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

ROJET DE FIN D'ETUDES

MASTER SCIENCES ET TECHNIQUES
SYSTÈMES INTELLIGENTS & RÉSEAUX

# LA RECONNAISSANCE AUTOMATIQUE DU LOCUTEUR PAR LA VOIX IP

LIEU DE STAGE : **LABORATOIRE DES SYSTÈMES INTELLIGENTS & APPLICATIONS**

RÉALISÉ PAR : HOUDA KADI

SOUTENU LE : 18/06/2014

ENCADRÉ PAR :

PR. JAMAL KHARROUBI

PR. FATIHA MRABTI

DEVANT LE JURY COMPOSÉ DE :

PR. J.KHARROUBI

PR. S.NAJAH

PR. R.BENABBOU

PR. L.LAMRINI

ANNÉE UNIVERSITAIRE 2013-2014

# Remerciements

Au terme de ce travail,

Après Dieu, je tiens à remercier de nombreuses personnes qui m'ont assuré un appui intellectuel, moral et matériel.

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

**Responsable MST SIR**

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

**A Mme F.MRABTI**

**Professeur à la fst**

Avec mon admiration pour votre rigueur dans le travail vos qualité humaines et professionnelles. Je vous prie de trouver dans ce travail toute la reconnaissance que je vous témoigne.

**A tous les membres de jury**

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

**A Mr Ayoub BOUZIANE**

**Doctorant au laboratoire SIA**

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

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

A mes amis (es) du MASTER systèmes Intelligents et Réseaux

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

# *Résumé*

La reconnaissance automatique de locuteur est le processus qui détermine automatiquement l'identité de la personne qui parle en se basant sur ses caractéristiques vocales. Actuellement, ce type de système est largement utilisé dans plusieurs domaines, essentiellement dans la sécurisation d'accès à des sites web protégés, pour faire passer des transactions bancaires, l'accès aux bases de données,… Le développement remarquable des moyens de communication vocale comme la VoIP a encouragé les chercheurs de migrer vers l'utilisation de cette dernière en reconnaissance automatique de locuteur ce qui définit la motivation principale de ce projet. Dans le cadre de ce travail, nous avons établi une étude bibliographique sur le système RAL par la voix IP, en mettant l'accent sur la tâche de l'identification dans la partie expérimentale.

***Mots clés :*** reconnaissance automatique du locuteur, Identification automatique du locuteur, VOIP, caractéristiques vocales.

# Abstract

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speech's voice to verify their identity and control access to services such as voice dialing, banking by telephone, database access services,…The fast development of technologies like VoIP has encouraged researchers to move to automatic speaker recognition using this technology, which defines the main motivation for this project. In the context of this work, we will establish a bibliographic study on the RAL system with VoIP, focusing on the task of identification in the experimental part.

***Keywords:*** Speaker recognition, VoIP, Speaker identification, speech waves

# Sommaire

**Introduction générale** ..... **10**

**Chapitre 1 : Etat de l'art sur les systèmes de reconnaissance automatique du locuteur** ..... **12**

1.1 Introduction ..... 13

1.2 Terminologie ..... 13

1.2.1 Identification & vérification automatique de locuteur ..... 13

- La Vérification Automatique de Locuteur (VAL) ..... 13
- L'Identification Automatique de Locuteur ( IAL) ..... 14

1.2.2 Dépendance et indépendance du texte ..... 15

1.2.3 Les variabilités du signal de la parole ..... 15

1.3 Fonctionnement d'un système de reconnaissance automatique du locuteur ..... 16

1.3.1 Paramétrisation ..... 16

- Paramètres de l'analyse spectrale ..... 16
- Paramètres prosodiques ..... 17
- Paramètres dynamiques ..... 17

1.3.2 Traitement post paramétrisation ..... 18

- La méthode RASTA ..... 18
- Feature warping ..... 19
- CMVN (Normalisation moyenne et variance des paramètres cepstraux) ..... 20
- Le feature mapping ..... 21
- VTLN (Vocal Tract Length Normalization) ..... 21
- HLDA (HETEROSCEDASTIC LINEAR DISCRIMINANT ANALYSIS) ..... 22
  - LDA (Analyse discriminante linéaire) ..... 23
  - HLDA (HeteroscedasticLinear Discriminant Analysis) ..... 24
- Speech enhancement ..... 26
  - La soustraction spectrale ..... 27
  - Le filtre de Wiener ..... 28
  - Filtrage de Kalman ..... 28

1.3.3 Modélisation ..... 29

- Approche vectorielle ..... 29
  - La programmation dynamique ..... 29
  - La quantification vectorielle ..... 29
- Approche statistique ..... 30
  - Méthodes statistiques du second ordre ..... 30

- Mélange de gaussiennes ..... 30
- Modèles de Markov cachés ..... 31

➢ Approche prédictive ..... 31

➢ Approche connexionniste ..... 32

1.3.4 Décision et mesures de performances ..... 32

➢ Identification automatique de locuteur ..... 32

➢ Vérification automatique de locuteur ..... 33

1.4 Evolution d'un système de reconnaissance automatique de locuteur ..... 33

1.5 Les approches classiques utilisées dans le système de reconnaissance automatique de locuteur ..... 34

1.6 Domaines d'application ..... 35

Conclusion ..... 36

**Chapitre 2 : La reconnaissance automatique du locuteur par la voix IP ..... 37**

2.1 Introduction ..... 38

2.2 La théorie de la voix sur IP ..... 38

2.2.1 Architecture de la transmission de la voix IP ..... 39

2.2.2 Protocoles de la voix ..... 41

➢ Le protocole H.323 ..... 41

- Présentation générale ..... 41
- Les limites du protocole ..... 42

➢ Le protocole SIP ..... 43

- Présentation générale ..... 43
- Fonctionnement ..... 43
- Avantages et inconvénients ..... 47

2.2.3 Points forts et limites de la voix sur IP ..... 48

2.3 Evolution de la reconnaissance automatique de locuteur par la voix IP ..... 49

2.4 Connexions internationales ..... 53

Conclusion ..... 54

**Chapitre 3 : Expériences et résultats ..... 55**

3.1 Introduction ..... 56

3.2 La paramétrisation ..... 56

3.2.1 MFCC_FB20 ..... 57

3.2.2 DavisSkowronski_MFCC_FB20 ..... 58

3.2.3 HTK_MFCC_FB24 ..... 58

3.2.4 HTK_MFCC_FB26 ..... 59

3.2.5 MFCC_FB40 ..... 60

3.2.6 HFCC_E_FB29 ..... 61
3.3 Le protocole expérimental ..... 63
3.3.1 Description de la base de données ..... 63
3.3.2 Décomposition parole/non parole ..... 63
3.3.3 La phase de la paramétrisation ..... 64
3.3.4 Apprentissage par GMM ..... 64
3.3.5 La phase de la décision ..... 64
3.4 Résultats et tests ..... 65
3.4.1 Identification du locuteur dans un milieu fermé ..... 65
3.4.2 Discussion des résultats obtenus ..... 65
3.4.3 Identification de locuteur en milieu ouvert ..... 67
3.5 Implémentation de l'interface graphique ..... 67
**Conclusion et perspectives ..... 71**
Bibliographie ..... 73
Webographie ..... 75

# Liste des figures

**Figure 1.1 : La vérification automatique de locuteur ..... 13**
**Figure 1.2 : L'identification automatique de locuteur en un groupe fermé ..... 14**
**Figure 1.3: L'identification automatique de locuteur en un groupe ouvert ..... 15**
**Figure 1.4: Extraction des paramètres avec MFCC ..... 17**
**Figure 1.5: Chaine de calcul de coefficient RASTA-PLP ..... 19**
**Figure 1.6: Les coefficients Cepstra avec la méthode RASTA-PLP ..... 19**
**Figure 1.7: Alignement des vecteurs acoustiques ..... 20**
**Figure 1.8: Feature mapping ..... 21**
**Figure 1.9: Limitation de LDA ..... 24**
**Figure 2.1: La Voix sur IP entre deux ordinateurs ..... 38**
**Figure 2.2: La voix sur IP entre un ordinateur et un téléphone ..... 39**
**Figure 2.3: La voix sur IP entre deux téléphones ..... 39**
**Figure 2.4: Architecture de la transmission de la voix IP ..... 39**
**Figure 2.5 : Pile de protocole H323 ..... 41**
**Figure 2.6: exemple d'établissement d'appel entre deux agents ..... 45**
**Figure 2.7 : Exemple d'enregistrement SIP ..... 45**
**Figure 2.8: Principe de proxy SIP ..... 46**
**Figure 2.9 : Session SIP à travers un proxy ..... 47**
**Figure 3.1 : banc de filtre de la variante MFCC_FB20 ..... 58**
**Figure 3.2: Le banc de filtre de la variante HTK_FB26 ..... 59**
**Figure 3.3 : Banc de 32 filtres ..... 61**
**Figure 3.4 : Comparaison entre le banc de filtres des MFCCs et des HFCCs ..... 61**
**Figure 3.5: Processus d'identification ..... 64**
**Figure 3.6 : Taux d'identification en utilisant la base de données basé sur la VOIP ..... 65**
**Figure 3.7: Taux d'identification avec une base de données de la voix enregistré par le microphone ..... 67**
**Figure3.8 : Interface d'accueil ..... 68**
**Figure 3.9 : identification dans un groupe fermé ..... 69**
**Figure 3.10: Identification dans un groupe ouvert ..... 70**

# Introduction générale

Depuis plusieurs années, la reconnaissance automatique du locuteur (i.e Reconnaissance par la voix) fait l'objet de travaux de recherches entrepris par de nombreuses équipes de recherches dans le monde, elle s'est limité longtemps à la détection ou la vérification de l'identité d'une personne à partir d'un échantillon de sa voix, la vérification consiste à accepter ou refuser l'identité proclamée par un locuteur, en se basant sur un modèle qui lui est associé. L'identification consiste en la reconnaissance d'un locuteur particulier parmi un ensemble fini de locuteurs possibles. Aussi bien la vérification, que l'identification du locuteur se fait en calculant un modèle stochastique sur la base de l'expression vocale du locuteur à reconnaitre. Une fois calculé, ce modèle est comparé à des modèles pré entraînés sur la base de différents enregistrements prononcés par les locuteurs.

Depuis quelques années le champ d'application des techniques de reconnaissance automatique de locuteur, s'est considérablement élargi suite au progrès à la fois des algorithmes utilisés, la puissance de traitement disponible, et l'évolution remarquable des technologies utilisées.

La tendance Technologique actuelle montre une évolution vers l'exécution de diverses transactions en utilisant la téléphonie sur IP qui n'était examiné avant comme une perspective de la technologie dans les travaux de recherches. L'évolution considérable de cette technique est due d'une part à l'évolution des infrastructures réseaux, d'autre part l'important débit fournit par cette technologie grâce à la nécessité d'un temps de latence très court.

Ce projet a pour but de faire une étude bibliographique sur le fonctionnement du système de reconnaissance automatique de locuteur et les différentes approches et méthodes qui en découlent, en passant à la présentation des différents travaux réalisés dans le cadre de la reconnaissance de locuteur par la voix IP, et terminant par une partie expérimentale qui porte essentiellement sur l'identification automatique de locuteur par la voix IP en se basant sur notre propre base de données construite au sein du laboratoire SIA.

Ce document est organisé comme suit : dans le premier chapitre, nous présentons l'état de l'art d'un système de reconnaissance automatique du locuteur, suivi d'une présentation de la théorie de la VOIP, et l'exposition des différents travaux réalisés dans le cadre de la

reconnaissance automatique du locuteur par la voix IP, ainsi que les laboratoires qui opèrent dans ce domaine dans le chapitre 2. Dans le chapitre 3 nous présentons la description du protocole expérimentale que nous avons utilisé dans les expériences sur les différentes variantes de la paramétrisation MFCC dans le cadre de l'identification de locuteur en mode indépendant du texte et les résultats obtenus, et terminant finalement par les conclusions et perspectives.

# Chapitre 1 : Etat de l'art sur les systèmes de reconnaissance automatique du locuteur

# 1.1 Introduction

La reconnaissance automatique du locuteur est une des branches de l'authentification biométrique, qui se réfère à la reconnaissance automatique de l'identité des personnes en utilisant certaines de leurs caractéristiques intrinsèques. Outre la voix, il y a beaucoup d'autres modèles physiques et comportementaux pour l'authentification biométrique, par exemple : l'iris, les réseaux veineux de la rétine, les réseaux veineux de la paume de la main, l'empreinte digitale, …etc. Pratiquement, la sélection d'un modèle biométrique adéquat devrait prendre en compte au moins les considérations suivantes : la robustesse, la précision, l'accessibilité et l'acceptabilité. Par rapport à ces critères de sélection, parmi toutes les technologies d'authentification biométriques, la reconnaissance du locuteur est probablement la plus naturelle et économique pour les systèmes de communication homme-machine parce que d'une part la collecte de données parole est beaucoup plus pratique que les autres motifs, et d'autre part, la parole est le mode dominant d'échange d'information pour les êtres humains et tend à être le mode dominant pour l'échange d'information pour les systèmes de communication homme-machine.

# 1.2 Terminologie

### *1.2.1 Identification & vérification automatique de locuteur*

La reconnaissance automatique de locuteur consiste à obtenir des renseignements concernant l'identité d'une personne à partir d'un enregistrement de sa voix. Pour qualifier précisément les différentes tâches entrant dans le cadre d'un système de reconnaissance automatique de locuteur, on distingue entre deux tâches différentes

- La Vérification Automatique de Locuteur (VAL)

Lorsqu'on cherche à décider si l'identité revendiquée par un locuteur est compatible avec sa voix. Dans ce type d'applications, il s'agit donc de trancher entre deux hypothèses, soit le locuteur est bien le locuteur autorisé, c'est à dire celui dont l'identité est revendiquée, soit nous avons affaire à un imposteur qui cherche à se faire passer pour un locuteur autorisé (*fig 1.1*).

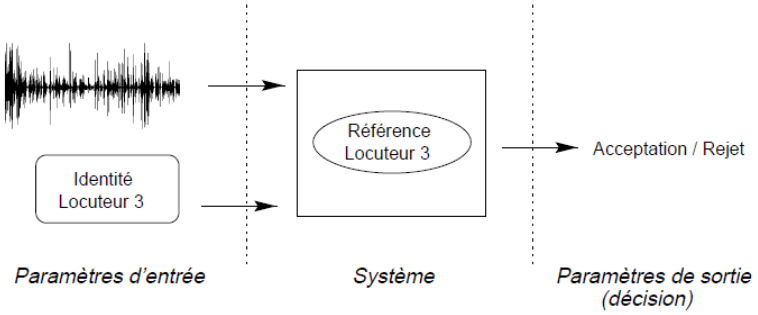


*Figure 1.1 : La vérification automatique de locuteur*

➢ L'Identification Automatique de Locuteur ( IAL)

Il s'agit de déterminer, parmi un ensemble de N locuteurs potentiels, à quel locuteur correspond un enregistrement vocal. En identification, la réponse apportée n'est pas de type binaire (acceptation ou rejet) comme pour la vérification, puisqu'il est nécessaire de distinguer un locuteur parmi un groupe. On distingue encore deux sous problèmes d'identification selon que l'on est sûr ou non du fait que l'enregistrement provient bien d'un des membres du groupe de locuteurs :

- Si l'on a affaire à **un ensemble fermé** (*closed set*) (*fig.1.2*) : le système IAL décide de l'identité la plus probable parmi les utilisateurs connus (dont il possède une référence). Ce mode de fonctionnement tend à considérer que seules des personnes référencées peuvent accéder au système. Un tel système ne doit alors être utilisé que dans un environnement au sein duquel tous les individus sont connus.

  L'identité $I_Y$ retournée, correspondant à la référence Y est obtenue par :

  **Y = argmax f(X|S).**

  f(X|S) est le score calculé lors de la comparaison des données S au modèle de référence de l'individu $I_X$.

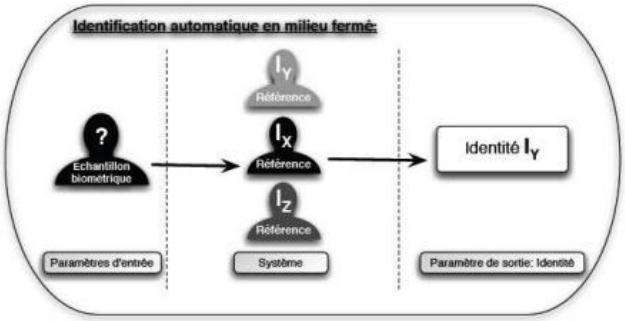


Figure 1.2 : L'identification automatique de locuteur *en un groupe fermé*

- Dans le cas d'**un ensemble ouvert** (*Open set*) (*fig.1.3*): le système IAL a la possibilité de rejeter le locuteur dont il teste les données si elles ne correspondent à aucune des identités répertoriées. Ce locuteur est alors considéré comme inconnu du système. Pour ce faire, les données S sont comparées à chaque référence X connue par le système. Chaque comparaison fournit un score f(X|S). Le score le plus élevé est alors comparé à un seuil fixé préalablement.

  Si le score est supérieur à ce seuil, le système décide qu'il s'agit de la personne correspondant à la référence sélectionnée. Si le score est inférieur à ce seuil, le système décide qu'il ne s'agit pas d'une personne «connue».

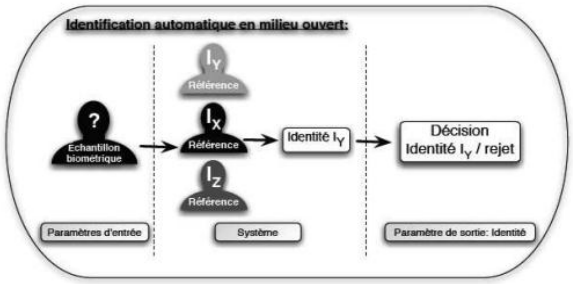


*Figure1.3: L'identification automatique de locuteur en un groupe ouvert*

Pour résumer, en identification en milieu ouvert, le système répond à deux interrogations: « *Quelle est l'identité la plus probable ?*» et « *Les données biométriques analysées correspondent-elles à cette identité ?*», alors qu'en milieu fermé il ne répond qu'à la première. L'identité la plus probable est obtenue, comme dans le cas de l'identification en milieu fermé, par **:** Y = argmax f(X|S). Après, cette probabilité maximale est comparé à un seuil Ω fixé au préalable, si la probabilité obtenue est supérieur à ce seuil, alors le locuteur est bien identifié, sinon, le système le déclare comme une identité non reconnue.

### *1.2.2 Dépendance et indépendance du texte*

On parle de reconnaissance de locuteur en mode dépendant du texte (Text-dependant ou fixed-text) lorsque le texte prononcé par le locuteur est fixé et connu à l'avance. Par contre, si le texte prononcé n'est pas connu à priori, on parle du mode indépendant du texte (text-independant ou free-text). La distinction entre ces deux modes de fonctionnement des applications de reconnaissance du locuteur est très importante car les techniques utilisées ainsi que les performances obtenues, dans les deux cas sont très différentes.

### *1.2.3 Les variabilités du signal de la parole*

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

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

Pour la reconnaissance de locuteur, on cherche à extraire des caractéristiques du signal de parole qui présente une forte variabilité interlocuteur (pour pouvoir différencier les locuteurs entre eux) et une faible variabilité intra-locuteur (pour garantir la robustesse du système).

La variabilité **intersession** (entre sessions d'enregistrements) fait apparaître l'influence de facteurs extérieurs sur le signal de parole. A la sortie du conduit vocal humain, l'onde de parole est considérée comme idéale, car aucune déformation/distorsion de l'environnement extérieur ne l'a modifiée. L'environnement sonore lors de l'enregistrement, le matériel d'acquisition ou le canal de transmission utilisé vont ensuite déformer l'onde sonore originelle. Le canal de transmission, par exemple, agit comme un filtre en fréquence sur l'onde sonore.

Ces facteurs rendent complexe la comparaison entre plusieurs échantillons d'un même

individu. De nombreux travaux expérimentaux ont montré que des variations de matériel entre les phases d'apprentissage et de test sont à l'origine de graves dégradations des performances. Par exemple, l'acquisition d'un signal de parole sur le réseau GSM [10] introduit les dégradations suivantes sur le signal de parole :

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

## 1.3 Fonctionnement d'un système de reconnaissance automatique du locuteur

La réalisation d'un système de reconnaissance automatique de locuteur passe par trois phases : La paramétrisation, la modélisation, et la décision.

### *1.3.1 Paramétrisation*

La variation de la nature du signal acoustique rend le traitement des données brutes issues de ce dernier très difficile. En effet, ces données contiennent des informations complexes, souvent redondantes et mélangées.

La phase de paramétrisation, qui traite le signal acoustique reçu, doit remplir plusieurs objectifs :

- Séparer le signal du bruit ;
- Extraire l'information utile à la reconnaissance;
- Convertir les données brutes à un format directement exploitable par le système.

Afin de concevoir un bon système RAL, il faut choisir des paramètres qui sont fréquents, (ne pas correspondre à des évènements ne survenant que très rarement dans le signal), facilement mesurables, robuste face aux imitateurs, ne pas être affecté par le bruit ambiant ou par les variations due au canal de transmission.

Pratiquement, il est très difficile de réunir tous ces éléments en même temps, la sélection des paramètres pose un problème très complexe, et influe fortement sur les résultats des systèmes RAL. D'après plusieurs recherches effectuées sur cette étape, les types de paramètres efficaces et utilisables sont les paramètres de l'analyse spectrale, les paramètres prosodiques, et les paramètres dynamiques.

- <u>Paramètres de l'analyse spectrale</u>

Les principaux paramètres de l'analyse spectrale utilisés en RAL sont les coefficients de prédiction linéaire et leurs différentes transformations (LPC, LPCC,..) ; ainsi que les

coefficients issus de l'analyse en banc de filtres et leurs différentes transformations (coefficients banc de filtres, MFCC...).

Plusieurs travaux ont été publiés pour comparer les différentes techniques en paramétrisation, l'enjeu de ces travaux était de cibler les meilleurs paramètres représentant de façon efficace les propriétés caractéristiques propres à chaque locuteur. Les meilleurs résultats ont été obtenus en utilisant la méthode MFCC.

Le codage MFCC est basé sur les variations des bandes critiques de l'oreille humaine avec la fréquence, les filtres sont espacés linéairement aux basses fréquences et logarithmiquement à hautes fréquences, ces filtres sont modélisés par une échelle non linéaire issue de connaissances sur la perception humaine : l'échelle de Mel. Pour les MFCC on utilise la fenêtre de Hamming durant la transformation du domaine temporel au domaine fréquentiel. Cette transformation est faite en utilisant la transformée de Fourier. Un filtrage est appliqué ensuite, par banc de filtres triangulaires espacés selon l'échelle de Mel. Cette échelle reproduit la séléctivité de l'oreille qui diminue avec l'accroissement de la fréquence. Après le calcul de log, une transformée en cosinus discrète est appliquée pour assurer un retour au domaine temporel (*figure 1.4*)

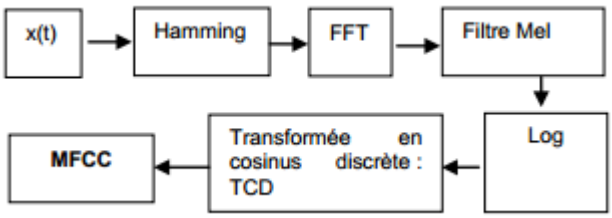


*Figure 1.4: Extraction des paramètres avec MFCC*

- Paramètres prosodiques

Le terme "paramètres prosodiques" réunit l'énergie, la durée et la fréquence fondamentale (ou pitch). Ces paramètres caractérisent en grande partie le style d'élocution d'un locuteur. L'énergie contient l'information liée au niveau acoustique moyen du signal. Ces paramètres s'avèrent fragiles en pratique et ne permettent pas, à eux seuls, de discriminer de manière fiable les locuteurs. En conséquent, ils sont souvent associés aux paramètres de l'analyse spectrale.

- Paramètres dynamiques

La prise en compte d'une information de type dynamique peut être un facteur d'amélioration des performances d'identification du locuteur.

Une première approche, employée pour utiliser cette information au niveau des paramètres, consiste à utiliser une concaténation de plusieurs trames successives de parole (méthodes prédictives). Cependant, cette approche nécessite plus de paramètres dans les modèles et est conduit à des problèmes d'estimation des modèles lors de l'apprentissage.

La seconde possibilité consiste à calculer les dérivées du premier et du second ordre appelé aussi coefficient de (Δ) ou (ΔΔ) qui sont désormais très répandue en raison de leur simplicité de mise en œuvre.

### *1.3.2 Traitement post paramétrisation*

Les paramètres MFCC sont sensibles au changement du canal ou d'environnement, pour cela il est nécessaire d'appliquer autres traitement pour remédier à ce changement, et pour réduire le bruit, dans cette partie on va présenter une étude sur les méthodes les plus connus pour ce traitement,

- La méthode RASTA

La méthode Rasta est une méthode a été proposé en 1994 [15], elle a pour but de supprimer les composantes spectrales dont l'évolution temporelle est plus rapide ou lente que celle du conduit vocal humain.

La méthode RASTA est intégrée à une analyse PLP. En effet, après avoir effectué la transformée de Fourier discrète à court terme, on calcule le spectre d'amplitude en bandes critiques. On applique le logarithme pour récupérer l'enveloppe spectrale du signal comme pour une analyse cepstrale. On effectue ensuite un filtre passe bande qui a pour conséquence de supprimer les composantes constantes ou lentes du signal. On réalise après une compression de l'amplitude par l'application d'une racine cubique. Enfin, on calcule les coefficients selon la méthode LPC classique (*figure 1.5*).

Signal sonore

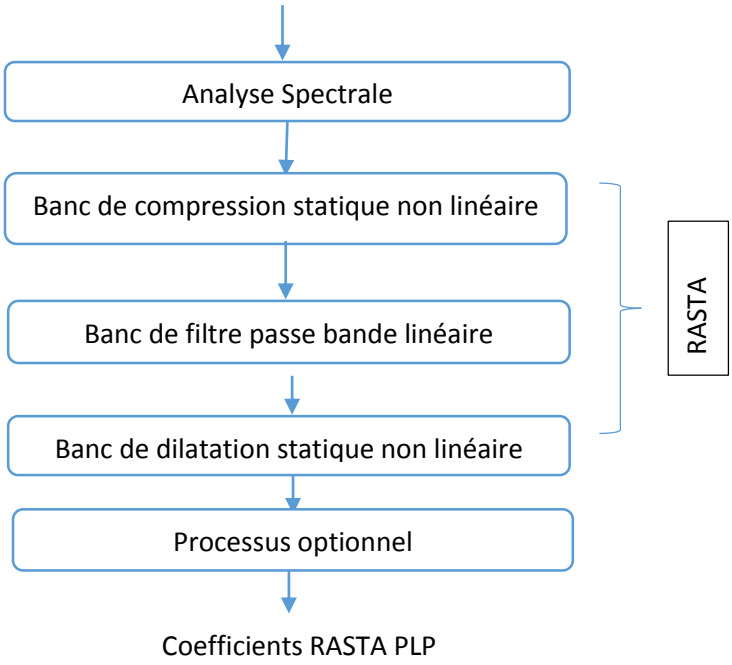


*Figure 1.5: Chaine de calcul de coefficient RASTA-PLP*

Un exemple de calcul de quelques paramètres d'un signal de parole, utilisant cette technique est illustré par la figure 1.6 :

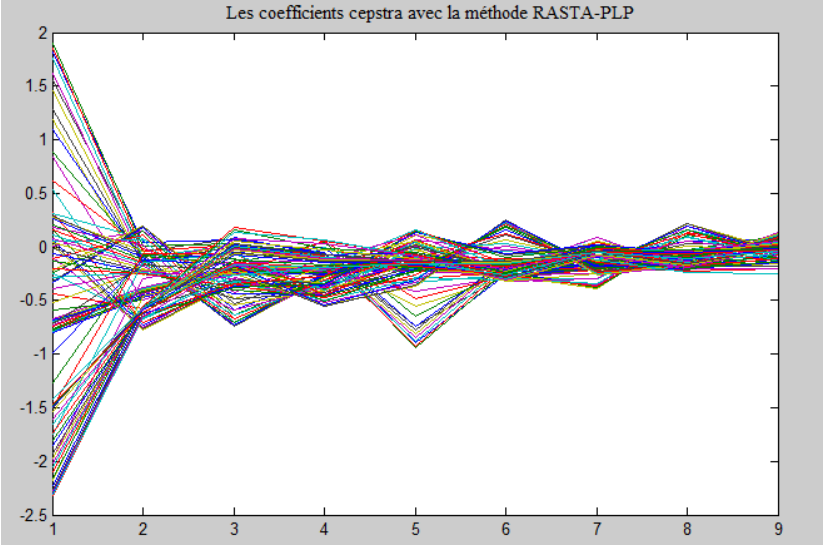


*Figure 1.6: Les coefficients Cepstra avec la méthode RASTA-PLP*

Plusieurs études réalisées pour mesurer la performance de cette technique, ont permis de confirmer ses bonnes qualités relativement aux distorsions et ses moindres qualités face aux bruits additifs, signe de la présence de plusieurs sources sonores dans un même environnement.

- Feature warping

Le feature warping est une technique, apparue en 2001 [20], qui a pour but de conditionner la distribution de chaque trajectoire cepstrale en alignant les coefficients cepstraux

d'une telle façon que sa distribution soit égale à une distribution donnée. Pour un coefficient cepstral $C_m(t)$, on cherche donc $\hat{C}_m$ satisfaisant : $\int_{-\infty}^{Cm(t)} h_m(x)\,dx = \int_{-\infty}^{\hat{C}m\,(t)} f(y)\,dy$ Où $h_m(x)$ est la distribution de la $m^{ème}$ trajectoire cepstrale et f(y) est la distribution cible. Généralement on prend comme distribution cible la distribution de la loi normale. Dans ce cas, cette méthode s'appelle aussi gaussianisation. Un schéma explicatif est représenté dans la figure 1.7:

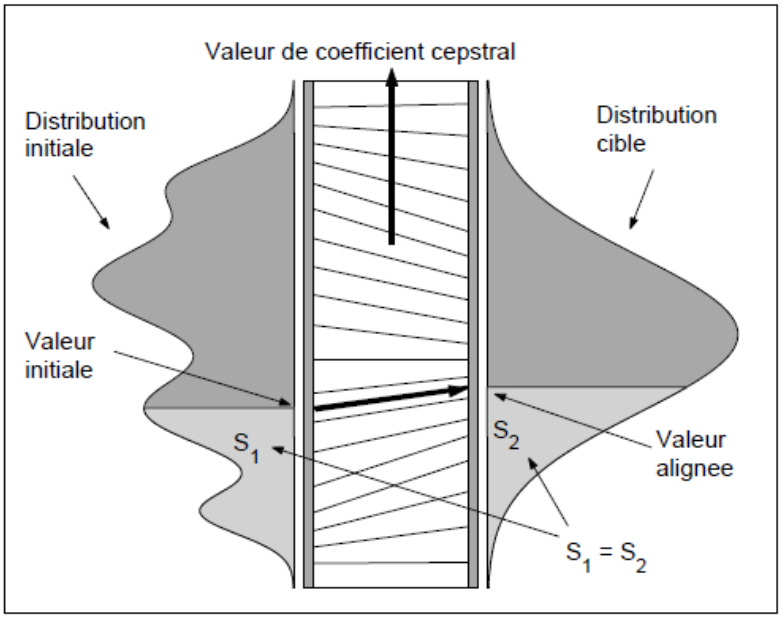


*Figure 1.7: Alignement des vecteurs acoustiques*

La mise en œuvre de cette méthode passe par ces quatre étapes :

- Pour chaque coefficient $C_m(t)$ fenêtrer sa trajectoire par une fenêtre de taille L de telle façon que ce coefficient soit au milieu de la fenêtre. Il est conseillé de prendre L= 3 sec.
- Ranger les coefficients obtenus dans l'ordre croissant de ces valeurs.
- Calculer le rang de chaque coefficient comme sa position dans ce rangement (Le plus petit coefficient obtient le rang 1 bien que le plus grand L). Cela revient à calculer la fonction de répartition de la distribution locale de la trajectoire cepstrale.
- Soit R, le rang de $C_m(t)$, calculer la nouvelle valeur $\hat{C}m\,(t)$ comme :

  $$\hat{C}m\,(t) = F^{-1}\left(\frac{R-1/2}{L}\right)$$

  Où F est la fonction de répartition de la distribution cible

La technique feature warping peut être combinée avec d'autres techniques pour améliorer la robustesse d'un système de reconnaissance sous certaines conditions.

- <u>CMVN (Normalisation moyenne et variance des paramètres cepstraux)</u>

La normalisation moyenne et variance des paramètres cepstraux est une technique très simple et très répondue en reconnaissance automatique de locuteur. Elle consiste à retirer

la moyenne de la distribution de chacun des paramètres cepstraux (la composante continue), et à ramener la variance à une variance unitaire en les divisant par l'écart type global des paramètres acoustiques [26]. Quand seule la moyenne est normalisée, on parle alors de la Cepstral Mean Subtraction (CMS) ou *Cepstral Mean Normalization* (CMN)

- Le feature mapping

La CMVN, le filtrage RASTA, et le feature warping sont toutes des techniques non supervisées qui n'utilisent aucune connaissance sur le canal, par contre le feature mapping, proposé par Reynolds en 2003 [22], est une technique supervisé de normalisation qui projette les paramètres acoustiques liées à une condition d'enregistrement donnée dans un nouvel espace de caractéristiques indépendant du canal. Cette transformation permet de réduire les effets de la variabilité du canal.

La technique de feature mapping est inspirée de l'approche Speaker Model Synthesis [24], elle se base sur la projection des caractéristiques des différents canaux, en un seul espace de caractéristiques indépendant du canal.

La figure 1.8 montre la structure de la technique de feature mapping, comme en SMS défini dans [24], le FM apprend un modèle GMM pour chaque canal connue par adaptation MAP d'un modèle indépendant du canal. Le modèle dépendant du canal modélise en réalité un sous espace de l'espace acoustique global. Il en est déduit une transformation représentant la relation entre le modèle indépendant du canal et le modèle dépendant du canal. Lors de la phase de l'apprentissage des modèles, on cherche en premier lieu à identifier le canal le plus vraisemblable pour l'enregistrement traité, et on applique par la suite la transformation qui lui est associé sur les vecteurs paramétriques, cette transformation permet de réduire les effets de la variabilité du canal, entre conditions de test et d'apprentissage.

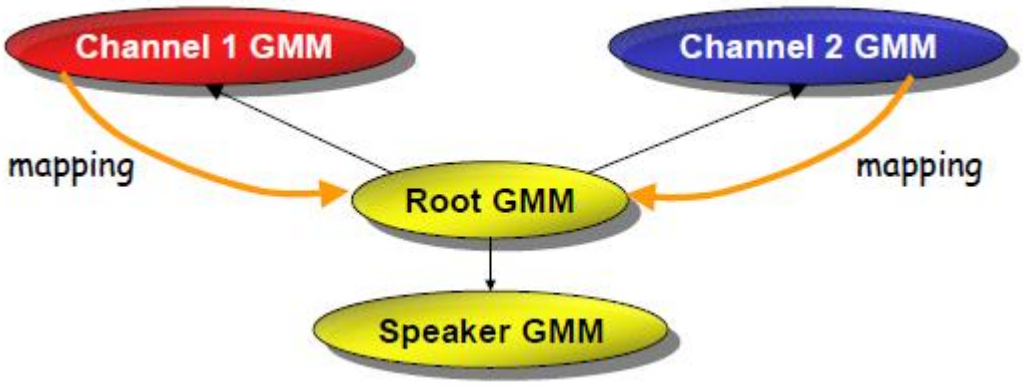


*Figure 1.8: Feature mapping*

Cette nouvelle approche a montré de bonnes performances au niveau de la réduction des effets du canal, elle peut être appliquée pour d'autres applications de reconnaissance de parole.

- VTLN (Vocal Tract Length Normalization)

La variabilité interlocuteur est l'un des problèmes majeurs dans un système de reconnaissance de locuteur, elle influence les performances globales du système à cause de la variabilité des locuteurs, de leurs accents, leurs styles de vocabulaires.

La normalisation de la longueur de conduit vocal (VTLN) est une technique de normalisation de locuteur intervenant au niveau des paramètres acoustiques. Elle est largement répandue dans les systèmes de reconnaissance à grands vocabulaires [28]. Cette normalisation repose sur une modification linéaire de l'échelle des fréquences afin de compenser les différences de longueur de conduit vocal entre les locuteurs.

En général, deux questions sont posées dans VTLN [23] : Premièrement, à partir d'un signal de parole donné, comment obtenir le facteur de normalisation ? La deuxième question est comment fait la normalisation en sachant le facteur de normalisation ?

Le facteur de normalisation (warping factor) reflète la différence du conduit vocal entre plusieurs locuteurs, il existe plusieurs techniques pour calculer l'estimer, en générale il est sélectionné parmi un ensemble de valeurs de 0,8 à 1,25.

Pour la deuxième question, l'utilisation de l'échelle de bark/mel [23]

La majorité des articles publiés sur la normalisation de conduit vocale s'intéresse sur l'une des sujets suivants :

- Type de l'échelle de fréquence (linéaire ou non linéaire), et son implémentation sur quel domaine (Domaine temporel, domaine fréquentiel, domaine cepstral).
- Estimation de facteur de normalisation (warpingfactors) pour l'apprentissage.
- Estimation d'un facteur de normalisation efficace pour le test
- Gain en taux de reconnaissance en utilisant la méthode VTLN et dans plusieurs situations (Environnement bruité ou propre, large corpus d'apprentissage ou limité, un grand vocabulaire ou petit).
- La comparaison de la technique VTLN avec des techniques d'adaptation

- HLDA (HETEROSCEDASTIC LINEAR DISCRIMINANT ANALYSIS)

De façon générale, l'objectif de la transformation discriminante des observations consiste à trouver un espace de projection de faible dimension mais conservant l'information discriminante, pour cela, l'approche de LDA et sa généralisation HLDA sont les plus utilisés.

De point de vue mathématique, on peut exprimer cela en appliquant la transformation linéaire suivante :

$$y = \theta^T x$$

Avec y représente l'ensemble des vecteurs du nouveau espace, x représente l'espace de vecteurs d'entrée, et $\theta$ est la matrice de transformation, de dimension $n \times p$.

- *LDA (Analyse discriminante linéaire)*

C'est une méthode utilisé pour les statistiques, la reconnaissance de formes, et l'apprentissage automatique pour trouver une combinaison linéaire qui caractérise ou sépare deux ou plusieurs classes d'objets ou d'événement, la combinaison résultante peut être utilisée comme un classificateur linéaire, ou plus couramment la réduction de la dimensionnalité plus tard avant la classification [17].La LDA consiste tout d'abord à calculer les vecteurs moyens et les matrices de covariances pour chaque classe j :

$$\mu_j = \frac{1}{N_j}\sum_{i\epsilon l^{-1}(j)} x_i \text{ et } \sum_j = \frac{1}{N}\sum_{i\epsilon l^{-1}(j)}(x_i - \mu_j)(x_j - \mu_j)^T$$

Où :

- $N_j$ le nombre de vecteurs dans la classe j, $\sum_{j=1}^{J} N_j = N$

Ensuite on calcule les deux matrices de covariance : inter-classe $\sum w$et intra-classe $\sum B$

$$\sum_w = \frac{1}{N}\sum_{j=1}^{J} N_j \sum_j \qquad \text{et } \sum_B = \frac{1}{N}\sum_{j=1}^{J} N_j(\mu_j - \mu\ )(\mu_j - \mu\ )^T$$

Enfin on cherche la transformation $\hat{\theta}$ satisfaisant le critère suivant :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{|\theta^T \Sigma_B \theta|}{|\theta^T \Sigma_W \theta|}$$

Le but de ce critère est de maximiser la variabilité entre les classes tout en minimisant la variabilité entre les classes.

Le problème de LDA, c'est qu'elle est essaie de séparer la moyennes des classes sans prendre en compte l'information discriminante présente dans la différence de matrice de covariance, elle est incapable de traiter les données dans le cas hétéroscédastique, càd le cas dans lequel les classes n'ont pas de matrices de covariance égale, cette limitation devient très évidente dans le cas de deux classes avec deux matrice de covariance différente (figure 1.9). D'où l'apparition de la méthode HLDA [17].

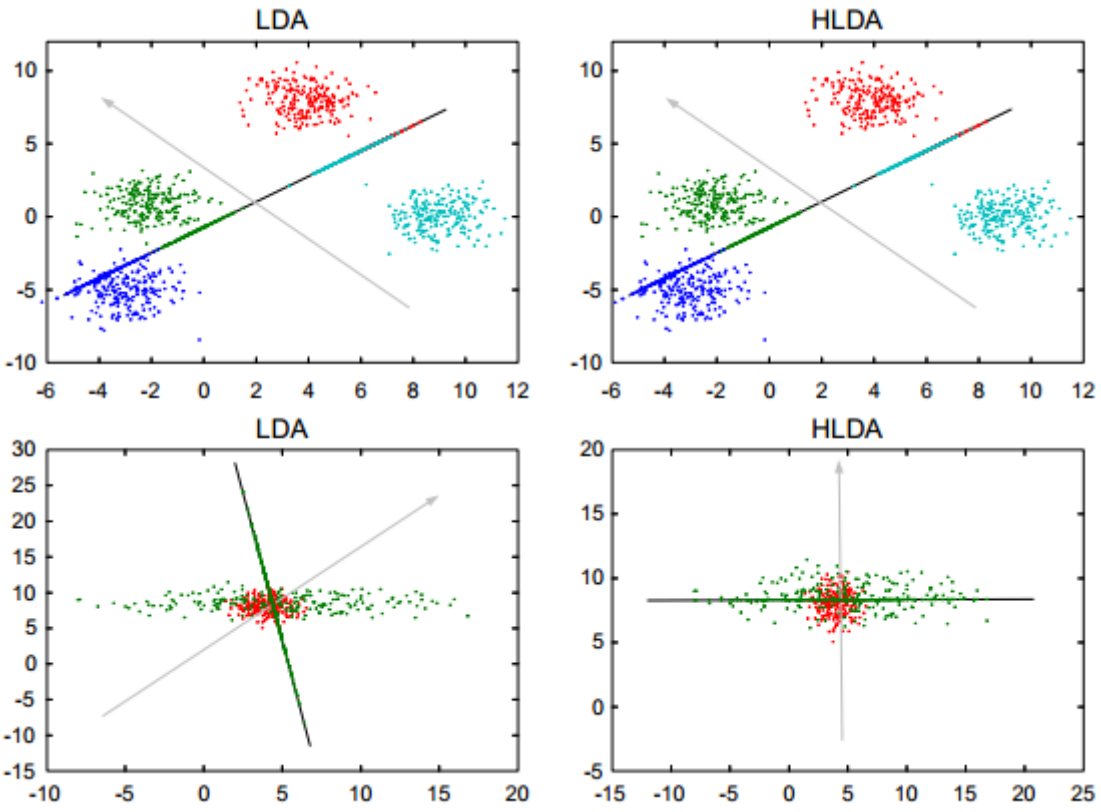


*Figure 1.9: Limitation de LDA*

- *HLDA (HeteroscedasticLinear Discriminant Analysis)*

L'idée de HLDA repose sur l'hypothèse qu'après la transformation linéaire de la matrice $\theta$, seulement les p-colonnes portent l'information discriminante et couvrent le sous espace de dimension p dans lequel les moyens et les variances différent, alors que les autre seront homogènes.

On écrit donc : $\theta = [\,\theta p \quad \theta 0]$ avec y = $\theta^T x = \begin{bmatrix} \theta p \\ \theta_0 \end{bmatrix}$ x

L'ensemble de données d'entrée est supposé une distribution gaussienne, après la transformation linéaire, l'ensemble de données de sortie est aussi une distribution gaussienne, la moyenne est donc :

$$\mu_j = [\ \mu_{j,1} \quad \mu_{j,2} \quad \cdots \quad \mu_{j,p}, \quad \mu_{0,p+1} \quad \mu_{0,p+2} \quad \cdots \quad \mu_{0,n}\ ]^T = \begin{bmatrix} \mu_{j,p} \\ \mu_{0,(n-p)} \end{bmatrix}$$

De même, l'expression de covariance pour une classe j est :

$$\Sigma_j = \underbrace{\begin{bmatrix} \Sigma_j^p \\ 0 \end{bmatrix}}_{p} \underbrace{\begin{bmatrix} 0 \\ \Sigma_0^{(n-p)} \end{bmatrix}}_{(n-p)}$$

En se basant sur ces deux définitions, l'expression de la densité de probabilité gaussienne devient :

$$P(y_i) = \frac{|\theta|}{\sqrt{(2\pi)^n |\Sigma_{g(i)}|}} \exp\left(-\frac{1}{2}(y_i - \mu_{g(i)})^T \Sigma_{g(i)}^{-1} (y_i - \mu_{g(i)})\right)$$

Avec g(i) = j représente la projection des vecteurs d'observation vers des classes, pour obtenir le meilleur estimateur pour $\theta$, il faut calculer la dérivée de la fonction de max de vraisemblance par rapport à $\theta$ et la mettre égale à 0

$$\log P(y_1, y_2, \ldots, y_N) = \sum_{i=1}^{N} \left( \log|\theta| - \frac{1}{2}\log((2\pi)^n|\Sigma_{g(i)}|) - \frac{1}{2}(y_i - \mu_{g(i)})^T \Sigma_{g(i)}^{-1}(x_i - \mu_{g_i}) \right)$$

On ne peut pas calculer directement la dérivée de cette expression, parce que les paramètres de covariance et de la moyenne dépende encore de $\theta$,
D'abord, on calcul d'abord :

- la dérivée par rapport à $\mu_j$

$$\begin{aligned}
\frac{\partial \log P}{\partial \mu_j} &= -\frac{1}{2}\sum_{i=1}^{N} \frac{\partial}{\partial \mu_j}(\theta^T x_i - \mu_j)\Sigma_j^{-1}(\theta^T x_i - \mu_j) = \\
&= -\frac{1}{2}\sum_{i=1}^{N}(\Sigma_j^{-1} + (\Sigma_j^{-1})^T)(\theta^T x_i - \mu_j) = 0 \\
&\Rightarrow N(\Sigma_j^{-1} + (\Sigma_j^{-1})^T)\mu_j = \sum_{i=1}^{N}(\Sigma_j^{-1} + (\Sigma_j^{-1})^T)\theta^T x_i \\
&\Rightarrow \mu_j = \frac{1}{N}\sum_{i=1}^{N}\theta^T x_i = \theta^T \frac{1}{N}\sum_{i=1}^{N} x_i = \underline{\theta^T \bar{X}_j}
\end{aligned}$$

- la dérivée par rapport à $\sum_j$

$$\begin{aligned}
\frac{\partial \log P}{\partial \Sigma_j} &= -\frac{1}{2}\sum_{i=1}^{N} \frac{\partial}{\partial \Sigma_j}\left((\theta^T x_i - \mu_j)\Sigma_j^{-1}(\theta^T x_i - \mu_j) + \log|\Sigma_j|\right) = \\
&= +\frac{1}{2}\sum_{i=1}^{N}(\Sigma_j^{-1}(\theta^T x_i - \mu_j)(\theta^T x_i - \mu_j)^T\Sigma_j^{-1} - (\Sigma_j^T)^{-1}) = \\
&= -\frac{1}{2}N(\Sigma^T)^{-1} + \Sigma_j^{-1}\sum_{i=1}^{N}(\theta^T x_i - \mu_j)(\theta^T x_i - \mu_j)^T\Sigma_j^{-1} = 0 \\
&\Rightarrow \Sigma_j = \frac{1}{N}\sum_{i=1}^{N}\theta^T(x_i - \bar{x}_j)(\theta^T(x_i - \bar{x}_j))^T \\
&\Rightarrow \Sigma_j = \frac{1}{N}\theta^T(\sum_{i=1}^{N}(x_i - \bar{x}_j)(x_i - \bar{x}_j)^T)\theta = \underline{\theta^T \bar{W}_j \theta}
\end{aligned}$$

En remplaçant ces valeurs par leurs expressions, on trouve :

$$\begin{aligned}\log P(x_i) &= \sum_{i=1}^{N}(-\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(\theta^T - \theta_p^T\bar{X}_j)^T(\theta_p^T\bar{W}_j\theta_p)^{-1}(\theta^T - \theta_p^T\bar{X}_j) - \\ &- \frac{1}{2}(\theta^T - \theta_{(n-p)}^T\bar{X}_j)^T(\theta_{(n-p)}^T\bar{W}_j\theta_{(n-p)})^{-1}(\theta^T - \theta_{(n-p)}^T\bar{X}_j) + N\log|\theta| = \\ &= \frac{-Nn}{2}\log 2\pi - \sum_{j=1}^{J}\frac{N_j}{2}\log|\theta_p^T\bar{W}\theta_p| - \frac{N}{2}\log|\theta_0^T\bar{T}\theta_0| + N\log\theta + \\ &-\frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{N_j}(x_i - \bar{X}_j)^T\theta_p(\theta_p^T\bar{W}_j\theta_p)^{-1}\theta_p^T(x_i - \bar{X}_j) + \\ &-\frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{N_j}(x_i - \bar{X})^T\theta_0(\theta_0^T\bar{T}\theta_0)^{-1}\theta_0^T(x_i - \bar{X})\ .\end{aligned}$$

Cette expression peut être simplifiée en remplaçant

$$\frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{N_j}(x_i - \bar{X}_j)^T\theta_p(\theta_p^T\bar{W}_j\theta_p)^{-1}\theta_p^T(x_i - \bar{X}_j) =$$

$$\frac{1}{2}\sum_{j=1}^{J}\text{trace}\,(\theta_p(\theta_p^T\bar{W}_j\theta_p)^{-1}\theta_p^T\bar{W}_j) = \sum_{j=1}^{J}\frac{N_jp}{2} = \frac{Np}{2}$$

Et

$$\frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{N_j}(x_i - \bar{X})^T\theta_0(\theta_0^T\bar{T}\theta_0)^{-1}\theta_0^T(x_i - \bar{X}) = \frac{N(n-p)}{2}$$

Et finalement on obtient :

$$\hat{\theta} = \arg\max_\theta\left(-\frac{N}{2}\log|\theta_0^T\bar{T}\theta_0| - \sum_{j=1}^{J}\frac{N_j}{2}\log|\theta_p^T\bar{W}_j\theta_p| + N\log|\theta|\right)$$

Les méthodes de la transformation linéaire comme le LDA ou HLDA ou leurs variantes (SHLDA proposé dans [6], MAP-SHLDA) sont largement utilisés pour réduire la dimension dans toute sorte de problème de classification. Les avantages de ces méthodes sont : elles peuvent être calculées sur des ensembles de données énormes, ainsi qu'elles ont un nombre de paramètres relativement faible, pourtant elles donnent des estimations robustes pour le modèle sous-jacent.

- <u>Speech enhancement</u>

Le « speech enhancement » est un ensemble de techniques de traitement du signal de parole qui filtre le bruit et met la parole en valeur. L'amélioration de la qualité de signal (Speech Enhancement) est un problème très difficile, les sources de complexité sont de natures différentes, d'une part la nature et les caractéristiques du signal bruité varie d'une application à une autre, et donc il n'est pas évident de trouver un modèle ou un algorithme qui fonctionne pour toutes les applications, et quel que soit l'environnement du travail, d'autre part il y a deux critères pour mesurer les performances d'une application : la qualité d'un signal de parole

(mesure subjective) et l'intelligibilité de la parole[(1)] (mesure objective), et donc il est plus difficile de satisfaire à ces deux critères en même temps.

Dans ce qui suit, on va présenter quelques méthodes qui ont pour objectif de restaurer un signal utile à partir des observations corrompues par un bruit supposé souvent additif, cette hypothèse est souvent utilisée, à la fois pour sa simplicité, mais aussi elle permet de modéliser un grand nombre de situations pratiques. Le signal observé est donc considéré comme la somme du signal de parole et du bruit ambiant. Ce modèle omet tout bruit convolutif, électrique ou de quantification.

- *La soustraction spectrale*

La soustraction spectrale est la méthode de débruitage la plus ancienne [7]. Elle opère dans le domaine fréquentiel et a pour principe de soustraire une estimé du bruit à partir du signal observé. Le bruit est supposé additif, stationnaire ou légèrement variant ce qui nous permet de l'estimer pendant les périodes de silence. Il existe deux versions de base de la soustraction spectrale se différenciant l'une de l'autre par l'utilisation soit de la puissance soit de l'amplitude.

- Si $\left|\hat{S}(v)\right| = |Y(v)| - |\hat{B}(v)|$ Il s'agit de la soustraction spectrale d'amplitude
- Si par contre, le bruit estimé est donné par son spectre de puissance, on aura la soustraction spectrale de puissance :

$$|\widehat{S}(\nu)|^2 = |Y(\nu)|^2 - |\widehat{B}(\nu)|^2$$

Vu que le second terme de l'équation précédente peut être négatif, on peut le rendre positif en changeant de signe ou bien en l'annulant comme l'équation suivante. Ceci fait partie des premières améliorations apportées à la soustraction spectrale

$$|\widehat{S}(\nu)|^2 = \begin{cases} |Y(\nu)|^2 - |\widehat{B}(\nu)|^2 & \text{si } |Y(\nu)|^2 > |\widehat{B}(\nu)|^2 \\ 0 & \text{sinon.} \end{cases}$$

Le passage dans le domaine temporel est réalisé par la transformée de Fourrier inverse en gardant la phase du signal bruité. On se permet de procéder ainsi, d'une part, parce que notre oreille est peu sensible aux variations de la phase et, d'autre part, parce qu'une estimation de la phase est une tâche très compliquée.

$$\widehat{s}(t) = \text{IFFT}\left[|\widehat{S}(\nu)|.e^{i\times \arg Y(\nu)}\right]$$ avec $|\hat{S}(v)|$ le spectre d'amplitude du signal rehaussé et arg Y(v) la phase du signal bruité

(1)Intelligibilité de la parole : signifie la capacité de comprendre un message linguistique contenu dans un signal de parole.

- *Le filtre de Wiener*

Une autre approche de débruitage consiste à appliquer au signal bruité un filtre optimal appelé filtre de Wiener [7]. Cette approche utilise également le spectre d'énergie et par conséquent, un estimé de signal dans le domaine spectral est requis. Comme dans le cas de la soustraction spectrale, cet estimé peut être obtenu par l'équation précédente. Ce type de filtrage consiste à obtenir une estimation du signal original suivant le critère minimum de l'erreur quadratique moyenne. Ce critère conduit au filtre de Wiener dont la réponse fréquentielle est donnée par :

$$H(f) = \frac{|X(f)|^2}{|Y(f)|^2} = \frac{|X(f)|^2}{|X(f)|^2 + |N(f)|^2}$$

Pour cette méthode, il est essentiel de connaître l'estimation du bruit pour évaluer$|N(f)|^2$, concernant$|X(f)|^2$, cela est obtenu par soustraction de l'estimation de $|N(f)|^2$du spectre d'énergie du signal bruité et du conduit au résultat :

$$H(f) = \begin{cases} 1 - \frac{1}{SNR} & \text{si} \quad 1 - \frac{1}{SNR} > 0 \\ 0 & \text{sinon} \end{cases}$$

Avec SNR est le rapport signal-bruit en entrée.

Ce filtre convient non seulement au son, mais aux images, car le spectre de fréquence de la plupart des images visuelles est souvent bien conditionné et peut être estimé facilement.

- *Filtrage de Kalman*

Le filtre de Kalman est une méthode visant à estimer des paramètres d'un système évoluant dans le temps à partir de mesures bruités, on retrouve ce filtre dans un bon nombre de domaines relatifs au traitement du signal, radar, traitement d'images,…

Le fonctionnement du filtre de Kalman est expliqué en détail en [3] peut se diviser en deux étapes :

- Une première étape de prédiction de l'estimation selon le modèle du système. Pour ce faire, le filtre de kalman reprend l'estimation précédente des paramètres et de l'erreur et prédit les nouveaux paramètres et la nouvelle erreur en fonction de la modélisation du système.
- La seconde étape va faire la mise à jour de cette prédiction grâce aux nouvelles mesures. Ces mesures (par définition bruités) vont permettre d'obtenir une estimation des paramètres et de l'erreur à partir de la prédiction faite. Si jamais le modèle comporte des erreurs, cette étape de mise à jour permettra de les rectifier.

De façon générale, l'extraction des paramètres acoustiques est une étape très importante dans les systèmes de reconnaissances automatique de locuteurs. Son but essentiel est d'extraire les données pertinentes à l'étape de modélisation statistique, et minimise ainsi les données redondantes et le bruit qui se présent dans un signal de parole. Il est à noter qu'il existe d'autre méthodes et variantes qui ne sont pas présenté dans cette partie.

### *1.3.3 Modélisation*

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

On peut distinguer quatre grandes approches pour la construction des modèles de locuteur : l'approche vectorielle, statistique, prédictive et connexionniste.

- Approche vectorielle

Dans l'approche vectorielle, un modèle de locuteur est un ensemble de vecteurs de paramètres représentatifs de l'espace acoustique construit lors de la phase de paramétrisation des signaux d'apprentissage. Lors de la reconnaissance, une distance entre cet ensemble de vecteurs et les vecteurs de paramètres issus des signaux de test est calculée.

L'approche vectorielle compte deux grandes techniques : la programmation dynamique et la quantification vectorielle.

- *La programmation dynamique*

La programmation dynamique (Dynamic Time Warping : DTW) consiste à aligner temporellement une séquence de vecteurs de paramètres de test avec une séquence de vecteurs d'apprentissage. Dans ce cas, le modèle de locuteur est tout simplement l'ensemble des vecteurs de paramètres obtenus après paramétrisation des signaux d'apprentissage. Une distance est calculée entre vecteurs d'apprentissage et de test et moyennée sur l'ensemble de la séquence.

La programmation dynamique est utilisée exclusivement en mode dépendant du texte, c'est une approche très rapide et fournie des résultats relativement bonne, mais elle est très sensible à la qualité d'alignement et notamment au choix du point de départ.

- *La quantification vectorielle*

La quantification vectorielle (Vector Quantization : VQ) repose sur un partitionnement de l'espace acoustique en sous-espaces. Chaque sous-espace est associé à leur vecteur centroïde (i.e. à un vecteur de paramètres représentant l'ensemble des vecteurs composant le sous-espace). Dans ces conditions, un modèle de locuteur est composé d'un ensemble de vecteurs centroïdes, appelé dictionnaire de quantification (codebook).

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

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

- Approche statistique

L'approche statistique consiste à représenter une séquence de vecteurs acoustiques issus de la paramétrisation par des statistiques à long terme. Les premiers travaux suggèrent d'utiliser les paramètres du spectre moyen à long terme comme un seul modèle des locuteurs. Lors de la reconnaissance, le spectre moyen estimé sur les vecteurs de test est comparé, à l'aide d'une distance spectrale, au spectre moyen issu de l'apprentissage.

Par la suite, l'approche statistique a été enrichie par l'introduction de statistiques d'ordre supérieur (statistiques d'ordre 2) qui permettent notamment de caractériser la variation des paramètres acoustiques (matrice de covariance).

- *Méthodes statistiques du second ordre*

Le principe des Méthodes Statistiques du Second Ordre (MSSO) est de représenter une séquence de vecteurs acoustiques par une distribution gaussienne multidimensionnelle. Le modèle d'un locuteur se résume alors par le triplet $\{\bar{x}, X_0, \mathrm{M}\}$ où $\bar{x}$ est un vecteur moyen, $X_0$ est une matrice de covariance, tous deux estimés à partir de la séquence de M vecteurs acoustiques.

L'avantage majeur des MSSO est leur simplicité de mise en œuvre, elles sont performantes sur de courtes durées (3 secondes), et ne capturent que les caractéristiques stables le long du signal de parole. Les variations locales sont, quant à elles, moyennées et ne sont pas prises en compte par les modèles.

- *Mélange de gaussiennes*

Un moyen de pallier ce problème (variations locales moyennées par les MSSO) est de considérer les modèles à mélanges de gaussiennes multidimensionnelles (Gaussian Mixture Model : GMM) [21]. Dans ce contexte, une séquence de vecteurs acoustiques d'apprentissage est représentée par un mélange de gaussiennes i.e. une somme pondérée de M distributions gaussiennes multidimensionnelles, chacune caractérisée par un vecteur moyen et une matrice de covariance.

Lors de l'apprentissage, les paramètres des modèles de locuteur (vecteur moyen $\bar{x}_i$, matrice de covariance $\Sigma_i$, pondération $p_i$ de chaque distribution gaussienne) sont généralement

estimés à l'aide de l'algorithme EM (Expectation-Maximization) couplé à l'approche par Estimation du Maximum de Vraisemblance (EMV).

Par les performances qu'ils obtiennent, les mélanges de gaussiennes sont considérés comme la modélisation « état de l'art » des systèmes de RAL en mode indépendant du texte. L'inconvénient majeur de cette technique est la quantité de signaux d'apprentissage requise pour une bonne estimation des paramètres des modèles.

- *Modèles de Markov cachés*

Les modèles de Markov cachés (Hidden Markov Models : HMM) permettent de caractériser les variations temporelles du signal de parole. Ils reposent sur une succession d'états associés à des probabilités de transition d'un état à l'autre. Une ou plusieurs distributions de probabilité associées à chaque état caractérisent les probabilités d'émission des vecteurs acoustiques par un état.

Lors de la reconnaissance, la vraisemblance pour qu'une séquence de vecteurs de test soit issue de la chaîne de Markov est calculée.

***NB :*** Les mélanges de gaussiennes peuvent être considérés comme un modèle de Markov caché à un seul état. De même, la quantification vectorielle décrite précédemment est souvent interprétée comme une dégénérescence des modèles de Markov cachés à un seul état pour lequel les probabilités d'émission sont remplacées par des mesures de distance.

Les modèles de Markov cachés s'appliquent parfaitement au mode dépendant du texte, obtenant d'excellents résultats. En revanche, l'utilisation des modèles HMM en mode indépendant du texte n'améliore pas les performances obtenues par des modèles plus simples à base de GMM.

- Approche prédictive

L'approche prédictive repose sur le principe qu'une trame de signal peut être prédite par la seule observation des trames précédentes. De par ce concept, cette approche est considérée dans la littérature comme une approche dynamique i.e. une approche tenant compte des informations dynamiques véhiculées par le signal de parole. Elle s'appuie principalement sur l'estimation d'une fonction de prédiction, propre à chaque locuteur et apprise sur les signaux d'apprentissage. Lors de la reconnaissance, une erreur de prédiction peut être calculée entre une trame prédite (par la fonction de prédiction) et la trame réellement observée dans la séquence de test.

- Approche connexionniste

L'approche connexionniste repose sur la discrimination entre locuteurs. Elle consiste à fournir à un réseau de neurones un ensemble de signaux de parole issus d'une population de locuteurs afin que ce dernier apprenne comment discriminer un locuteur des autres. L'approche connexionniste se résume, par conséquent, à une tâche de classification. Un modèle se présente sous la forme d'un ou plusieurs réseaux de neurones pour lequel la séquence de vecteurs d'apprentissage du locuteur concerné ainsi que celles des autres locuteurs du système sont fournies en entrée. Différents types de modèles de réseaux sont proposés dans la littérature : Les réseaux multicouches (MLP) utilisés au départ ont rapidement présenté des problèmes lors de l'apprentissage, qui devient long et complexe quand le nombre de locuteurs est grand. Pour éviter ce problème, la tâche de classification est divisée en plusieurs sous-tâches de complexité moindre. On peut aller jusqu'à construire un classificateur pour chaque paire de locuteurs. Un apprentissage plus rapide peut également être obtenu en remplaçant les réseaux multicouches par des réseaux RBF (Radial Basis Function). Les réseaux TDNN (Time Delay Neural Networks) permettent quant à eux de prendre en compte l'information dynamique en réalisant la classification sur des segments de plusieurs trames concaténées. Enfin, l'approche LVQ (Learning Vector Quantization) est une méthode de type quantification vectorielle avec apprentissage discriminant des vecteurs de référence à l'aide d'un réseau de neurones.

Le principal inconvénient de l'approche connexionniste en reconnaissance est la modularité. En effet, dans le cas d'un apprentissage discriminant, les modèles de tous les locuteurs doivent être réappris quand une nouvelle personne est ajoutée dans la base.

### *1.3.4 Décision et mesures de performances*

La stratégie mise en jeu dans cette partie dépend essentiellement des deux processus : la vérification et l'identification automatique de locuteur.

- Identification automatique de locuteur

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

$$I_c = \frac{Nombre\ de\ tests\ correctement\ identifiés}{Nombre\ total\ de\ tentatives}$$

Et

$$I_i = \frac{Nombre\ de\ tests\ mal\ identifiés}{Nombre\ total\ de\ tentatives}$$

Avec :

$$I_c + I_i = 100\%$$

- Vérification automatique de locuteur

Consiste à vérifier l'adéquation du message vocale avec la référence acoustique du locuteur qu'il prétend être. C'est une décision en tout ou rien. Les performances de vérification de locuteur sont données en termes des faux rejets $f_r$, et de fausses acceptations $f_a$.

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

$$f_r = \frac{Nombres\ de\ tentatives\ d'abonnés\ rejetées}{Nombre\ total\ de\ tentatives\ d'abonnés}$$

***Fausse acceptation*** : erreur commise lorsqu'un imposteur est malencontreusement accepté en tant qu'utilisateur légitime (i.e. erreur commise lors d'un test imposteur) ;

$$f_a = \frac{Nombres\ de\ tentatives\ d'imposteurs\ acceptés}{Nombre\ total\ de\ tentatives\ d'imposteurs}$$

## 1.4 Evolution d'un système de reconnaissance automatique de locuteur

Les recherches sur la reconnaissance du locuteur ont été entreprises depuis plus de 50 ans, et continues d'être un domaine actif de traitement de la communication parlée. Le développement de la technologie de la reconnaissance du locuteur est étroitement concomitant avec l'avancement dans la connaissance de la parole, le traitement du signal et la technologie des ordinateurs.

La reconnaissance du locuteur par les humains a été largement étudiée dans les années 1960. La motivation de ces études était d'apprendre comment l'homme reconnait les locuteurs et la fiabilité d'un humain à reconnaitre un locuteur. Le travail le plus important qui a stimulé la recherche sur la reconnaissance du locuteur par la machine a été réalisé par Kersta qui a introduit le spectrogramme (ou il l'a noté comme empreinte vocale) en tant que moyen d'identification personnelle.

Dans les années 1970, l'attention a été tournée vers la reconnaissance du locuteur par ordinateur et devient la reconnaissance automatique du locuteur. A cette époque, les systèmes de reconnaissance du locuteur, en général, ne portait que sur une petite population (moins de 20 locuteurs). La transformée de Fourier, les techniques de prédiction linéaire et d'analyse cepstrale ont été appliquées pour générer des paramètres du locuteur. Les moyennes long-terme de ces paramètres ont été utilisées comme références des locuteurs.

Dans les années 1980, des méthodes statistiques de reconnaissance des formes plus compliquées ont été investiguées, par exemple, l'alignement temporel dynamique(DTW) et la quantification vectorielle (VQ), pour des systèmes de reconnaissance du locuteur à grande échelle (>100 locuteurs). La contribution des caractéristiques statiques et dynamiques pour la reconnaissance du locuteur a également été étudiée.

Depuis les années 1990, la mise à disposition de bases de données de parole plus importantes (par exemple, corpus YOHO) a boosté les études sur des modèles plus compliqués

pour la représentation des locuteurs. Ces modèles comprennent les modèles stochastiques (par exemple, les modèles de Markov cachés (HMM)), le modèle de mélange de Gaussiennes (GMM), les réseaux de neurones (par exemple, Perceptron Multicouches MLP), fonctions a base radiale (RBF)) et les machines a vecteurs de support (SVM)), …etc. Parmi ces techniques de modelisation, la GMM a été reconnu comme la plus efficace à caractériser la distribution de la densité des données de la parole et a été considérée comme la technique de modélisation dominante pour les systèmes de reconnaissance du locuteur. En ce qui concerne l'extraction des caractéristiques, les coefficients cepstraux incorporant le modèle auditif, connus sous le nom de Coefficients Cepstraux à Fréquence Mel (MFCC) et leurs coefficients dynamiques ont été les caractéristiques ou paramètres dominants. Un système avec les paramètres MFCC, une modélisation GMM est considérée comme un système de référence en mode indépendant du texte pour comparer les nouvelles technologies [25].

Les années 2000 ont connues l'apparition d'une nouvelle famille de techniques de normalisation de scores [12], dans laquelle des scores sont normalisé par la soustraction de la moyenne, divisés par l'écart type des scores imposteurs : $\check{S} = \frac{s-\mu_I}{\sigma_I}$ (avec $\check{S}$ est le score normalisé, s le score original, $\mu_I \; et \; \sigma_I$ sont la moyenne et l'écart type) Les techniques les plus couramment utilisées sont : Znorm, Hnorm, Tnorm, Htnorm, Cnorm, et Dnorm, en outre, les paramètres de haut niveau (phonèmes, paramètres idiolectaux, sémantique, accent, prononciation…) proposé en 2001 par Doddington, ont été largement utilisées en vérification de locuteur en mode indépendant de texte [11].

## 1.5 Les approches classiques utilisées dans le système de reconnaissance automatique de locuteur

Différentes méthodologies sont utilisées en RAL pour réaliser les références de locuteurs. Les approches génératives regroupent des méthodes qui utilisent les données d'apprentissage pour modéliser les densités de probabilité de chaque classe, par une famille de fonctions paramétriques. L'approche générative dominante pour représenter la référence du locuteur, en RAL indépendante du texte, est le modèle de mélanges de Gaussiennes (GMM, Gaussian Mixture Model) qui constitue l'état de l'art des systèmes de RAL.

Il existe d'autres approches génératives comme les modèles de Markov cachés (HMM, Hidden Markov Model). Les HMM sont très employés en RAL dépendante du texte car ils sont capables de capturer les dépendances temporelles entre différentes variables aléatoires.

Les approches à base de quantification vectorielle ont été utilisées en RAL. Elles proposent une représentation minimale d'une classe de paramètres observes : un représentant (dans un dictionnaire) pour chaque classe. Chaque classe de paramètres est déterminée par un algorithme de classification du type K-moyennes. Cette représentation est choisie en minimisant la distance entre le centroïde et les paramètres de la population observée. Ces approches ne sont plus très employées depuis l'apparition des GMM en RAL.

L'approche discriminante : les Support Vector Machine (SVM) est largement utilisé en RAL. A l'origine, ils ont été conçus comme une fonction discriminante permettant de séparer au mieux des régions complexes dans des problèmes de classification à 2 classes. Ils démontrent aujourd'hui des performances similaires à l'approche GMM. Ces deux méthodes sont aussi combinées dans un nouveau formalisme, le GMM/SVM Super-Vecteur qui profite des capacités génératives du GMM et discriminantes du SVM.

## 1.6 Domaines d'application

Nos voix ne sont pas seulement un moyen de communiquer. Elles offrent également un moyen fiable de nous reconnaître, et font partie intégrante de notre identité. C'est la raison pour laquelle les banques et d'autres grandes entreprises se tournent aujourd'hui vers l'authentification vocale.

La voix humaine est unique. Elle est avec nous tout le temps contrairement à nos clés de voitures, et aux mots de passes ou codes PIN qu'on peut très souvent oublier. C'est à la fois cette sécurité et cette simplicité d'usage offerte par l'authentification biométrique vocale qui poussent les banques, les opérateurs de télécommunications et autres grandes organisations à choisir ce mode d'authentification.

La biométrie vocale, tout comme la reconnaissance et la synthèse vocale, s'est d'abord propagée dans les serveurs vocaux automatiques des centres d'appels. Mais aujourd'hui, elle est également utilisée dans des domaines aussi variés que l'authentification mobile et le paiement par cartes de crédit.

- **Sécurisation des applications mobiles**

Les grandes entreprises voient désormais leurs clients utiliser massivement les canaux mobiles pour prendre contact et effectuer les opérations courantes. C'est même devenu une attente forte des clients et des consommateurs. Mais la multiplication des applications et services en ligne fait qu'il devient difficile de gérer tous ces mots de passes, de forme et de tailles différentes. L'authentification vocale devient dès lors le mode d'authentification mobile idéal. Il suffit simplement de donner une simple phrase clé à prononcer à un client pour vérifier son identité.

En plus d'éliminer la frustration née des mots de passe difficiles à mémoriser ou à saisir, le 'login vocal' réinvente véritablement l'authentification mobile. Le mobile devenant de plus en plus le point de contact principal entre un consommateur et un fournisseur de services, améliorer l'expérience utilisateur et la sécurité deviennent une priorité.

- **Sécurisation des transactions à risque par carte de crédit**

La reconnaissance de locuteur constitue aussi une solution sûre et pratique pour vérifier les transactions à risque par carte de crédit (par exemple celles en dehors des habitudes de consommation du client ou de son emplacement géographique habituel). Quand une opération à risque est détectée, une demande de vérification de la transaction peut être envoyée au titulaire de la carte de crédit, via un appel sortant automatique, sur son téléphone portable. Le détenteur est alors invité à prononce une phrase clé : "J'autorise cette transaction par ma signature vocale."

A l'inverse, si la transaction est suspecte, il peut tout aussi facilement rejeter celle-ci, ce qui permet alors à l'institution financière d'investiguer sur les transactions marquées comme suspectes.

- **Paiement en ligne**

La reconnaissance de la voix peut être utilisée pour sécuriser des paiements en ligne, typiquement des paiements à risque tels que le premier paiement en ligne sur un site d'e-commerce, par exemple le transfert de l'argent ou des opérations importantes. Lorsque ces opérations sont effectuées, un appel sortant automatique est émis vers le téléphone portable du titulaire du compte effectuant l'opération. Si cette opération est valide, l'utilisateur est invité à confirmer le paiement de la même façon qu'il peut confirmer l'achat par carte de crédit.

- **Aide aux handicapés**

La reconnaissance de locuteur est très utile dans ce cas, elle offre la possibilité de saisir les données à la voix, commandes vocales (ouverture porte, contrôle des équipements au domicile).

# Conclusion

Dans ce chapitre, nous avons présenté de façon générale l'état de l'art d'un système de reconnaissance automatique de locuteur, nous avons présenté, également, la structure générale d'un système RAL et ses composants modulaires. Pour chaque module, nous avons décrit les différentes techniques utilisées en citant leurs avantages et leurs faiblesses, et nous avons terminé par la présentation des domaines d'application de cette discipline

# Chapitre 2 : La reconnaissance automatique du locuteur par la voix IP

## 2.1 Introduction

La voix sur IP connait aujourd'hui une croissance remarquable, il est nécessaire de développer un système de reconnaissance automatique de locuteur basé sur la voix IP, pour cela plusieurs laboratoires et organismes de recherches ont été spécialisés sur l'amélioration de cette discipline.

Dans ce chapitre, nous allons présenter un aperçu générale sur la théorie de la voix IP, en commençant par le concept de la voix IP, passant aux avantages et inconvénients de cette nouvelle technologie, et les protocoles utilisés, et terminant par la présentation des différents travaux de recherches et publications faites dans le cadre de la reconnaissance automatique du locuteur par la voix IP.

## 2.2 La théorie de la voix sur IP

La Voix sur IP (en anglais, *Voice over IP* ou VoIP) est le nom d'une nouvelle technologie de télécommunication vocale en pleine émergence qui transforme la téléphonie. Cette technologie marque un tournant dans le monde de la communication en permettant de transmettre de la voix sur un réseau numérique et sur Internet. C'est en 1996 que naquit la première version Voix sur IP, appelée H323. Depuis, la technologie Voix sur IP a progressé à mesure que les entreprises découvraient ses avantages pour accroître la productivité et l'efficacité de leurs réseaux.

La voix sur IP (Voice over IP) est une technologie de communication vocale qui consiste à encapsuler un signal audio numérisé (en général la voix) dans des paquets IP circulant sur internet. Selon le type de terminal utilisé (un Ordinateur ou un Téléphone), on distingue trois modes d'accès possibles de la voix sur IP :

- ***La voix IP entre deux ordinateurs*** : Cela nécessite que les deux interlocuteur soient équipés informatiquement et dialoguent en utilisant de simples applications genre « NetMeeting » ou « Skype » utilisant pour cela un simple micro et des hauts parleurs. Ce genre de communication est gratuit, exception faite du coût du logiciel. (*figure 2.1*)

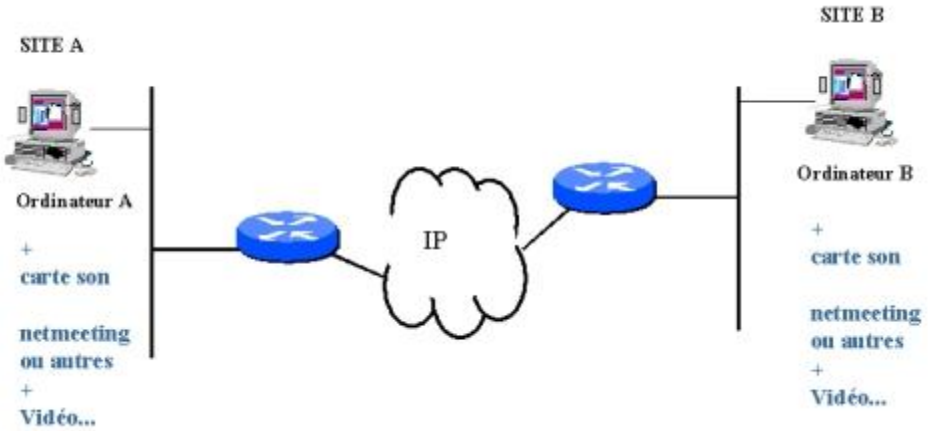


*Figure 2.12: La Voix sur IP entre deux ordinateurs*

- ***La voix sur IP entre un ordinateur et un téléphone :*** Cela nécessite la mise en œuvre d'une passerelle soit au départ de l'appel soit a l'arrivée afin de faire transiter la communication d'un réseau IP à un réseau téléphonique (RTC). L'appel est taxé uniquement pour la traversée du réseau téléphonique. Ainsi, pour les appels internationaux, plus la proportion du segment IP est grande, plus l'économie réalisée est importante. (*figure 2.2*)


*Figure2.2: La voix sur IP entre un ordinateur et un téléphone*

- ***La voix sur IP entre deux téléphones :*** Lorsque l'appelant et l'appelé sont tous les deux sur téléphone, le réseau de transport devient transparent, cela nécessite la mise en œuvre de plusieurs passerelles et la taxation de l'appel dépend de l'opérateur (non taxé dans le cas d'un réseau privé). C'est cette dernière qui réalise le plus l'intégration voix/données (figure 2.3)


*Figure 33: La voix sur IP entre deux téléphones*

### *2.2.1 Architecture de la transmission de la voix IP*

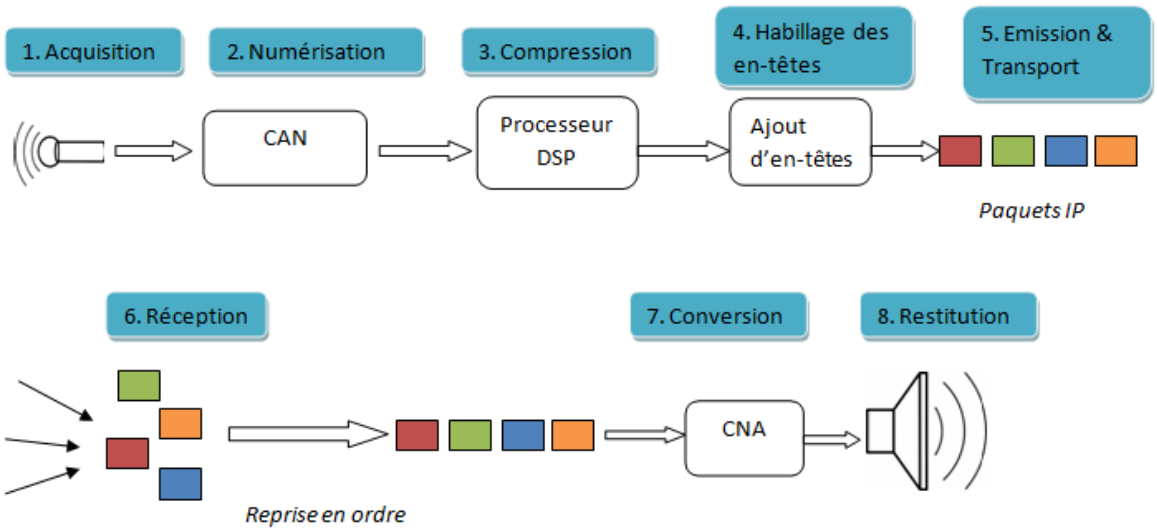


*Figure 2.4: Architecture de la transmission de la voix IP*

La chaîne de transmission de la voix IP commence par la numérisation de la voix, puis par reconversion des paquets numériques en voix à l'arrivée. Le format numérique est plus facile à contrôler, il peut être compressé, routé et converti en un nouveau format meilleur (*figure 2.4*). Le signal numérique est plus tolérant au bruit que l'analogique.

Comme montre la figure ci-dessus, l'architecture de transmission de la voix IP est découpée en 8 étapes :

- ***Acquisition de signal de parole***

  La première étape consiste naturellement à capter la voix à l'aide d'un micro, qu'il s'agisse de celui d'un téléphone ou d'un micro-casque.

- ***Numérisation***

  La voix passe dans un convertisseur analogique numérique qui réalise deux tâches distinctes :

  - Echantillonnage du signal : un prélèvement périodique de ce signal, il s'agit d'enregistrer à des intervalles très rapprochés la valeur d'un signal afin de pouvoir disposer d'un enregistrement proche de la valeur réelle de ce signal.
  - Quantification du signal : consiste à affecter une valeur numérique (en binaire) à chaque échantillon.

- ***Compression :*** Le signal une fois numérisé peut être traité par un DSP (Digital Signal Processor) qui va le compresser, c'est-à-dire réduire la quantité d'informations (bits) nécessaire pour l'exprimer. Plusieurs normes de compression et décompression (Codecs) sont utilisées pour la voix. L'avantage de la compression est de réduire la bande passante nécessaire pour transmettre le signal.

- ***Habillage des en-têtes :*** Les données brutes qui sortent du DSP doivent encore être enrichies en informations avant d'être converties en paquets de données à expédier sur le réseau. Trois couches sont utilisées pour cet habillage :
  - La couche IP : Correspond à l'assemblage des données en paquets. Chaque paquet commence par un en-tête indiquant le type de trafic concerné
  - La couche UDP : Consiste à formater très simplement les paquets. Si l'on restait à ce stade, leur transmission serait non fiable : UDP ne garantit ni le bon acheminement des paquets, ni leur ordre d'arrivée.
  - La couche RTP /RTCP[(1)]: Pour palier l'absence de fiabilité d'UDP, un formatage RTP est appliqué de surcroît aux paquets. Il consiste à ajouter des entêtes de synchronisation pour s'assurer du réassemblage des paquets dans le bon ordre à la réception. RTP est souvent renforcé par RTCP qui comporte, en plus, des informations sur la qualité de la transmission et l'identité des participants à la conversation.

[(1)] RTP : Real-time Transport Protocol
[(1)] RTCP : Real-time Transport Control Protocol

- ***Emission et transport :*** Les paquets sont acheminés depuis le point d'émission pour atteindre le point de réception sans qu'un chemin précis soit réservé pour leur transport. Ils vont transiter sur le réseau en fonction des ressources disponibles et arriver à destination dans un ordre indéterminé.
- ***Réception :*** Lorsque les paquets arrivent à destination, il est essentiel de les replacer dans le bon ordre et assez rapidement, sinon une dégradation de la voix se fera sentir
- ***Conversion numérique analogique*** : C'est l'étape réciproque de l'étape 2, qui permet de transformer les données reçues sous forme de série discrète en un signal électrique continu
- ***Restitution :*** La voix peut être retranscrite par le haut-parleur du casque, du combiné téléphonique ou de l'ordinateur.

### *2.2.2 Protocoles de la voix*

Il existe plusieurs protocoles qui peuvent supporter la voix sur IP tel que le H.323, SIP, MGCP/MEGACO. Les deux protocoles les plus utilisées actuellement dans les solutions voix IP sont le H.323 et le SIP.

- Le protocole H.323

  - *Présentation générale*

H.323 est un protocole de communication englobant un ensemble de normes utilisées pour l'envoi de données audio et vidéo sur internet. Il existe depuis 1996 et a été développé par l'ITU. Concrètement, il est utilisé dans des programmes tels que Microsoft NetMeeting ou encore dans des équipements tels que les routeurs Cisco. H323 est un regroupement de plusieurs protocoles qui concerne trois catégories distinctes : la signalisation, la négociation de codecs et le transport de l'information [*w5*].

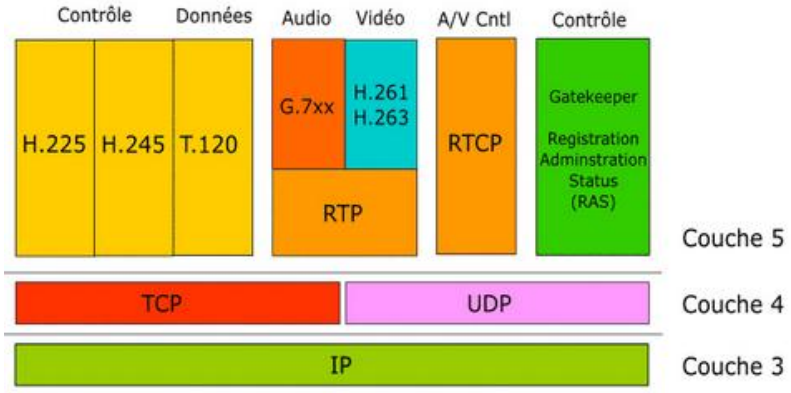


*Figure 2.5 : Pile de protocole H323*

Le respect du standard H.323 permet de garantir un contrôle sur l'utilisation des ressources réseaux et des contraintes de qualité de service. Tous les terminaux H.323 doivent supporter :

- Le protocole H.245 (*figure 2.5*) qui négocie l'ouverture et l'utilisation des canaux ainsi que les paramètres de la communication voix. La négociation est utile pour mettre d'accord les terminaux et les équipements voix qui

communiquent entre eux sur les choix du type des données transportées. Les langages utilisés entre les équipements doivent s'adapter aux contraintes imposées par le support de transmission notamment et par les équipements eux-mêmes. Le choix du codec est très important (G7xx et H26x de la *figure 3.5*), du moins gourmand en bande passante à celui qui offre la meilleur qualité vocale.

- Le protocole H.225 (SIG) pour la signalisation et l'établissement des appels.
- Le protocole H.225 (RAS) (Registration/Admission/Status), qui est le protocole utilisé par le terminal pour communiquer avec le serveur de contrôle d'appels.
- Les protocoles RTP/RTCP transportent les flux audio et vidéo.
- Le T.120 permet l'ouverture d'un canal pour le partage d'applications.

▪ *Les limites du protocole*

Le standard mis de l'ITU permet une interopérabilité entre des équipements de constructeurs différents et est très largement utilisé encore aujourd'hui, il présente toutefois les inconvénients suivants :

- ***Interopérabilité avec les autres normes de visioconférence***

Le fonctionnement de la visioconférence entre les équipements utilisant les protocoles H.320[(2)] et H.323 posent des problèmes et nécessite des gateways (passerelles).

H.323 et IP Multicast ne sont, en règle générale, pas compatibles, sauf dans le cadre de VRVS[(3)] qui permet un certain degré d'interopérabilité, mais ne gère pas la norme T.120.

Le développement de l'H.323 a été basé sur la téléphonie et ses différents standards ne sont pas compatibles avec la plupart des protocoles du monde IP (contrairement au protocole SIP)

- ***Problème d'interopérabilité entre équipements***

L'H.323 comprend de nombreuses options susceptibles d'être implémentées de façon différente par les constructeurs et donc de poser des problèmes d'interopérabilité ou de plus petit dénominateur commun (par exemple : le choix du codec). D'autre part, comme le seul codec obligatoire est le codec G.711 (64 Kps) et que le support des autres codecs plus efficaces est optionnel, l'interopérabilité entre produits provenant de constructeurs différents ne signifie pas qu'ils feront un usage optimal de la bande passante. En effet, dans le cas où les codecs à bas débits sont différents, le transport de la voix se fera à 64 Kbps, ce qui, en termes de bande passante, ne présente guère d'avantages par rapport à un système téléphonique classique.

---

[(2)]**H.320** : protocole assurant le transport de la visioconférence sur le réseau RNIS/ISDN/Numéris. Ce protocole garantit une qualité de service au détriment d'un coût d'utilisation important. Il est souvent le plus adapté pour les échanges internationaux mais son usage tend à diminuer du fait de la maturité du protocole H323.

[(3)]**VRVS :** permet de faire de la vidéoconférence multipoint à l'aide d'un simple ordinateur personnel, pratiquement de n'importe quel endroit où l'on dispose d'un accès Internet avec un débit raisonnable (typiquement ADSL).

Le protocole H.323 est une des normes envisageables pour la voix sur IP à cause de son développement inspiré de la téléphonie. Cependant, elle est pour l'instant employée par des programmes propriétaires (Microsoft, etc.). La documentation est difficile d'accès car l'ITU fait payer les droits d'accès aux derniers développements de cette technologie, en dehors des efforts faits par le projet OpenH.323 pour rendre cette technologie accessible à tous. Ainsi son adaptation au réseau IP est assez lourde. C'est pourquoi au fil des recherches sont nées le SIP.

- ***Protocole complexe***

L'H.323 est un protocole complexe, créé initialement pour les conférences multimédia et qui incorpore des mécanismes superflus dans un contexte purement téléphonique. Ceci a notamment des incidences au niveau des terminaux H.323 (téléphones IP, par exemple) qui nécessitent de ce fait une capacité mémoire et de traitement non sans incidence au niveau de leur coût et du délai d'établissement de l'appel.

- Le protocole SIP

  - *Présentation générale*

Le protocole SIP est un protocole normalisé et standardisé par l'IETF, il a été conçu pour établir, modifier et terminer des sessions multimédia. Il se charge de l'authentification et de la localisation des multiples participants, et également de la négociation sur les types de média utilisables par les différents participants en encapsulant des messages SDP (Session Description Protocol). SIP ne transporte pas les données échangées durant la session comme la voix ou la vidéo. SIP étant indépendant de la transmission des données, tout type de données et de protocoles peut être utilisé pour cet échange. Cependant le protocole RTP (Real-time Transport Protocol) assure le plus souvent les sessions audio et vidéo. SIP remplace progressivement H323.

SIP est le standard ouvert de VoIP (Voice Over IP, voix sur IP) interopérable le plus étendu et vise à devenir le standard des télécommunications multimédia (son, image, etc..). Skype par exemple, qui utilise un format propriétaire, ne permet pas l'interopérabilité avec un autre réseau de voix sur IP et ne fournit que des passerelles payantes vers la téléphonie standard. SIP n'est donc pas seulement destiné à la VoIP mais pour de nombreuses autres applications telles que la visiophonie, la messagerie instantanée, la réalité virtuelle ou même les jeux vidéo.

  - *Fonctionnement*

- ***Modèle d'échange***

Le protocole SIP repose sur un modèle Requête/Réponse. Les échanges entre un terminal appelant et un terminal appelé se font par l'intermédiaire de requêtes. La liste des requêtes échangées est la suivante :

INVITE : Permet à un client de demander une nouvelle session
ACK : Confirme l'établissement de la session
CANCEL : met fin à une session pendante
OPTION : permet d'indiquer un certains nombre de paramètres permettant notamment de pouvoir faire de gestion de présence sur le réseau.
REGISTER : enregistrement d'une entité auprès du serveur
BYE : termine une session en cours

- ***Codes d'erreurs***

Une réponse à une requête est caractérisé, par un code et un motif, appelés respectivement code d'état et raison phrase. Un code d'état est un entier codé sur 3 chiffres indiquant un résultat à l'issue de la réception d'une requête. Ce résultat est précisé par une phrase, text based (UTF-8), expliquant le motif du refus ou de l'acceptation de la requête. Le code d'état est donc destiné à l'automate gérant l'établissement des sessions SIP et les motifs aux programmeurs. Il existe 6 classes de réponses et donc de codecs d'état, représentées par le premier digit :

- 1xx = Information - La requête a été reçue et continue à être traitée.
- 2xx = Succès - L'action a été reçue avec succès, comprise et acceptée.
- 3xx = Redirection - Une autre action doit être menée afin de valider la requête.
- 4xx = Erreur du client - La requête contient une syntaxe erronée ou ne peut pas être traitée par ce serveur.
- 5xx = Erreur du serveur - Le serveur n'a pas réussi à traiter une requête apparemment correcte.
- 6xx = Echec général - La requête ne peut être traitée par aucun serveur.

Pratiquement, l'architecture SIP comprend deux types de composants : les User Agent et les serveurs.

- ***Les User Agent***

Le premier type de composant SIP est l'application de l'utilisateur final. Ce peut être, par exemple, un terminal de téléphonie ou de visioconférence sur IP, un serveur audio ou vidéo ou encore une passerelle vers un autre protocole. Ce type de composant est appelé User Agent (UA). Il se décompose en une partie cliente et une partie serveur. La partie cliente, appelée User Agent Client (UAC), envoie les requêtes SIP, et la partie serveur, appelée User Agent Server (UAS), les reçoit. Le principal objectif de SIP est de permettre l'établissement de sessions entre User Agents (*figure 2.6*).

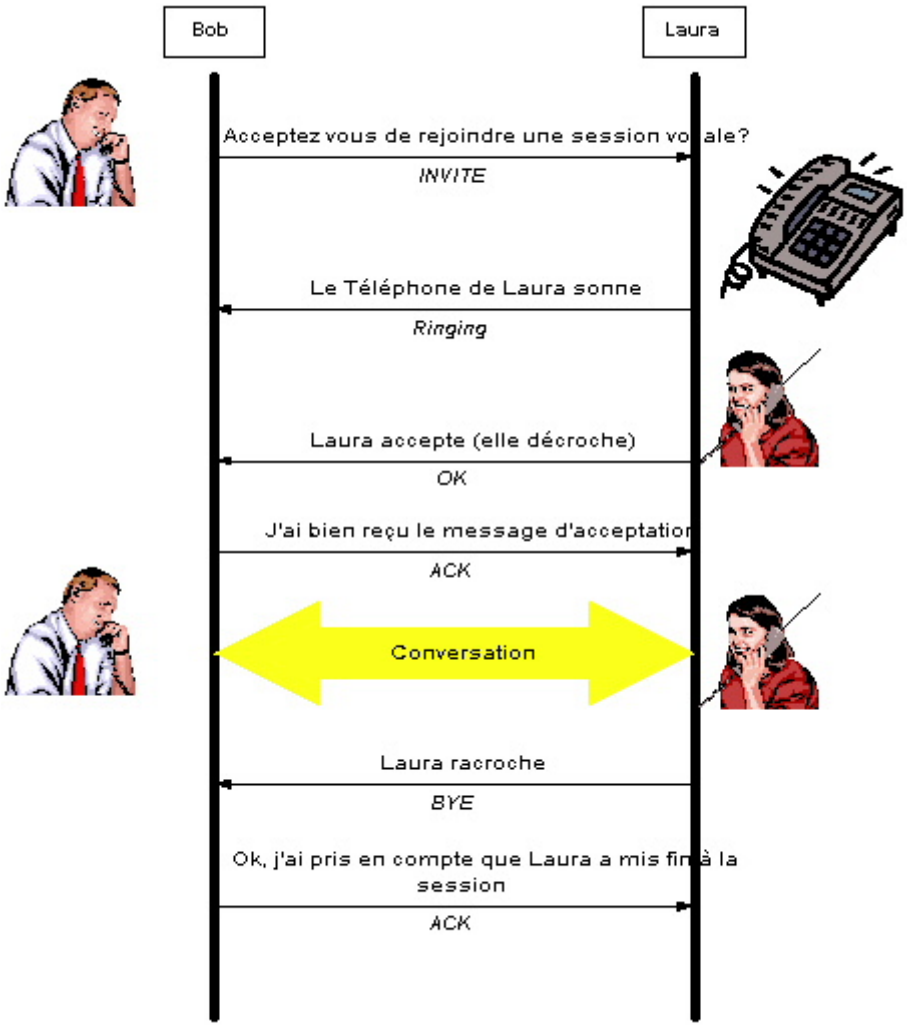


*Figure 2.6: exemple d'établissement d'appel entre deux agents*

- ***Serveur Registrar***

Afin de pouvoir joindre une personne à partir de son adresse SIP, une entité dans le réseau doit maintenir une correspondance (mapping) entre les adresses IP et les adresses SIP : c'est le rôle du serveur Registrar [w5]. Un utilisateur peut donc changer d'adresse, il lui suffit de s'inscrire auprès du Registrar en lui indiquant son adresse SIP et son adresse de machine sur le réseau (*figure 2.7*).

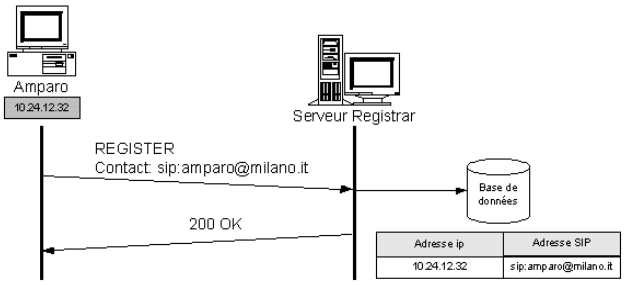


*Figure 2.7 : Exemple d'enregistrement SIP*

La figure (*figure 2.7)* montre l'enregistrement du terminal d'Amparo sur un serveur Registrar. A la réception du message REGISTER, le serveur Registrar a accès à l'adresse IP de la source, Amparo, dans l'en-tête IP du message. Il enregistre alors la correspondance entre cette adresse IP et l'adresse SIP donnée dans le champ « Contact : », soit ici « sip:amparo@milano.it ».

- ***Serveur Proxy***

Un Proxy SIP sert d'intermédiaire entre deux User Agents qui ne connaissent pas leurs emplacements respectifs (adresse IP) [w5] . En effet, l'association URI-Adresse IP a été stockée préalablement dans une base de données par un serveur Registrar. Le Proxy peut donc interroger cette base de données pour diriger les messages vers le destinataire (figure 2.8).


*Figure 2.8: Principe de proxy SIP*

Le serveur Proxy se contente de relayer uniquement les messages SIP pour établir, contrôler et terminer la session (*figure 2.9*). Une fois la session établie, les données, par exemple un flux RTP pour la VoIP, ne transitent pas par le serveur Proxy. Elles sont échangées directement entre les User Agents

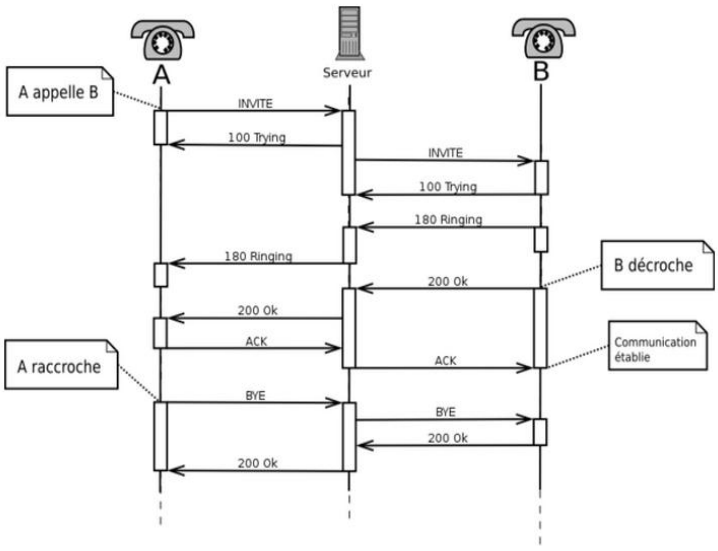


*Figure 2.9 : Session SIP à travers un proxy*

- *Avantages et inconvénients*

Ouvert, standard, simple et flexible sont les principales atouts du protocole SIP, voilà en détails ces différents avantages :

- *Ouvert :* les protocoles et documents officiels sont détaillés et accessibles à tous en téléchargement.
- *Standard* : l'IETF a normalisé le protocole et son évolution continue par la création ou l'évolution d'autres protocoles qui fonctionnent avec SIP.
- *Simple* : SIP est simple et très similaire à HTTP.
- Flexible : SIP est également utilisé pour tout type de sessions multimédia (voix, vidéo, mais aussi musique, réalité virtuelle, etc.).
- Téléphonie sur réseaux publics : il existe de nombreuses passerelles (services payants) vers le réseau public de téléphonie (RTC, GSM, etc.) permettant d'émettre ou de recevoir des appels vocaux.
- Points communs avec H323 : l'utilisation du protocole RTP et quelques codecs son et vidéo sont en commun.

Par contre une mauvaise implémentation ou une implémentation incomplète du protocole SIP dans les User Agents peut perturber le fonctionnement ou générer du trafic superflu sur le réseau. Un autre inconvénient du protocole SIP est sa vulnérabilité face à des attaques de types DoS (dénis de service), détournement d'appel,…

### *2.2.3 Points forts et limites de la voix sur IP*

Différentes sont les raisons qui peuvent pousser les entreprises à s'orienter vers la Voix sur IP comme solution, les avantages les plus marqués sont :

***Réduction des coûts :*** Aujourd'hui, la position des opérateurs téléphoniques classique est rapidement menacée par l'arrivée massive de la téléphonie sur IP, dont la tarification tend vers la gratuité, les coûts des communications interurbaines ont chuté de manière considérable ce qui laisse croire qu'elle a encore de beaux jours devant elle.

***Standards ouverts :*** La voix sur IP n'est plus uniquement H323, mais un usage multi-protocoles selon les besoins de services nécessaires. Par exemple, H323 fonctionne en mode égale à égale alors que MGCP fonctionne en mode centralisé. Ces différences de conception offrent immédiatement une différence dans l'exploitation des terminaisons considérées.

***Un réseau voix, vidéo et données*** (à la fois) : Grâce à l'intégration de la voix comme une application supplémentaire dans un réseau IP, ce dernier va simplifier la gestion des trois applications (voix, réseau et vidéo) par un seul transport IP. Une simplification de gestion, mais également une mutualisation des efforts financiers vers un seul outil.

***Un service PABX distribué ou centralisé :*** *Les* PABX en réseau bénéficient de services centralisés tel que la messagerie vocale et la taxation, etc… Cette même centralisation continue à être assurée sur un réseau Voix IP sans limitation du nombre de canaux. Il convient, pour en assurer une bonne utilisation, de dimensionner convenablement le lien réseau. L'utilisation de la voix IP met en commun un média qui peut à la fois offrir à un moment précis une bande passante maximum à la donnée, et dans une autre période une bande passante maximum à la voix, garantissant toujours la priorité à celle-ci.

Les points faibles de la voix sur IP sont :

***Fiabilité et qualité sonore :*** Un des problèmes les plus importants de la téléphonie sur IP est la qualité de la retransmission qui n'est pas encore optimale. En effet, des désagréments tels la qualité de la reproduction de la voix du correspondant ainsi que le délai entre le moment où l'un des interlocuteurs parle et le moment où l'autre entend peuvent être extrêmement problématiques. De plus, il se peut que des morceaux de la conversation manquent (des paquets perdus pendant le transfert) sans être en mesure de savoir si des paquets ont été perdus et à quel moment.

***Dépendance de l'infrastructure technologique et support administratif exigeant :*** les centres de relations IP peuvent être particulièrement vulnérables en cas d'improductivité de l'infrastructure. Par exemple, si la base de données n'est pas disponible, les centres ne peuvent tout simplement pas recevoir d'appels. La convergence de la voix et des données dans un seul système signifie que la stabilité du système devient plus importante que jamais et l'organisation doit être préparée à travailler avec efficience ou à encourir les conséquences.

***Vol :*** Les hackers qui parviennent à accéder à un serveur voix IP peuvent également accéder aux messages vocaux stockés et au même au service téléphonique pour écouter des conversations ou effectuer des appels gratuits aux noms d'autres comptes.

***Attaque de virus* :** Si un serveur voix IP est infecté par un virus, les utilisateurs risquent de ne plus pouvoir accéder au réseau téléphonique. Le virus peut également infecter d'autres ordinateurs connectés au système.

## 2.3 Evolution de la reconnaissance automatique de locuteur par la voix IP

En 2001, ils ont proposé une approche pour l'extraction des vecteurs caractéristiques directement du signal de parole codé (sans décoder le signal, et puis le traiter) (figure 2.10), en se basant sur le codec G.723.1 qui est principalement utilisé dans la voix sur IP[19], puis ils ont comparé cette approche avec deux approches classiques de la reconnaissance automatique de locuteur : la reconnaissance automatique des chiffres isolés et la reconnaissance automatique de parole continue en mode indépendant de locuteur, les résultats obtenus montrent que cette nouvelle approche est plus performante que les deux approches classiques, parce que cette méthode a deux avantages, d'une part le système n'est affecté que par la distorsion de l'enveloppe spectrale, et d'autre part, dans le cas de perte de paquets, cette approche devient plus efficace puisqu'elle n'est pas limité à l'erreur de manipulation de codecs.

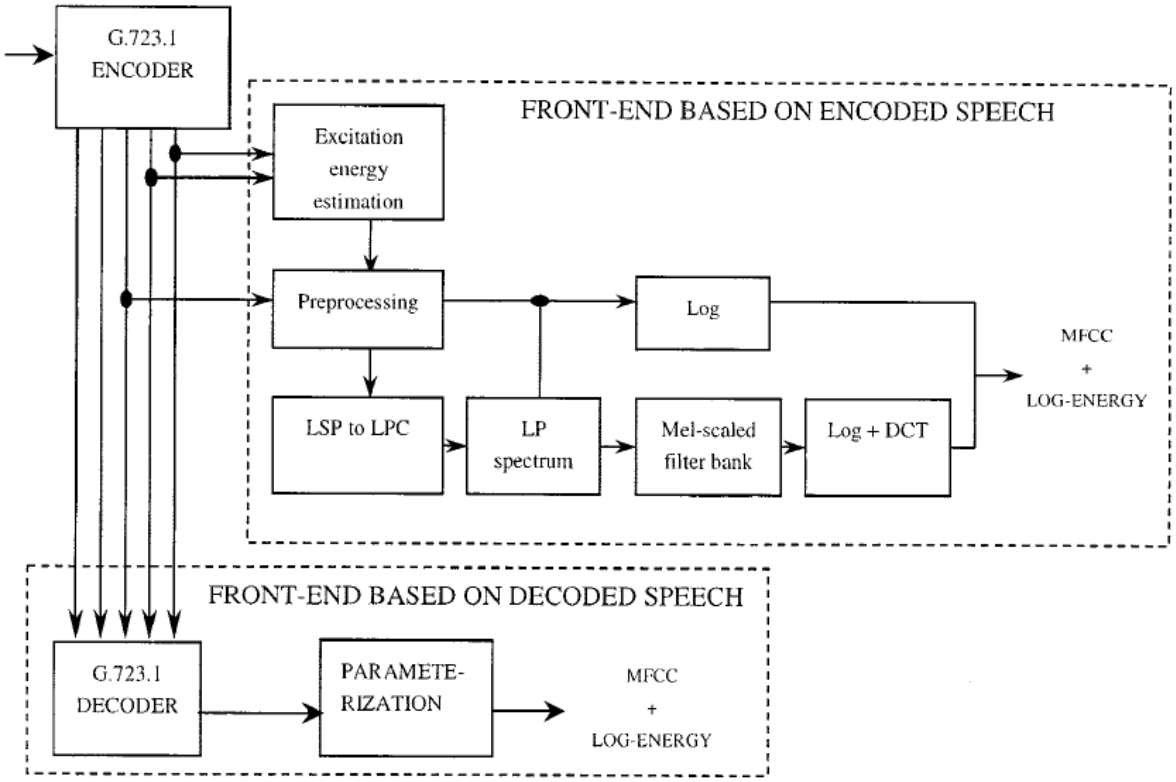


*Figure2.10: la procédure de la paramétrisation, en bas l'approche classique, le block en haut présente la nouvelle procédure proposée*

Il est clair qu'avec l'explosion d'internet et de la téléphonie mobile, les moyens de communication ont totalement changé et appellent de nouveaux moyens pour la sécurisation des accès et des échanges de données, c'est sur cet aspect que s'est focalisé l'article [4] qui propose une nouvelle méthodologie pour évaluer les performances de la vérification de

locuteur qui peut être affecté par la transmission des données via l'internet (perte de paquets par exemple), en se basant sur la base de donnée XM2VTS qui est considérée comme la norme dans la communauté biométrique audio et visuelle de vérification multimodale (parole et image), ce travail s'est effectué dans le cadre de l'action européenne COST-275.

Dans d'approche classique de l'identification de locuteur par la voix sur IP, après la numérisation de la voix, elle doit être compressée pour l'insérer dans les paquets IP, à la réception, un processus de décompression est nécessaire pour restituer l'information et la transformer en signal sonore (figure 2.11), ce processus de compression et décompression peut générer des problèmes et des limitations en termes de ressources processeur ou mémoire, influencer le débit de flux après décompression ou la taille de fichier résultant, un temps de latence très élevé. Pour remédier à ces problèmes, une nouvelle approche a été proposée [14] qui fait la reconnaissance de locuteur par voix IP en direct, en utilisant une méthode de clustering pour rassembler les vecteurs caractéristiques similaires appelée micro-clustering, le taux de précision de ce nouveau système est de 80%, et il est trois fois plus rapide que l'approche classique basée sur la modélisation des GMM.

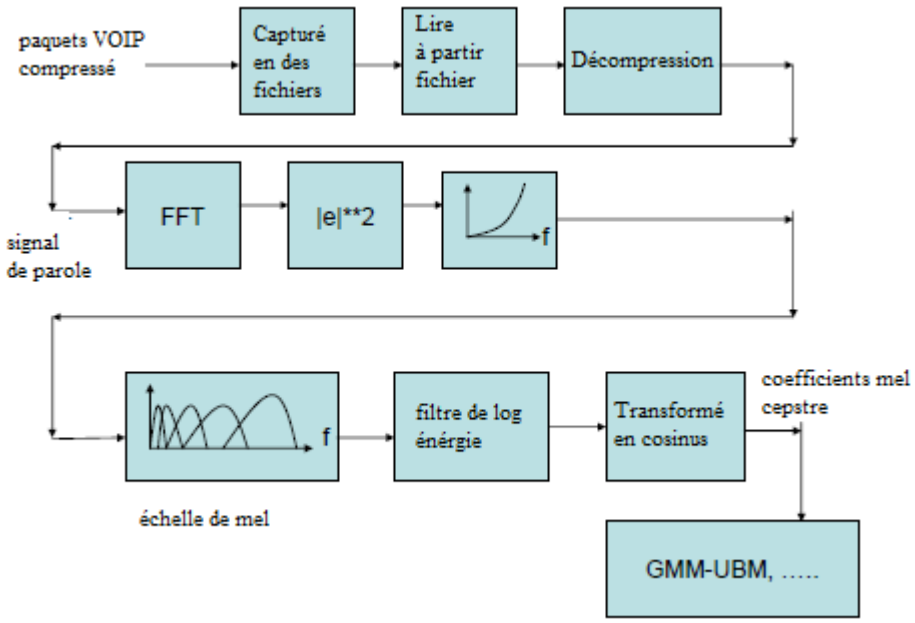


*Figure 2.11: Système d'identification de locuteur classique*

Un réseau VOIP est une cible potentielle de diverses menaces. Des attaques de déni de services peuvent empêcher les utilisateurs à communiquer et entraîner des pertes économiques considérables. L'écoute clandestine d'une communication entre deux personnes est facilement réalisable que dans des services traditionnels (e-mail, web). Pour cela, il est nécessaire de développer un mécanisme pour sécuriser le réseau VOIP, ainsi que l'audit de la voip (figure 2.12).

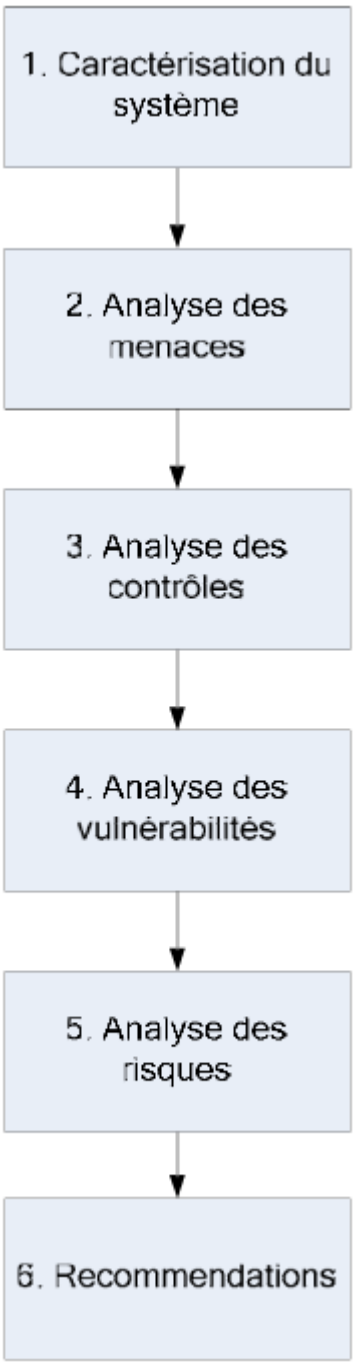


*Figure 2.12: procédure générale de l'audit*

L'article [29] proposé en 2007 présente une approche qui combine entre l'audit de la voip et la reconnaissance de locuteur, elle est basée sur une paramétrisation MFCC, avec combinaison de la DTW avec VQ comme méthode de modélisation, la performance du système obtenue est réduite à cause de la perte de paquets et de la qualité de l'information.

En 2008, l'article [9] présente un nouveau système de reconnaissance automatique de locuteur dans le domaine compressé basé sur un algorithme PSH (Probabilistic Stochastic Histogram), afin de tester l'efficacité de cet algorithme, ils l'ont appliqué sur deux techniques de modélisation (figure 2.13)

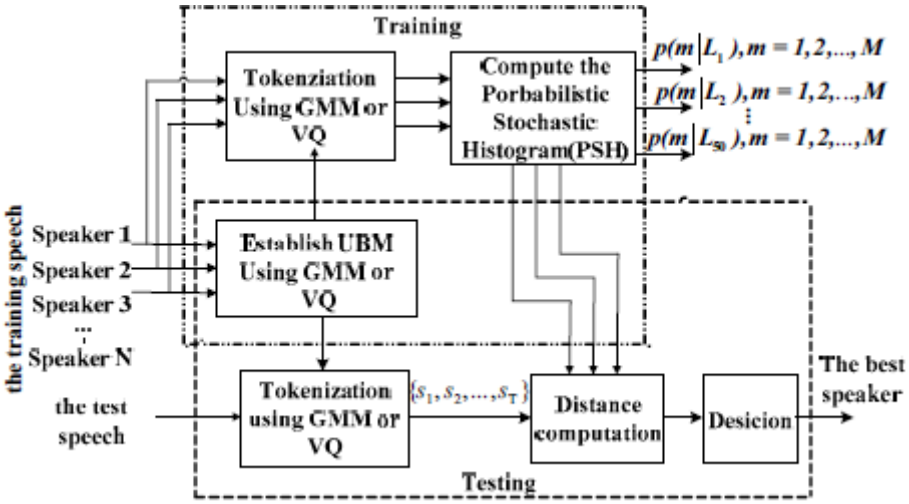


*Figure 43: reconnaissance de locuteur basé sur PSH*

après l'extraction des vecteurs caractéristiques en utilisant différents codecs *G.729, G.723.1 6.3k, et G723.1 5.3k,* la première technique est fondée sur la quantification vectorielle utilisant l'algorithme PSH, et la deuxième technique est l'application d'une modélisation GMM basée sur PSH, l'expérience effectuée a montré que ces deux nouvelles approches basée sur PSH donnent une bonne discrimination, pour valider l'efficacité de nouveau algorithme, ils l'ont comparé avec l'approche GMM classique (figure 2.14).

| Bit Stream | Recognition rate for Classical GMM ( % ) | | Recognition Rate for GMMPSH (%) | |
|---|---|---|---|---|
| | 15-dimension( G723.1) or 11-dimension( G729) | 7-dimension | 15-dimension (G723.1) or 11-dimension( G729) | 7-dimension |
| G723.1 6.3k | 76.64 | 70.57 | 87.74 | 79.25 |
| G723.1 5.3k | 76.28 | 70.12 | 89.62 | 82.55 |
| G729 | 19.26 | 23.92 | 78.77 | 84.43 |

*Figure 2.14: Comparaison entre GMM et GMMPSH*

Un article publié en 2010 [1], s'intéresse au problème d'authentification et la détection de l'ID de l'appelant en se basant sur sa signature vocale dans un appel voip, la solution proposée c'est le développement d'un système de reconnaissance de locuteur avec MATLAB qui fait l'extraction des paramètres par la méthode LPC, avec une modélisation basé sur le principe de réseaux de neurones, et puis utiliser ce système pour développer un simple prototype basé sur le protocole SIP, qui a pour mission essentiel la gestion des appels entrants et sortants dans le réseaux, cette solution va remédier surtout aux attaques ID spoofing.

La communication par la voix sur un réseau non sécurisé n'offre aucune garantie de confidentialité ou d'intégrité, les informations échangées peuvent être facilement interceptées, pour cela il est nécessaire de crypter la voix pour éviter toute écoute indésirable sur le réseau, l'approche proposée dans l'article [2] vise à dévoiler l'identité de locuteur qui participe à la communication, cette approche exploite le concept de la détection de l'activité vocal (VAD), les résultats ont montré que même en présence de cryptage de l'information, uniquement 48% des cas ont été correctement identifiés.

En 2012, une étude a été faite sur l'effet de la transmission de données par la voix IP sur les performances d'un système de vérification automatique de locuteur en mode indépendant de texte [18], le système est basé sur une modélisation EM-MAP-GMM, et testé avec quatre variant codeur du standard H.323.

Une autre étude menée en 2012 pour mesurer également les performances d'un système de vérification automatique de locuteur, en utilisant une classification GMM-UBM et sous deux conditions différentes, la première est de le tester pour un codec large bande, et la deuxième condition est de le tester pour un codec de transmission à bande étroite [27]. Le système de vérification de locuteur basé sur un codec de transmission à bande large est plus performant que l'utilisation du codec à bande étroite.

## 2.4 Connexions internationales

Le sujet de la reconnaissance automatique de locuteur basée sur la voix IP est à la fois intrinsèquement lié à la recherche scientifique et un enjeu majeur pour les développements technologiques, c'est cependant une problématique complexe qui requiert des efforts importants. En effets, le développement comme l'évaluation de nouvelles techniques nécessitent des ressources importantes dont la production est difficilement accessible aux laboratoires. De plus, les métriques d'évaluation doivent prendre en compte tous les aspects de la tâche à évaluer tout en restant simples et homogènes pour ne pas nuire à la comparaison des résultats, dans ce qui suit on va présenter une liste non exhaustive des laboratoires et centre de recherches opérant dans ce domaine :

Aux Etats-Unis d'Amérique, des compagnes annuelles NIST [*w6*] spécialisées dans la reconnaissance automatique de locuteur ont fortement contribué à dynamiser la recherche dans ce domaine, il y a aussi plusieurs laboratoires et organismes qui ont participé à l'évolution de ce domaine, on cite par exemple: MIT-LL, qui a travaillé sur l'identification de locuteur, et surtout sur la compensation des distorsions que peut générer l'encodage de la parole, ce travail a été évalué sur le corpus TIMIT, et dans le cadre de l'identification de locuteur, et en utilisant la méthode LPC les résultats obtenus montrent la méthode utilisée a diminué le taux d'erreur et c'est la méthode la mieux placé pour compenser les distorsion de l'encodage de parole *(les activités et les publications de ce laboratoire sont plus détaillées dans [w1])*

En France, il y a le laboratoire d'informatique LIA du centre d'enseignement et de recherche en informatique associé à l'école doctorale Argosciences et Sciences de l'université d'Avignon et des Pays de Vaucluse,il fait partie du laboratoire d'excellence : Brain and Language Research Institute et participe à la fédération de recherche Agorantic. Il a été classé A+ à l'issue de sa dernière évaluation par l'AERES en 2012, Différents aspects du traitement automatique de la parole sont étudiés au LIA : reconnaissance de la parole, du locuteur, des langues, indexation audio, caractérisation de pathologies de la voix... Ces recherches s'appuient sur un environnement logiciel développé au LIA et librement distribuable sous licence GPL ou LGPL, notamment :

- SPEERAL : Moteur de reconnaissance automatique de la parole continue grand vocabulaire, basé sur un algorithme A*. Ce système est sous licence LGPL.
- MISTRAL : Plateforme open source d'authentification biométrique (ANR)

Le laboratoire participe à des campagnes d'évaluation internationales sur ces domaines et participe - ou porte - de nombreux projets collaboratifs nationaux et européens (ANR en cours : Sumacc, OT-media, Decoda, PI, ASH, DesphoApady, RPM2, Avison, PERCOL, ...) *(Pour plus de détails : [w2])*. Parmi les différents travaux effectué, on cite par exemple : dans le cas de la vérification de locuteur, ils ont proposé une méthodologie pour améliorer la dégradation des performances d'un système de vérification automatique du locuteur causée par la transmission des données par la voix IP, ils se sont basé pour cela sur une base de données multimodales X2MVTS et en utilisant deux codecs G723.1 et G711.

L'Algérie a également mené un ensemble de travaux dans ce contexte par le laboratoire Speech communication and Signal processing [5], ce laboratoire a effectué plusieurs recherches dans différentes disciplines, surtout dans l'authentification biométrique, (vous pouvez les consulter en se référant à [*w3*]) il y'avait une variété de sujets proposés dans le contexte de la reconnaissance automatique de locuteur par la voix IP, un sujet a été proposé en master, sur l'influence des pertes de paquets sur les performances d'un système de reconnaissance automatique de locuteur pour les transmissions basées sur la voip, un autre travail a été effectué dans le cadre du doctorat, sur l'influence du codec G729 sur le système de reconnaissance automatique de locuteur par la voix IP,

Ainsi la Chine contribue au développement des recherches dans ce domaine grâce au centre de recherche Intelligent Computing Research Center de l'Institut Harbin de Technologie, parmi les travaux effectués, dans le domaine criminalistique, ils ont proposé d'hybrider entre l'audit de la VOIP et la reconnaissance de locuteur, ils ont constaté une dégradation au niveau des performances du système, cela est expliqué par le problème de perte de paquets [*w4*].

## Conclusion

Après avoir présenté les différentes phases qui constituent un système de reconnaissance automatique de locuteur, en commençant par l'extraction des paramètres, la modélisation et finalement la décision, nous avons passé à la présentation d'un aperçu sur la théorie de la voip, et finalement, nous avons présenté brièvement, quelques travaux effectués dans le cade de la reconnaissance automatique du locuteur en utilisant la voix IP, ainsi que les laboratoires qui opèrent dans ce secteur.

# Chapitre 3 : Expériences et résultats

# 3.1 Introduction

Pour développer notre système d'identification automatique de locuteur, nous allons passer par trois phases : la paramétrisation, la modélisation et finalement l'étape de la décision. Nous allons commencer par la définition des six variantes qui découlent de la méthode MFCC pour l'extraction des paramètres, en décrivant après le protocole expérimental que nous avons utilisé, passant à l'analyse et lecture des résultats obtenus et terminant par l'implémentation d'une interface graphique en Matlab basée sur la meilleure configuration obtenue dans les tests

# 3.2 La paramétrisation

Après l'introduction de la méthode MFCC pour l'extraction des paramètres, plusieurs variantes et améliorations de la méthode d'origine ont été proposé, elles différent dans le nombre de filtres nécessaires, la forme de ces filtres, la manière dans laquelle ces filtres sont espacés, et la bande passante utilisée.

Cette diversité d'implémentation est due à l'intérêt de suivre le progrès dans le domaine de la psycho-acoustique, par exemple, il existe diverses approximations de la perception non linéaire du pitch du système auditoire humain, l'approximation de koenig [13] est de deux types soit linéaire au-dessous de 1000 Hz, soit logarithmique au-delà, cette approximation a l'avantage de fournir une représentation peu couteuse des calculs, mais elle ne donne pas une meilleur précision, pour cela, une autre approximation plus précise proposée en 1949 par Fant :

$$\hat{f}_{mel} = k_{const} \,.\, log_n(1 + \frac{f_{lin}}{F_b}) \qquad (4.1)$$

Dans le cas où $F_b = 1000$, cette relation devient : $\hat{f}_{mel} = \frac{1000}{log_n 2} \,.\, log_n(1 + \frac{f_{lin}}{1000})$ (4.2)

Cette approximation, en la comparant avec celle de koenig, fournit une représentation très proche de l'échelle de mel, elle est plus convenable, vu que les valeurs de $\hat{f}_{mel}$ ne sont pas affecté par le choix de la nature du logarithme. Autres approximations qui dérivent de l'équation (4.1) et qui varient entre le choix de la nature du logarithme ont conduit à ces deux représentations :

$$\hat{f}_{mel} = 2595 \,.\, log_{10}(1 + \frac{f_{lin}}{700}) \qquad (4.3)$$

et $\hat{f}_{mel} = 1127 \,.\, log_e(1 + \frac{f_{lin}}{700})$ ou $\hat{f}_{mel} = 1127 \,.\, Ln(1 + \frac{f_{lin}}{700})$ (4.4)

qui sont largement utilisé dans les différentes implémentations de MFCC.

Dans ce qui suit, nous allons présenter les différentes implémentations les plus populaires de la MFCC : MFCC_FB20, HTK_MFCC_FB24, HTK_MFCC_EB26, MFCC_FB40, HFCC_E_FB29, Skowronski_MFCC_FB20.

### *3.2.1 MFCC_FB20*

C'est un paradigme introduit par Davis et Mermelstein en 1980 [13], il est issu d'une transformation de fourier inverse appliqué au logarithme du spectre de puissance.

Les MFCC par Davis et Mermelstein s'obtiennent en considérant, pour le calcul du cepstre, la représentation fréquentielle selon l'échelle perceptive de mel exprimé par l'expression (4.3), avec f la fréquence en valeurs linéaires.

Pour faire l'extraction des paramètres, ils ont utilisé un banc de filtres triangulaires de hauteurs égales appliqué à une transformée de fourier discrète d'une suite de N termes du signal d'entrée.
La suite de N termes X(0) …. X(N-1) est définis par :

$$X(k) = \sum_{n=0}^{N-1} x(n)\exp(\frac{-j2\pi nk}{N}) \qquad (4.5)$$

Chaque filtre est définit par :

$$H_i(k) = \begin{cases} 0 & k < f_{b_{i-1}} \; ou \; k > f_{b_{i+1}} \\ \frac{(k - f_{b_i})}{(f_{b_i} - f_{b_{i-1}})} & f_{b_{i-1}} \leq k \leq f_{b_i} \\ \frac{(f_{b_{i+1}} - k)}{(f_{b_{i+1}} - f_{b_i})} & f_{b_i} \leq k \leq f_{b_{i+1}} \end{cases} \qquad (4.6)$$

Avec i présente le i[ème] filtre, $f_{b_i}$ présente les points limites des filtres, k correspond au k[ème] coefficient des N-termes de la suite DFT.

Le banc de filtres utilisé par cette variante contient vingt filtres d'une hauteur égale (figure 4.1) qui couvre l'intervalle [0 4600] Hz, les fréquences centrales des dix premiers filtres sont espacées de façon linéaire entre 100 et 1000Hz, et les fréquences centrales des dix derniers filtres sont espacés de façon logarithmique entre 1000 et 4000 Hz, les MFCCs s'obtiennent alors par une transformée en cosinus inverse :

$$C_j = \sum_{i=1}^{M} X_i . \cos(j . \left(i - \frac{1}{2}\right) . \frac{\pi}{M}$$

avec j = 1…M le nombre des coefficients cepstrals

$$et \; X_i = Log(\sum_{k=0}^{N-1} |X(k)| \, . H_i(k))$$

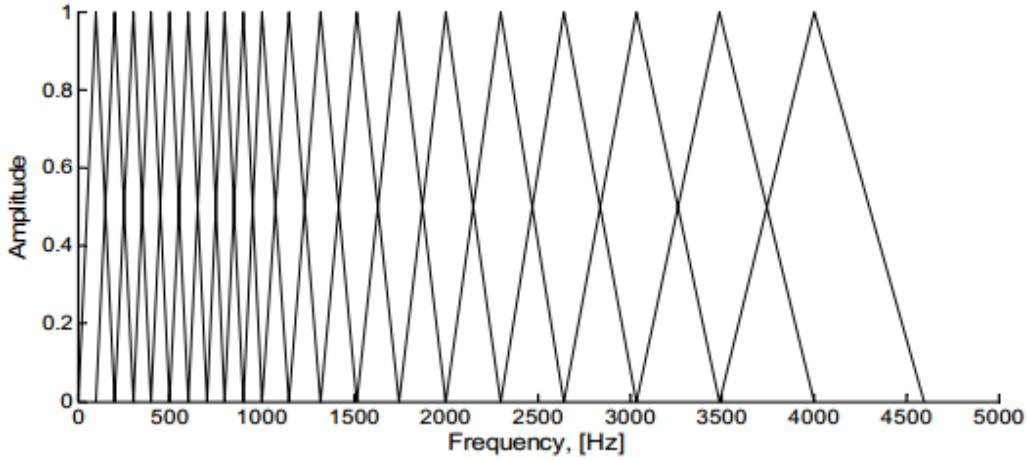


*Figure 3.1 : banc de filtre de la variante MFCC_FB20*

### *3.2.2 DavisSkowronski_MFCC_FB20*

C'est une variante de la paramètrisation MFCC proposé en 2004 [13], elle utilise un banc de 20 filtres, étalé sur une gamme de fréquence de [0 8000] Hz, c'est la même variante Davis_MFCC_FB20, sauf qu'on 2004 Skowronski a proposé une méthode pour calculer les fréquences centrales de chaque filtre, le choix de la fréquence centrale $f_{c_i}$ de l'i$^{ème}$ filtre peut être approximé par :

$$f_{c_i} = \begin{cases} 100.i & i = 1 \dots 10 \\ f_{c_{10}}.2^{0.2(i-10)} & i = 11 \dots..20 \end{cases} \qquad (4.7)$$

### *3.2.3 HTK_MFCC_FB24*

C'est une autre mise en œuvre de la MFCC qui est largement utilisée, elle a été créé dans la plateforme HMM du HTK qui est originalement mise au point à l'université de cambridge, la désignation HTK_MFCC_FB24 reflète le nombre de filtres (24) recommandé par HTK, pour une bande passante de 8000 Hz du signal, l'implémentation de cette variante est similaire à l'originale approche de Davis et Mermelstein.

Cette variante utilise aussi la représentation fréquentielle selon l'échelle de mel donné dans l'équation (4.3), les limites de la gamme de fréquences sont les paramètres qui définissent la base de la conception du banc de filtres. Pour cela, il est nécessaire de déterminer d'abord les fréquences minimales et maximales qui limitent la gamme de fréquence utilisée, après on calcul un facteur $\Delta\hat{f}$ qui sert à ajuster la définition des fréquences centrales de chaque filtre, il est exprimé par : $\Delta\hat{f} = \frac{\hat{f}_{high} - \hat{f}_{low}}{M+1}$ avec M le nombre des filtres, la fréquence centrale individuel est donnée en mel par l'équation : $\hat{f}_{c_i} = \hat{f}_{low} + i.\Delta\hat{f}$, La fréquence centrale de chacun des filtres est alors transformé et exprimée en Hz :

$$f_{c_i} = 700\left(10^{\frac{\hat{f}_{c_i}}{2595}} - 1\right)$$

Avec $\hat{f}_{c_i}$ la fréquence centrale du i[ème] filtre.

Les paramètres HTK_MFCC_FB24 sont calculés comme suit: la DFT de la suite X(k) calculé pour un signal d'entrée x(n) exprimé en (4.4) est utilisé pour calculer le spectre de puissance $|X(k)|^2$ qui réagit comme une entrée du banc de filtre $H_i(k)$, après, le banc de filtre de sortie est donné par : $X_i = Ln(\sum_{k=0}^{N-1}|X(k)|^2 . H_i(k))$, et les paramètres HTK_MFCC_FB24 sont extraite par l'application de la transformé en cosinus inverse DCT.

### *3.2.4 HTK_MFCC_FB26*

Cette variante est une version de la méthode HTK_MFCC_FB24 [13], elle utilise 26 filtres pour une gamme de fréquence de [0 8000] Hz (figure 4.2), cette variante présente la version la plus récente de HTK selon Young en 2006.

| Filter no. | Lower frequency [Hz] | Higher frequency [Hz] | Center frequency [Hz] | Filter bandwidth [Hz] |
|---|---|---|---|---|
| 1 | 0 | 144 | 68 | 72 |
| 2 | 68 | 226 | 144 | 79 |
| 3 | 144 | 317 | 226 | 87 |
| 4 | 226 | 416 | 317 | 95 |
| 5 | 317 | 525 | 416 | 104 |
| 6 | 416 | 645 | 525 | 115 |
| 7 | 525 | 777 | 645 | 126 |
| 8 | 645 | 921 | 777 | 138 |
| 9 | 777 | 1080 | 921 | 152 |
| 10 | 921 | 1254 | 1080 | 167 |
| 11 | 1080 | 1445 | 1254 | 183 |
| 12 | 1254 | 1655 | 1445 | 201 |
| 13 | 1445 | 1886 | 1655 | 221 |
| 14 | 1655 | 2139 | 1886 | 242 |
| 15 | 1886 | 2416 | 2139 | 265 |
| 16 | 2139 | 2721 | 2416 | 291 |
| 17 | 2416 | 3056 | 2721 | 320 |
| 18 | 2721 | 3423 | 3056 | 351 |
| 19 | 3056 | 3827 | 3423 | 386 |
| 20 | 3423 | 4270 | 3827 | 424 |
| 21 | 3827 | 4756 | 4270 | 465 |
| 22 | 4270 | 5289 | 4756 | 510 |
| 23 | 4756 | 5875 | 5289 | 560 |
| 24 | 5289 | 6519 | 5875 | 615 |
| 25 | 5875 | 7225 | 6519 | 675 |
| 26 | 6519 | 8000 | 7225 | 741 |

*Figure 3.2: Le banc de filtre de la variante HTK_FB26*

### *3.2.5 MFCC_FB40*

Dans la bibliothèque Auditory Toolbox développée par Malcolm Slaney en 1998 [13], les paramètres MFCCs sont calculés par un banc de 40 filtres. En supposant que la fréquence d'échantillonnage est de 16000 Hz, Slaney a mis en place un banc de 40 filtres qui couvrent l'intervalle de fréquences [133,3 6855] Hz, les fréquences centrales $\hat{f}_{c_i}$ des 13 premiers filtres sont espacés de façon linéaire dans l'intervalle [200 1000] Hz, avec un pas de 66,67 Hz, or les 27 derniers filtres sont espacés de façon logarithmique dans [1071 6400] Hz avec un pas d'une taille de 1,0711703 calculé par :

logStep = $\exp\left(\frac{\ln\left(\frac{f_{c_{40}}}{1000}\right)}{numLogFilt}\right)$ avec $f_{c_{40}}$ = 6400 Hz est la fréquence centrale du derniers filtres espacés de façon logarithmique, et numLogFilt = 27 est le nombre totale de ces derniers.

En se référant à la représentation des filtres (6), chaque filtre dans ce cas est définit par :

$$H_i(k) = \begin{cases} 0 & f_{b_{i-1}} \; ou \; k > f_{b_{i+1}} \\ \frac{2(k - f_{b_i})}{(f_{b_i} - f_{b_{i-1}})} & f_{b_{i-1}} \leq k \leq f_{b_i} \\ \frac{2(f_{b_{i+1}} - k)}{(f_{b_{i+1}} - f_{b_i})} & f_{b_i} \leq k \leq f_{b_{i+1}} \end{cases} \quad (4.8)$$

Les points limites $f_{b_i}$ sont exprimés en termes de leurs positions, comme indiqué ci-dessus. Le banc de filtre exprimé en (4.8) est normalisé d'une façon tel que la somme des coefficients de chacun des filtres soit égale à un : $\sum_{k=1}^{N} H_i(k) = 1$ avec i = 1….M (le nombre des filtres).

Il sera utilisé par la suite pour le calcul du spectre d'énergie, les MFCCs s'obtiennent alors par une transformée en cosinus inverse.

Dans le cas où la fréquence d'échantillonnage est de 8000 Hz, seulement les filtres inférieurs à 4000 Hz sont considérés. La figure 4.3 illustre 32 filtres qui couvrent l'intervalle [133 3954] Hz, selon la représentation d'origine de Slaney, les fréquences centrales des 13 premiers filtres sont espacés de façon linéaire sur [200 1000] Hz et les 19 filtres suivant dont espacés de manière logarithmique sur [1071 3692] Hz avec un pas de 1.0711703, cette version de la paramétrisation MFCC de Slaney dans le cas où la fréquence d'échantillonnage est 8000 Hz est nommé par MFCC FB32.

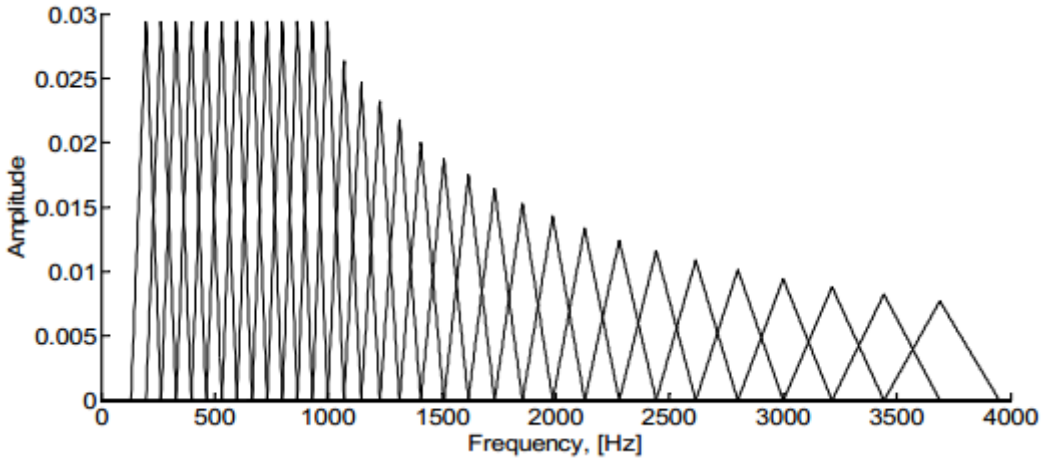


*Figure 3.3 : Banc de 32 filtres*

### *3.2.6 HFCC_E_FB29*

La variante HFCC introduite par Skowronski et Harris en 2004 [13], représente l'implémentation de bande mel la plus récente, cette variante n'est pas censé être un modèle perceptif du système auditoire humain comme pour toutes les versions de la paramétrisation MFCC.

Supposant la fréquence d'échantillonnage de 12500 Hz, Skowronski et Harris ont proposé cette mise en œuvre du banc de filtres HFCC composé de 29 filtres, et qui couvre la gamme de fréquence [0 6250] Hz, comme montre (la figure 4.4), les filtres HFCC se chevauchent de façon différente que l'approche traditionnelle, un filtre peut se chevaucher non seulement avec les filtres adjacents, mais avec les plus distants aussi

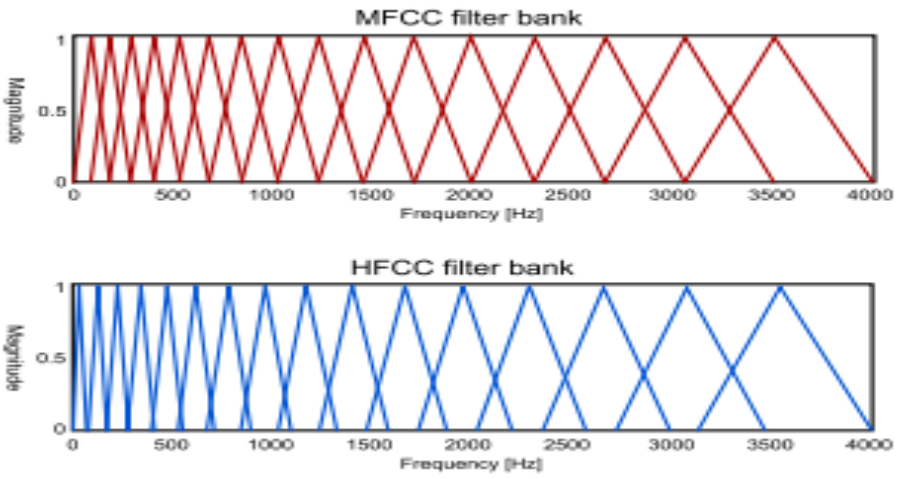


*Figure3.4 : Comparaison entre le banc de filtres des MFCCs et des HFCCs*

La différence entre cette variante et les MFCCs réside dans le fait que les HFCCs sont basé sur une mesure de la largeur des filtres appelée « Largeur de bande réctangulaire équivalente, ou Equivalent Rectangular Bandwidth(ERB) » proposé par Moore et Glasberg en 1983, pour chaque filtre, la valeur ERB est définie comme la largeur d'un filtre passe bande idéal de même fréquence centrale, la mesure de ces ERB illustre la résolution fréquentielle du système auditif, elle est donnée par la formule suivante :

$$ERB = 6.23.10^{-6}\,.f_c^2 + 93.39\,.10^{-3}.f_c + 28.52 \qquad (4.9)$$

Avec $f_c$ la fréquence centrale du filtre exprimé en Hz. Le filtre passe bande calculé en (4.9) est pondéré par une constante appelé (selon Skowronski et Harris) E-factor.

La conception du banc de filtre de HFCC est décrite comme suit, d'abord on choisit le nombre de filtres M ainsi que la fréquence minimale $f_{low}$ et maximale $f_{high}$ du banc de filtre tout en entier, les fréquences centrales $f_{c_1}$ et $f_{c_M}$ sont calculées comme suit :

$$f_{c_i} = \frac{1}{2}.(-\bar{b} + \sqrt{\bar{b}^2 - 4.\bar{c}}) \qquad (4.10)$$

Avec i l'indice de la fréquence centrale 1 ou M, les coefficients $\bar{b}$ et $\bar{c}$ sont définis par :

$\bar{b} = \frac{b-\hat{b}}{a-\hat{a}}$ (4.11) et $\bar{c} = \frac{c-\hat{c}}{a-\hat{a}}$ (4.12)

Les valeurs des constantes a, b, et c sont ceux exprimé en (4.9) : $6.23.10^{-6}$, $93.39\,.10^{-3}$ et $28.52$ respectivement et elles varient dans les deux cas, pour le premier filtre, les coefficients $\hat{a}$, $\hat{b}$, et $\hat{c}$ sont calculé comme suit :

$$\hat{a} = \frac{1}{2}.\frac{1}{700+f_{low}} \qquad (4.13)$$

$$\hat{b} = \frac{700}{700+f_{low}} \qquad (4.14)$$

$$\hat{c} = -\frac{f_{low}}{2}.\left(1 + \frac{1}{700+f_{low}}\right) \qquad (4.15)$$

Pour le dernier filtre, ces coefficients sont donnée par :

$$\hat{a} = -\frac{1}{2}.\left(\frac{1}{700+f_{high}}\right) \qquad (4.16)$$

$$\hat{b} = -\left(\frac{700}{700+f_{high}}\right) \qquad (4.17)$$

$$\hat{c} = \frac{f_{high}}{2}.\left(1 + \frac{700}{700+f_{high}}\right) \qquad (4.18)$$

Une fois les fréquences centrales du premier et dernier filtre sont calculé, la génération des fréquences centrales des filtres au milieu est facile parce qu'elles sont équidistant sur l'échelle de mel, le pas $\Delta\hat{f}$ entre les fréquences centrales des filtres adjacents est calculé par $\Delta\hat{f} = \frac{\hat{f}_{c_M} - \hat{f}_{c_1}}{M-1}$

Le passage de $f_{c_1} \rightarrow \hat{f}_{c_1}$ et $f_{c_M} \rightarrow \hat{f}_{c_M}$ est donnée par la formule (4.3), les fréquences centres des filtres adjacents est calculé donc par : $\hat{f}_{c_i} = \hat{f}_{c_1} + (i-1).\Delta\hat{f}$ pour i = 2….M-1
Finalement, les fréquences maximales et minimales de chaque filtre i sont exprimé par :

$$f_{low_i} = -(700 + ERB_i) + \sqrt{(700 + ERB_i)^2 + f_{c_i}\left(f_{c_i} + 1400\right)}$$

Et $f_{high_i} = f_{low_i} + 2.ERB_i$ Avec $ERB_i = \frac{1}{2}.(f_{high_i} - f_{low_i})$

Après le calcul de ces paramètres, la conception du filtre HFCC est complète, l'étape suivante c'est le calcul de la transformé en cosinus inverse pour extraire les paramètres HFCC.

## 3.3 Le protocole expérimental

Notre but dans ce chapitre est de développer un système d'identification automatique de locuteur par la voix IP, et de le valider sous l'environnement Matlab. Pour cela nous allons suivre les étapes suivantes pour la construction et la validation du système : on utilisera tout d'abord une base de données (ou corpus) utile à l'apprentissage du système et ensuite à son évaluation, après une étape de la décomposition de parole/non parole sera nécessaire afin de ne conserver que les zones de paroles correspondants aux locuteurs, puis on passera à l'étape de l'extraction des paramètres par différentes variantes de la méthode MFCC, ensuite on fera l'apprentissage des modèles de mélanges de lois gaussiennes, finalement on passera à l'étape de la reconnaissance, décider si le locuteur testé est le locuteur cible.

### *3.3.1 Description de la base de données*

La base de données utilisée dans l'expérience est une base de données basée sur la voix IP construite au sein du laboratoire SIA de la FST, d'abord tous les enregistrements passent par l'intermédiaire du logiciel de communication Skype, nous avons fait l'expérience sur vingt locuteurs (15 hommes et 5 femmes), chaque locuteur possède 7 enregistrements, la durée de ces fichiers varie entre 3 à 26 secondes, nous avons réservé quatre enregistrements d'une durée d'une minute pour l'apprentissage, et un trois enregistrement de dix secondes pour le test, le choix de cette répartition s'est porté sur les locuteurs qui ont le plus de segment de test disponibles, les fichiers wav sont nommés de la façon suivante : $Loc_x\_fich_y.wav$ avec x l'id du locuteur et y le numéro de fichier.

### *3.3.2 Décomposition parole/non parole*

Chaque fichier son n'est pas uniquement composé de parole. En effet, des zones de silence sont présentes. Afin d'effectuer l'apprentissage des modèle, il est nécessaire d'isoler les zones de parole.

### *3.3.3 La phase de la paramétrisation*

Cette phase correspond à l'extraction des paramètres du signal, nous avons implémenté les six variantes de la paramétrisation MFCC : Davis_FB20, Davis_Skowronski_FB20, HTK_FB24, HTK_FB26, Slaney_FB40, Skowronski_HFCC_FB29, nous avons fixé le nombre de coefficients MFCC à 19, avec l'utilisation du fenêtrage de Hamming d'une taille de 25ms, un chevauchement (overlap) de 10 ms et un facteur de préaccentuation du signal de 0.95.

### *3.3.4 Apprentissage par GMM*

Dans cette étape, nous avons choisis d'utiliser la méthode GMM est justifiable par le fait qu'elle a fait ses preuves dans le domaine de la reconnaissance automatique de locuteur, nous avons testés pour un nombre de gaussiens qui varie de 16 à 1024.

Un modèle de mélange gaussien correspond à une densité de la forme :

$$\mathbf{p(x|y) = \sum_{k=1}^{K} p(k) \ . N(x, \mu_{yk}, \Sigma_{yk})}$$

Avec $\mathbf{N(x, \mu_{yk}, \Sigma_{yk})}$ représente la loi gaussienne de moyenne $\boldsymbol{\mu_{yk}}$ et de matrice de covariance $\boldsymbol{\Sigma_{yk}}$ et p(k) représente la probabilité a priori que x soit produite par la $k^{ieme}$ composante du mélange.

### *3.3.5 La phase de la décision*

Afin d'effectuer cette étape, il convient de calculer les coefficients MFCC pour le locuteur de test, après, comparer entre les coefficients MFCC extraite et les différents modèles GMM issues de la phase d'apprentissage et stockés dans la base de données (figure 4.5), le modèle qui fournit un meilleur score est bien le modèle de locuteur qui correspond au locuteur testé.

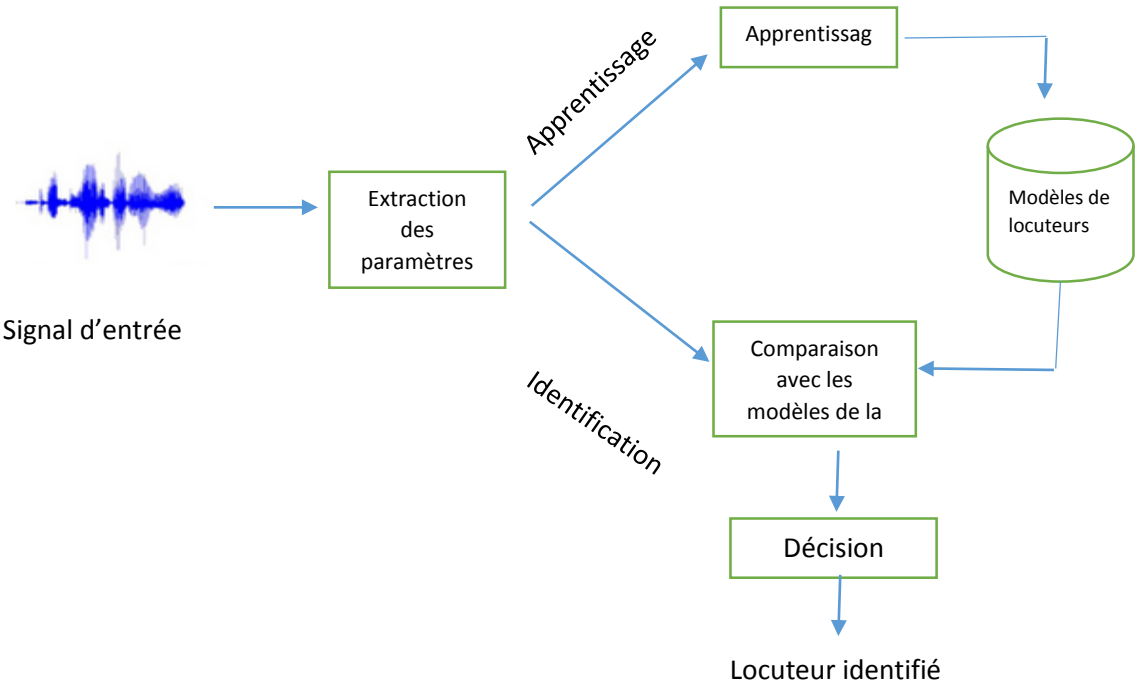


*Figure 3.5: Processus d'identification*

## 3.4 Résultats et tests

Avant de construire un système d'identification de locuteur par la voix IP, nous avons commencé par l'étape de la paramètrisation, nous avons testé les six variantes de la méthode MFCC citées précédemment, après nous avons utilisé une modélisation GMM des paramètres extraites en passant par 7 classes de mélange de gaussiennes (16, 32, 64, 128, 256, 512, et 1024). Pour en déduire la configuration qui donne les meilleurs résultats.

### *3.4.1 Identification du locuteur dans un milieu fermé*

Dans cette section, le résultat fourni par le système est de donner l'ID de locuteur qui a le score maximal. Le nombre de tests correctement identifié et mal identifié pour les six variantes et dans les différentes classes GMM sont représenté dans le graphe suivant (figure 3.6) :

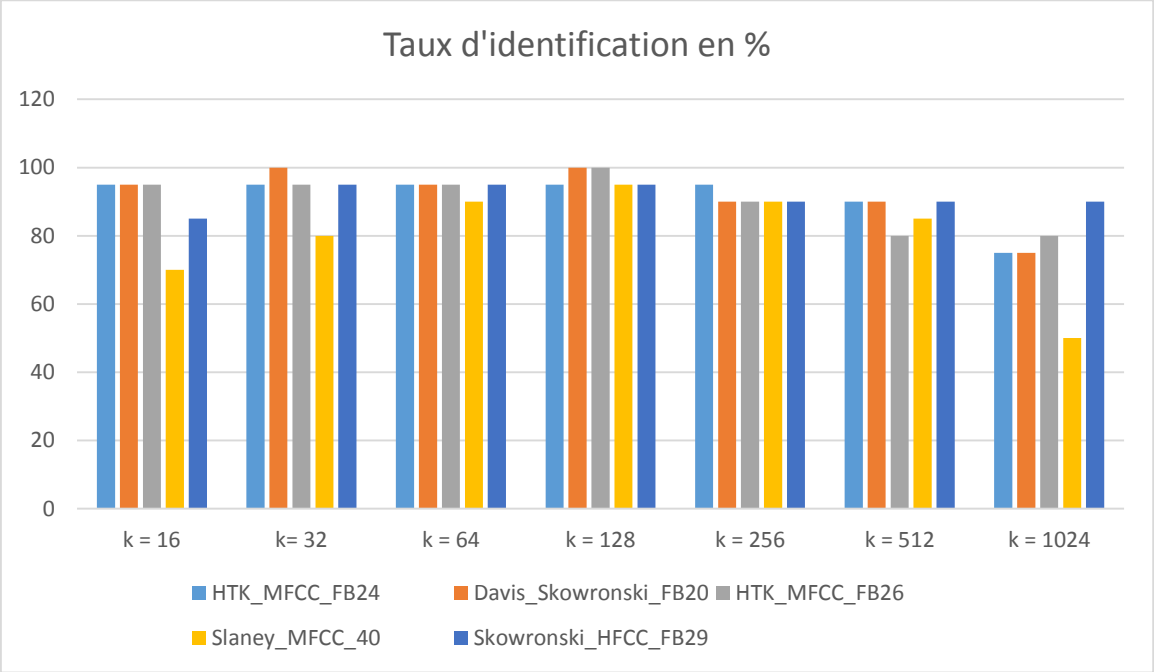


*Figure 3.6 : Taux d'identification en utilisant la base de données basé sur la VOIP*

### *3.4.2 Discussion des résultats obtenus*

L'implémentation des différentes variantes de la méthode MFCC ont été évalué en utilisant notre propre base de données de la voix IP, le même protocole expérimentale a été adapté dans toutes les expériences effectué dans le cadre de l'identification de locuteur en mode indépendant du texte, pour chacun des 20 locuteurs de la base de données, on a créé sept modèles GMM d'ordre différent (16, 32, 64, 128, 256, 512, 1024 gaussiennes). La durée des fichiers d'apprentissage est d'une seule minute, pour les fichiers de tests, la durée est de 10 secondes. Les modèles GMMs sont entraînés par 19 coefficients. Les vecteurs

MFCCs sont extraits en appliquant un fenêtrage de Hamming d'une longueur de 25ms, avec un facteur de chevauchement de 10 ms. Nous avons aussi procédé à l'élimination du silence.

A travers les tests effectués, nous avons vu les performances du système d'identification de locuteur ne varie pas considérablement pour les variantes qui utilisent un petit nombre de filtres, les approximations qui utilisent une large bande passante donnent un taux de reconnaissance élevé, comme on peut le remarqué pour la variante Davis_Skowronski_FB20, suivi de la variante HTK_MFCC 26 et 24 qui utilisent une bande passante de [0 8000] Hz, après vient pour la méthode Skowronski HFCC 29 qui couvrent l'intervalle [0 6250] Hz, et finalement vient la méthode Slaney_MFCC_FB40 qui utilise la bande de fréquences [133,3 6855] Hz, cette dégradation de taux de reconnaissance en minimisant la largeur de la bande passante peut être expliqué par le fait que la voix IP nécessite une large bande passante pour transmettre un grand flux de données, nous remarquons aussi que l'augmentation de l'ordre du modèle GMM influence la performance globale du système d'identification de locuteur.

La variante Davis_Skowronski_FB20 a donné un taux de reconnaissance de 100% dans le cas où le modèle GMM correspond à 32, et à 128 gaussiennes, or la variante HTK_MFCC_FB26 a donné un taux de reconnaissance de 100% dans le cas où le modèle GMM correspond à 128 gaussiennes, pour l'approximation HTK_MFCC_FB24, elle a donné un taux de reconnaissance de 95% pour les modèles GMM allant de 16 à 256 gaussiennes, au délà de cela, on remarque une dégradation de performances, pour la variante Slaney_MFCC_FB40, taux de reconnaissance a dégradé de 70% pour 16 gaussiennes jusqu'à 50% pour 1024 gaussiennes, pour la variante Davis_MFCC_20, le taux a dégradé de 90% pour 16 gaussiennes à 75% pour 1024 gaussiennes. Le choix du modèle GMM est très important, en effet si nous choisissons un grand ordre GMM, nous pouvons avoir le problème de sur-apprentissage du modèle GMM, càd présenter des données qui n'existe pas dans l'espace de paramètres acoustiques du locuteur en question.

D'après les expériences effectués, un modèle GMM composé de 32, 64 ou 128 gaussiennes est largement suffisant pour représenter la distribution des vecteurs acoustiques d'un locuteur.

En comparant les résultats obtenus par la voix IP avec des résultats qui ont été effectué au sein du laboratoire SIA sur une base de données appelée ELSDSR de 24 locuteurs basé sur des enregistrements par microphone, et qui ont été effectué dans le même protocole expérimentale que nous avons utilisé (figure 3.7), les performances du système d'identification de locuteur obtenues sont complétement différentes de celles obtenues en voip. D'après la figure 3.7, on constate que le taux de reconnaissance ne varie pas considérablement lorsque les différentes approximations de la perception de pitch du système auditoire humain ont été utilisées.

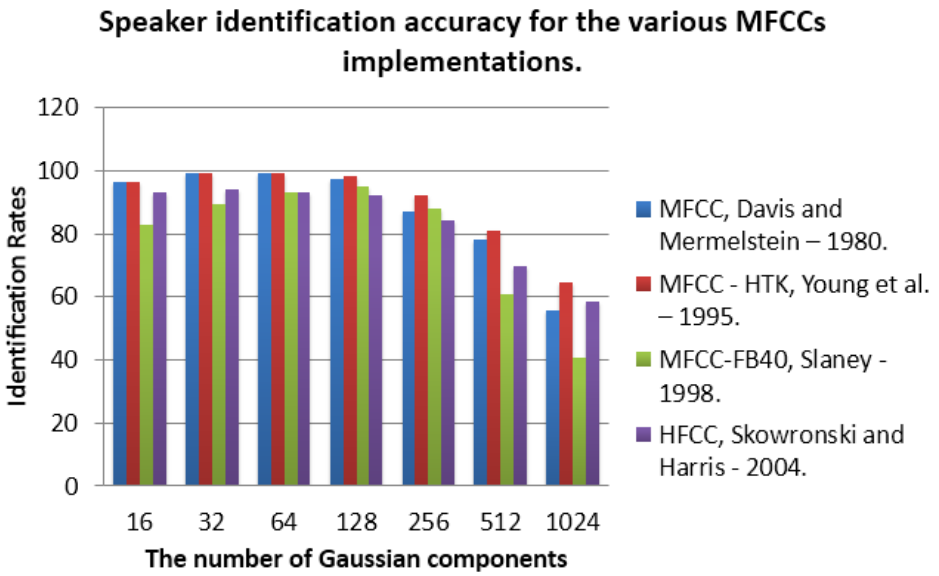


*Figure 3.7: Taux d'identification avec une base de données de la voix enregistré par le microphone*

Dans le test qui va suivre, nous allons utiliser la variante avis_Skowronski_MFCC_20 comme étant la meilleure méthode d'extraction des paramètres avec un modèle de 32 gaussiennes.

### *3.4.3 Identification de locuteur en milieu ouvert*

L'identification de locuteur en milieu ouvert est différente à celle du milieu fermé, le résultat retourné par le système n'est pas uniquement le locuteur qui a le score maximale de vraisemblance, l'identification de locuteur en milieu ouvert offre la possibilité de rejeter un locuteur qui n'est pas dans la base de références.

Pour ce faire, nous allons tester sur des locuteurs qui ne figurent pas dans la base de données, en utilisant la méthode Skowronski_MFCC_20 comme méthode d'extraction des paramètres, avec un modèle GMM de 32 gaussiennes.

Nous avons changé deux locuteurs dans la base de données, et puis nous avons fixé un seuil, qui en comparant le score avec ce seuil, s'il est supérieur alors le locuteur est bien identifié, sinon le système le rejette immédiatement, le taux de reconnaissance que nous avons obtenu a diminué vers 40%.

## 3.5 Implémentation de l'interface graphique

Pour présenter les résultats finals en utilisant la meilleure approximation trouvée dans la partie des tests, nous avons développés une interface graphique avec Matlab qui fait l'identification automatique de locuteur en mode indépendant du texte dans un groupe fermé et dans un groupe ouvert.

L'interface d'accueil est présentée comme suit :

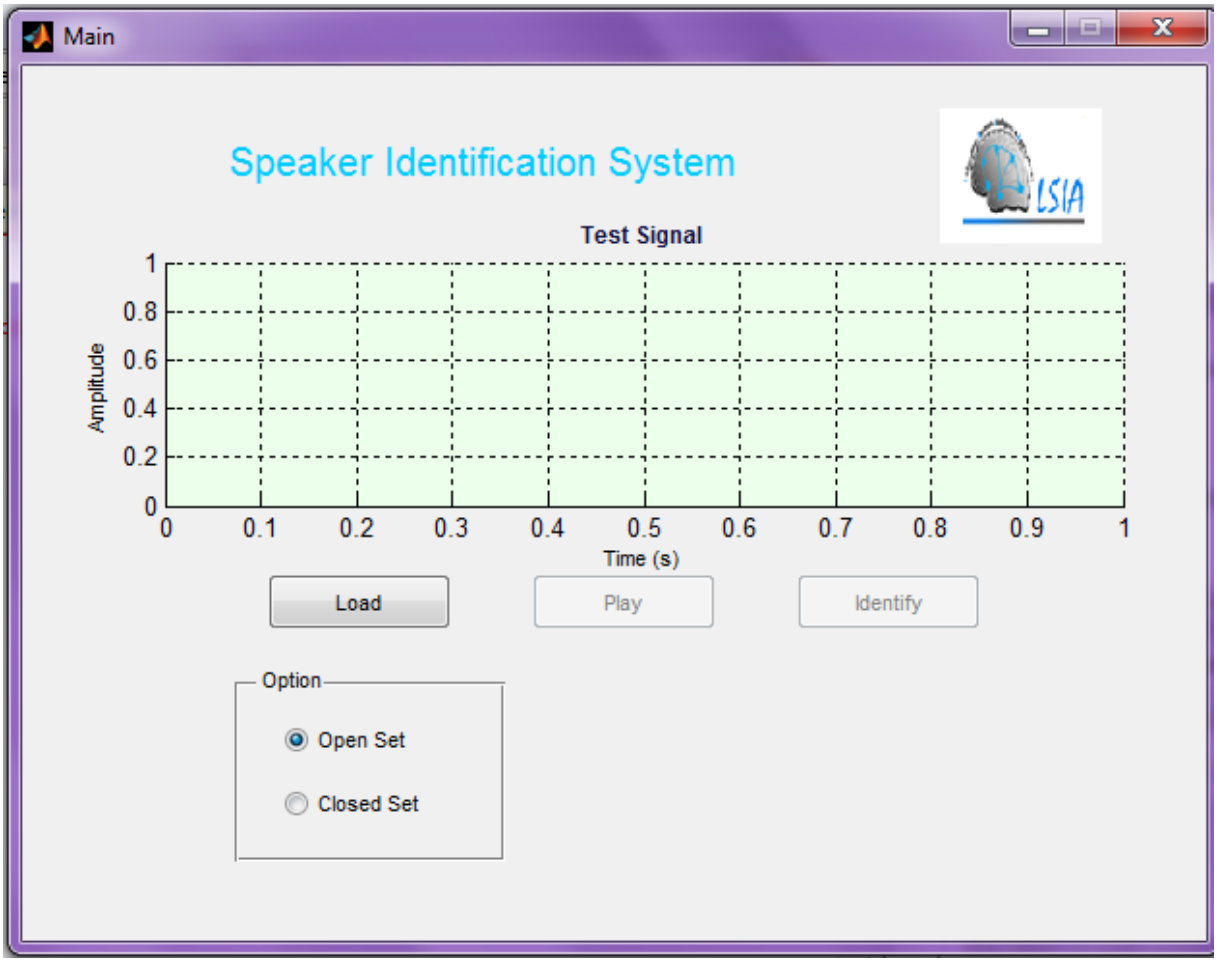


*Figure3.8 : Interface d'accueil*

Si on veut faire le test dans le cas d'un groupe fermé, on utilise load pour charger un fichier wav pour le test, puis on coche le bouton Closed set, le système affiche l'identité du locuteur testé (figure 3.9)

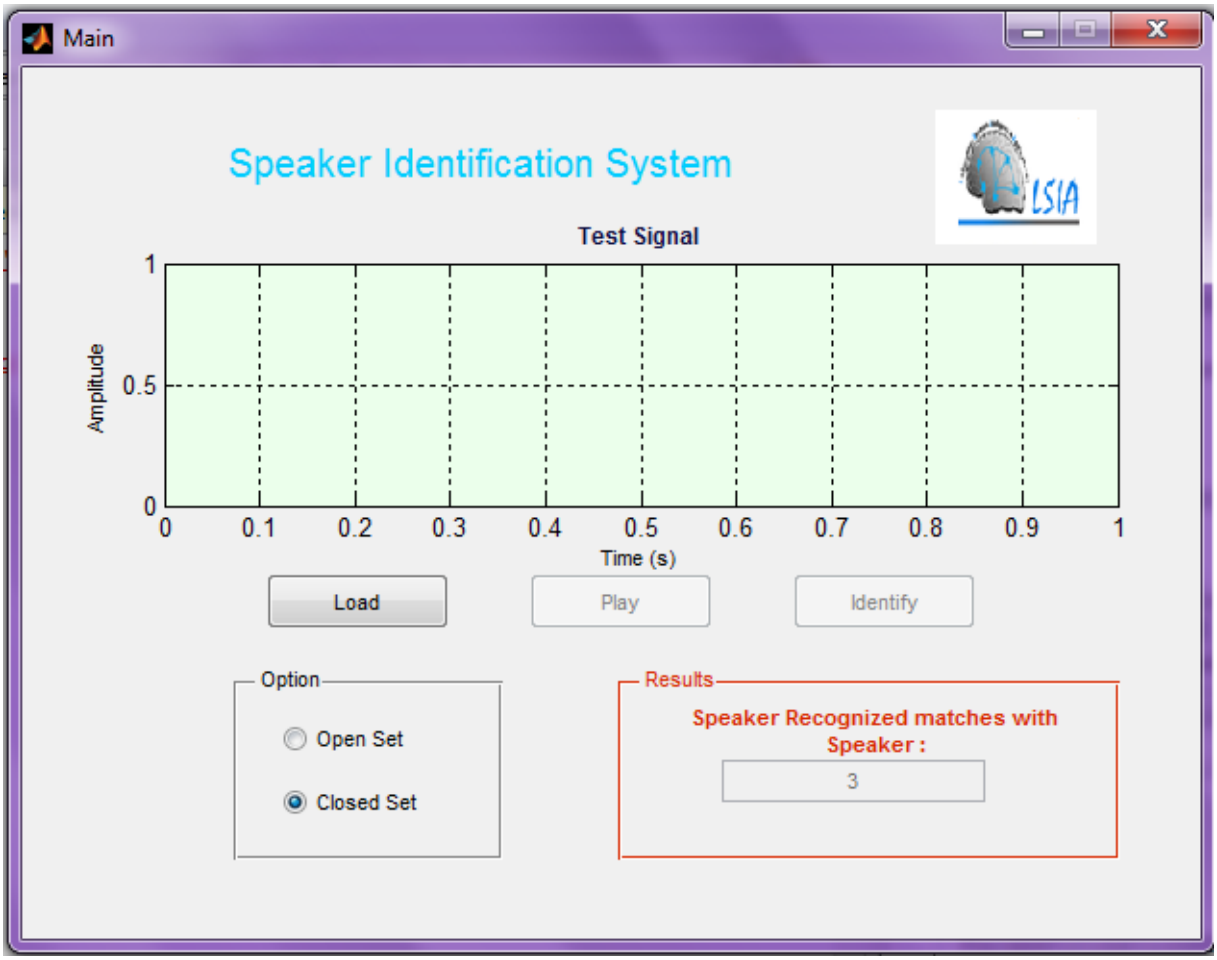


*Figure 3.9 : identification dans un groupe fermé*

Dans le cas d'un groupe ouvert, le système peut rejeter le locuteur de test s'il n'est pas présent dans la base de données (figure 3.10).

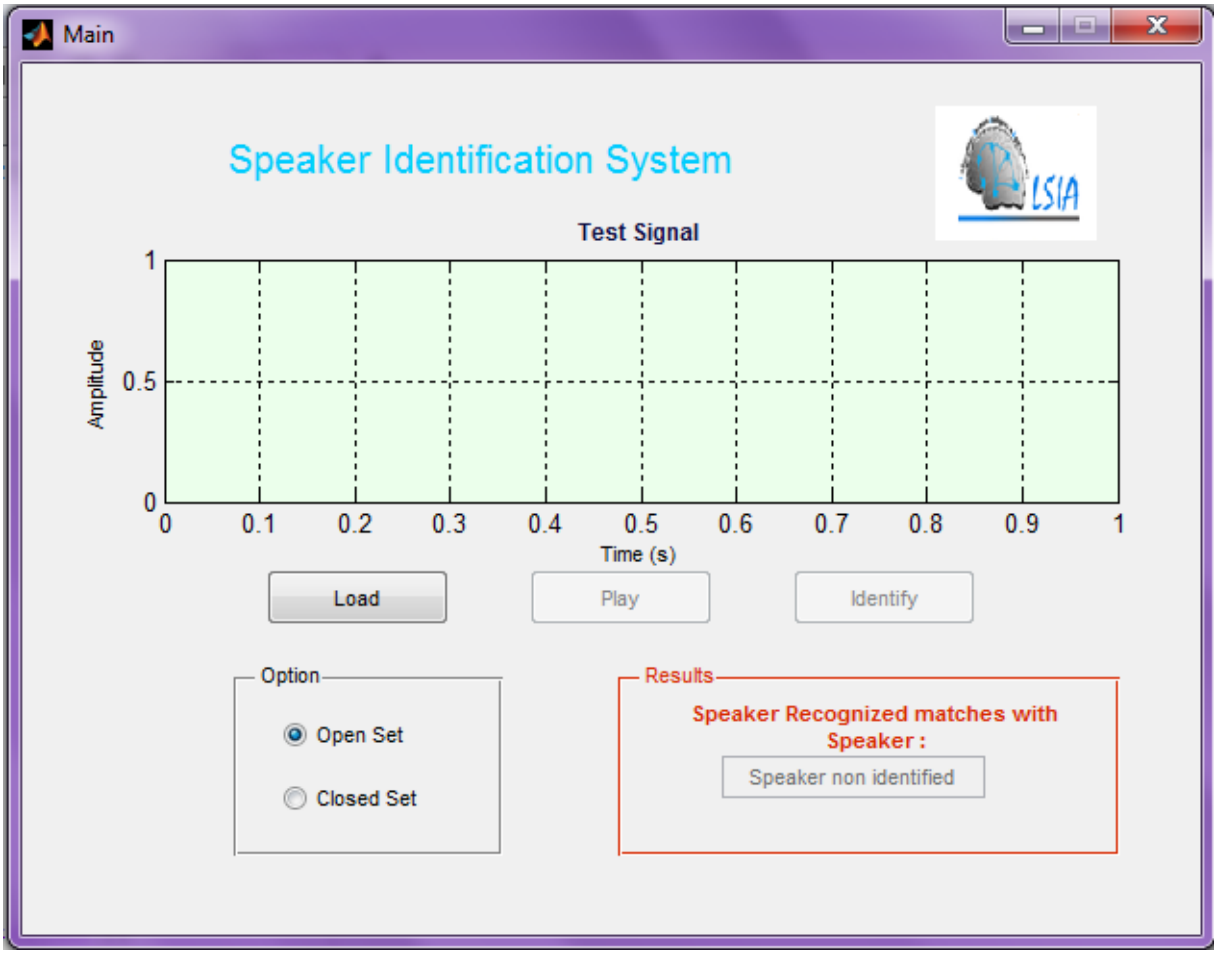


*Figure 3.10: Identification dans un groupe ouvert*

# Conclusion et perspectives

L'objectif principal du sujet abordé dans ce mémoire est de faire une étude bibliographique sur le fonctionnement d'un système RAL, de présenter les différents travaux et recherches effectués spécialement dans le cadre de la voix sur IP vu qu'elle présente la tendance actuelle des technologies de communication, et de terminer par une étude expérimentale dans le cadre de l'identification automatique de locuteur par la voix IP.

Pour pouvoir faire cette étude expérimentale, nous avons tout d'abord construit une base de données de la voix sur IP, nous avons utilisé 20 locuteurs (15 hommes et 5 femmes), avec 1 minute dans l'apprentissage, et 10 secondes dans le test. Cette base de données nous a permet de faire plusieurs tests d'évaluation sur les étapes de l'identification automatique du locuteur en mode indépendant du texte dans le cas du milieu ouvert et fermé, d'abord l'étude des six variantes majeures de la paramétrisation par MFCCs pour en déduire la meilleure, les résultats obtenus montrent que la variante Davis_Skowronski_MFCC_FB20 a donné de meilleurs performances, elle a donné un taux de reconnaissance de 100% dans le cas où on utilise 32 ou 128 modèles GMMs, cette performance élevée a été justifié par le fait que la voix IP utilise une large bande passante pour transmettre un flux important de données, nous avons par la suite comparé ces résultats avec l'identification de la voix enregistré par le microphone, nous avons remarqué une dégradation de performances de la méthode Davis_Skowronski, et un taux de reconnaissance élevé pour HTK_MFCC 26 et 24, cela a été expliqué par le fait que ces approximation sont basé sur la perception du pitch du système auditoire humain.

La technique de modélisation GMM a été largement étudiée dans ce travail. A travers les expériences que nous avons effectuées, nous pouvons dire que le modèle GMM en mode indépendant du texte 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 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, càd présenter des données qui n'existent pas dans l'espace de paramètres acoustiques du locuteur en question comme pour le cas de la variante Slaney_FB40. L'ordre du modèle GMM doit être suffisamment grand pour présenter l'ensemble des vecteurs acoustiques d'une population de données, d'après les expériences que nous avons effectuées, un modèle GMM composé de 32 ou 128 gaussiennes est largement suffisant pour représenter la distribution des vecteurs acoustiques d'un seul locuteur.

La reconnaissance automatique de locuteur par la voix IP est en plein développement, et elle comporte plusieurs axes de recherches qui peuvent être explorés dans les futurs travaux, parmi ces axes nous pouvons citer : Les systèmes RAL sont très couteux en termes de temps d'exécution. En exploitant les architectures parallèles disponibles dans les PC modernes, nous proposons de définir des algorithmes parallèles qui permettent de minimiser le temps

d’exécution, et de concevoir des systèmes de reconnaissance automatique de locuteur par la voix IP en temps réel, l’étude des différents facteurs qui peuvent influencer les performances du système de reconnaissance automatique de locuteur : le nombre des MFCCs utilisé, les conditions d’enregistrement, la langue, le support de transmission lors de la phase de l’extraction des paramètres, de l’apprentissage du modèle GMM et celle de décision.

# Bibliographie

[1] Ali, Q., & Ghani, N. A. (2010). Reviewing Speaker Recognition Influence in Strengthening VOIP Caller Authentication, *1*(4), 147–153.

[2] Backes, M., Doychev, G., Markus, D., & Boris, K. (n.d.). Speaker Recognition in Encrypted Voice Streams, 1–16.

[3] Basu, A. paliwal, Systems, C., Group, C., & Road, H. B. (n.d.). A Speech ehancement method based on Kalman Filtering.

[4] Besacier, Laurent. Pedro, Mayorga. Jean-François, Bonastre. Corinne, F. (n.d.). METHODOLOGY FOR EVALUATING SPEAKER VERIFICATION.

[5] Bilan Triennal des activités de recherche,DG-RSDT. (2011).

[6] Cernock, J., Karafi, M., & Schwarz, P. (n.d.). Robust heteroscedastic linear discriminant analysis and LCRC posterior features in large vocabulary continuous speech recognition, 2–5.

[7] Amehraye, A. (2009). débruitage perceptuel de la parole.

[8] Dan, Q., Honggang, Y., Hui, T., & Bingxi, W. (2008a). Two Schemes for Automatic Speaker Recognition over VOIP. *2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, 695–699.

[9] Dan, Q., Honggang, Y., Hui, T., & Bingxi, W. (2008b). VOIP compressed-domain automatic speaker recognition based on probabilistic stochastic histogram. *2008 9th International Conference on Signal Processing*, 692–696.

[10] Fredouille, C. (2000). Approche Statistique pour la Reconnaissance Automatique du locuteur: Informations dynamiques et Normalisation Bayesienne des vraisemblances.

[11] Fredouille, C., Gravier, G., Magrin-chagnolleau, I., Meignier, S., Merlin, T., Ortega-garc, J., … Reynolds, D. A. (2004). A Tutorial on Text-Independent Speaker Verification Fr ed, 430–451.

[12] Furui, S. (2004). 50 years of progress in speech and speaker recognition Sadaoki Furui Department of Computer Science Tokyo Institute of Technology, (1).

[13] Ganchev, T. (2011). *Contemporary Methods for Speech Parametrization* (Amy Neuste.).

[14] Heights, Y. (2005). IBM Research Report CSR : Speaker Recognition from Compressed VoIP Packet Stream, *23499*.

[15] Hermansky, H., & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, *2*(4).

[16] Joki, I., Joki, S., Gnjatovi, M., Se, M., & Deli, V. (2011). The Impact of Telephone Channels on the Accuracy of Automatic Speaker Recognition, *3*(2), 100–104.

[17] Kumar, N. (1997). Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition.

[18] Maciejko, W. (2012). The Effect of Voice Transmission over IP on Text-Independent Speaker Verification Performance, 6–9.

[19] Peláez-moreno, C., Gallardo-antolín, A., & Díaz-de-maría, F. (2001). Recognizing Voice Over IP : A Robust Front-End for Speech Recognition on the World Wide Web, *3*(2), 209–218.

[20] Pelecanos, J., & Sridharan, S. (2001). Feature Warping for Robust Speaker Verification School of Electrical and Electronic Systems Engineering.

[21] Reynolds, D. A. (1995). Automatic Speaker Recognition Using Gaussian Mixture Speaker Models, (2).

[22] Reynolds, D. a. (2003). Channel robust speaker verification via feature mapping. *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, *2*, II–53–6. doi:10.1109/ICASSP.2003.1202292

[23] Saraswathi, S., & Geetha, T. V. (2007). Time scale modification and vocal tract length normalization for improving the performance of Tamil speech recognition system implemented using language independent segmentation algorithm. *International Journal of Speech Technology*, *9*(3-4), 151–163.

[24] Teunen, R., Shahshahani, B., & Heck, L. (2000). A MODEL-BASED TRANSFORMATIONAL APPROACH TO ROBUST SPEAKER RECOGNITION, (Icslp), 1–4.

[25] Velho, F. (2006). Reconnaissance automatique du locuteur à l'aide de la transformée en ondelettes continue (TOC).

[26] Viikki, O., & Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, *25*(1-3), 133–147.

[27] Wagner, M. (2012). Analysis of Automatic Speaker Verification Performance over Different Narrowband and Wideband Telephone Channels, (December), 157–160.

[28] Waibel, P. Z. and A. (1997). Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition, (May).

[20] Wang, X., & Lin, J. (2007). APPLYING SPEAKER RECOGNITION ON VOIP AUDITING, (August), 19–22.

# Webographie

[w1] - https://www.ll.mit.edu/mission/communications/ist/publications/

[w2] - http://lia.univ-avignon.fr/thematiques/

[w3] - http://www.lcpts.usthb.dz/

[w4] - http://en.hit.edu.cn/

[w5] - http://wapiti.telecom-lille1.eu/

[w6] - http://www.itl.nist.gov/iad/mig/tests/spk/