanc aux séquences de parole de test. Différents paramètres ont été testés comme le nombre de paramètres MFCC, le nombre de GMM et le nombre d'itérations. Ces simulations visaient à obtenir le système le plus performant possible.

Nous avons montré que l'utilisation des coefficients MFCC par le modèle de mélange de gaussienne donne un meilleur taux. Et nous avons remarqué que l'augmentation du nombre de coefficient d'adaptation au-delà de 10 n'apporte pas une amélioration sensible du taux d'identification en milieu sans bruit (calme). En revanche, les performances du système sont dégradées, quand le milieu où notre système d'identification est opérationnel devient bruité.

Perspectives :

Comme perspective, nous proposons :

- ✓ Calculer le taux d'identification par une étude comparative basée sur deux méthodes d'extraction de caractéristiques (MFCCs et LPCs) ou bien plus, dans le cas réel ou le milieu est fortement bruité.
- ✓ Proposer un déploiement de notre système conçu, sur une carte FPGA pour qu'il soit portable et commerçable.

Références

Références

[1] Bousquet. P. M, « Bénéfices et limites des représentations en facteur de variabilité totale pour la reconnaissance du locuteur», Spécialité : Informatique, Université d'Avignon et des Pays de Vaucluse, pour obtenir le diplôme de DOCTORAT, (2014).

[2] Toumi. Y, Bouziane. T, 'Utilisation des paramètres i-vecteur pour la reconnaissance des marques de téléphone', Thèse de master, Génie Electrique Filière de télécommunication, Bouira, (2018)

[3] Filipe Velho, ' La reconnaissance du locuteur à l'aide de la transformée en ondelettes continue', Thèse à l'obtention de la maitrise en génie électrique, l'école de technologie supérieure Montréal, (2006)

[4] Alouache. L, Louggani,Y, « Compensation de la variabilité du canal en reconnaissance du locuteur ». Thèse de master, Génie Electrique, Bouira, (2019).

[5] Alouache. L, Louggani,Y, « Compensation de la variabilité du canal en reconnaissance du locuteur », Mémoire diplôme de master système des télécommunications, Télécommunication, Université-Bouira, (2019).

[6] Aziza. Y, Modélisation AR et ARMA de la Parole pour une Vérification Robuste du Locuteur dans un Milieu Bruité en Mode Dépendant du Texte. Mémoire de magister. Université Ferhat Abbas –Sétif (Algérie) (2013).

[7] Hadjer. A et Sabrina .B : "La Reconnaissance Automatique du Locuteur(RAL) par réseaux de neurones et GMM" Thèse d'ingénieur D'Etat, USTHB, (2009).

[8] Ayad. M.Y, Traka. B : "Reconnaissance de Cibles par Radar a écho Doppler" Mémoire de fin d'études pour l'obtention du diplôme d'ingénieur d'état en défense aérienne, (2015).

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

[10] Kim. D.S, Lee. S.Y, Kil. R.M, «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) (1999).

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

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

[13] Asbai. N, '*Reconnaissance automatique de locuteurs en environnement bruité par les méthodes à noyaux* ' Faculté d'électronique et informatique USTHB, Alger, (Magister dissertation) (2010).

[14] Guemmour. S, Bouzidi. S, « Reconnaissance automatique de locuteurs en sciences forensiques (Criminalistiques)», Télécommunication, Université Mohamed el bachir el ibrahimi, Bordj Bou arréridj, (2019).

[15] Fréderic. C, « Théorie et traitement des signaux. presses polytechniques et universitaires romandes », Suisse, (2013).

[16] Douib. O, « reconnaissance automatique de la parole arabe par cmu sphinx 4 » magister option : communication, Électronique, Université Ferhat Abbas –Sétif (Algérie), (2013).

[17] B. S. Atal, "Automatic recognition of speakers from their voice", *Proceeding of the IEEE*, vol, 64(4), pages 460-475, (1976).

[18] D. O'Shaughnessy, "Speaker recognition", *IEEE Transactions Acoustics, Speech, and Signal Processing ASSP*, pages 4–17, (1986).

[19] Reynolds. D and Richard. C, "Robust test –independent speaker identification using Gaussian mixture models", IEEE Transaction on speech and audio processing vol.3, no.1, January, (1995).

[20] Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech communication, 17(1), 91-108.

[21] Bing.X and Berger.T, "Efficient Text-Independent speaker Verification with Structural Gaussian Mixture Models and Neutral Network", IEEE Transaction on speech and audio processing, vol.11, no.5, September, (2003).

[22] DUDA/HART. (1973). Pattern classification and scene analysis. John Wiley.

[23] Li, S. Z., Zhang, D., Ma, C., Shum, H. Y., & Chang, E. (2003, September). Learning to boost GMM based speaker verification. In *INTERSPEECH*.

[24] Duxans, H., & Bonafonte, A. (2003, September). Estimation of GMM in voice conversion including unaligned data. In *INTERSPEECH*.

[25] Reynolds, D.A. (2008). Universal Background Models. Encyclopedia of Biometric Recognition, Springer, Journal Article. Available online: http://www.ll.mit.edu/mission/communications/ist/publications/0802_Reynolds

[26] TIMIT Acoustic-Phonetic Continuous Speech Corpus, speech Recognition LDC93S1. Web Download. the University of Pennsylvania: Linguistic Data Consortium, 1993.