الجمهورية الجزائرية الديمقراطية الشعبية
REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
وزارة التعليم العالي والبحث العلمي
MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
جامعة محمد بوضياف –المسيلة
UNIVERSITE MOHAMED BOUDIAF– M'SILA

كلية التكنولوجيا
Faculté de Technologie

PROJET DE RECHERCHE – FORMATION UNIVERSITAIRE (PRFU)
Année : 2018

- **Chef de projet :Dr.YESSAD DALILA**

Grade :master 2

Filière :électronique Spécialité :STLC

- **Membres**

Nom &Prénoms	Grade	Etablissement
BOUZIDI ADNANE	MASTER 2	
GUENDOUZ OUARD	MASTER 2	

Intitulé du projet :

Vérification automatique du locuteur sur IP

N° de CODE du PRFU

REMERCIEMENTS

Le travail présenté dans cette thèse a été effectué au Département d'Electronique, la Faculté de Technologie de l'Université de M'sila.

Ce travail n'a pas été possible sans notre encadreur, ***Dr. YESSAD DALILA****, pour nous encourager et nous aider pendant cette modeste étude*

Nous remercions très sincèrement Monsieur ***LAADJAL MOHAMED****, Président du département d'électronique del'Université de M'sila,pour ses encouragements.*

Nous exprimons mes reconnaissances aux membres de jury pour bien vouloir accepter de présider et examiner ce travail.

Mes remerciements vont aussi à tous les membres du département d'électronique.

Enfin, nous remercions tous ceux qui ont contribué de près ou de loin, pour Leurs soutiens moraux ou matériels, dans la réalisation de ce projet.

Table des matières

Liste des figures..... i
Sigles et abréviations..... ii
Introduction générale..... 1

Chapitre I : La Reconnaissance Automatique du Locuteur

I.1. Introduction..... 5
I.2. Mécanismes de production de la parole..... 5
I.2.1. Production de la parole..... 5
I.2.2. La perception de la parole sur la bande téléphonique. 7
I.3. Sources de variabilité du signal de la parole..... 8
I.3.1. Variabilité intra-locuteur..... 8
I.3.2. Variabilité inter-locuteurs..... 9
I.3.3. Variabilité due au matériel..... 9
I.3.4. Robustesse en environnements difficiles..... 9
I.3.5. Tentatives d'imposture - locuteurs non coopératifs..... 10
I.4. La reconnaissance vocale..... 10
I.5. Différentes Tâches en RAL..... 11
I.5.1. Identification Automatique du Locuteur..... 11
I.5.2. Vérification Automatique du Locuteur..... 12
I.5.3. Détection de Locuteurs..... 13
I.5.4. Indexation par Locuteur et ses variantes..... 14
I.6. Structure d'un système de RAL..... 16
I.7. Analyse acoustique du signal de parole..... 17
I.7.1. Paramètres prosodiques..... 17
I.7.1.1. Energie totale..... 17
I.7.1.2. Fréquence fondamental..... 17
I.7.2. Analyse spectrale du signal de parole..... 18
I.7.2.1. La transformée de Fourier discrète..... 18
I.7.2.2. La transformée de Fourier à courte terme..... 19
I.7.2.3. Les coefficients MFCC (Mel Frequency Cepstral Coefficients)..... 19
I.7.3. Paramètres exploitant la dynamique du signal de parole 23
I.8. Modélisation des locuteurs..... 23
I.8.1. L'approche vectorielle..... 24
I.8.2. L'approche statistique..... 25
I.8.3. L'approche connexionniste..... 27
I.8.4. L'approche prédictive..... 27
I.8.5. L'approche discriminante..... 28
I.9. Prise de décision..... 28
I.9.1. Décision en identification..... 28
I.9.2. Décision en vérification..... 28
I.10. Mesures de performances..... 29
I.11. Les courbes DET (Detection Error Tradeoff)..... 30
I.12. Conclusion..... 31

Chapitre II : La reconnaissance automatique du locuteur sur IP

II.1. Introduction..... 32
II.2. Système distribué..... 32
II.3. La reconnaissance automatique du locuteur distribuée sur IP(DSR)..... 33
II.4. Architecture client-serveur..... 35
II.5. Sockets..... 36
II.6. Protocoles réseau et transport..... 38
II.6.1. Le protocole IP..... 38
II.6.2. Le protocole TCP..... 39
II.6.3. Le protocole UDP..... 40
II.7. Conclusion..... 42

Chapitre III : Evaluation expérimentale

III.1. Introduction..... 44
III.2. Les outille de programmation utilisés..... 44
III.2.1. MATALB..... 44
III.2.2. C++..... 44
III.3. Description des bases de données..... 44
III.3.1. La base de données ARADIGIT..... 44
III.3.2. Les bases de données extraites d'ARADIGIT..... 45
III.4. Architecture client/serveur avec le codec G.729..... 45
III.4.1. Côté client..... 45
III.4.2. Les étapes de transmission..... 46
III.4.2. Côté serveur..... 47
III.5. Extraction des caractéristiques..... 48
III.6. Evaluation des performances..... 48
III.6.1. Influence de l'ordre de modèle sur ARADIGIT8K..... 48
III.6.2. Influence de nombre des paramètres sur ARADIGIT8K..... 50
III.6.3. Influence du G.729 sur ARADIGIT8K..... 53
III.7. Conclusion..... 56

Conclusion Générale et perspectives..... 57
Bibliographies..... 58

Liste des figures

Chapitre I

Figure I.1 Organes de production de la parole…… 6
Figure I.2 Principaux axes du traitement automatique de locuteur……. 11
Figure I.3 Identification automatique du locuteur……..……. 12
Figure I.4 La vérification automatique du locuteur……...…….. 13
Figure I.5 La tâche d'Indexation par Locuteur d'un flux audio……..…… 14
Figure I.6 La tâche de suivi de locuteurs……..…….. 15
Figure I.7 Schéma typique d'un système de reconnaissance automatique……..……... 16
Figure I.8 La fréquence fondamentale……..…….. 18
Figure I.9 Le champ auditif humain…..……..…… 20
Figure I.10 La transformation du Hz en Mel……...…… 21
Figure I.11 Calcul des coefficients MFCC avec une échelle Mel……... 21
Figure I.12 Banc de Filtres Triangulaires équidistance en échelle Mel……..…… 22
Figure I.13 Structure du paradigme GMM-UBM en RAL…… 27
Figure I.14 Types d'erreurs dans un système RAL……..……... 30
Figure I.15 Exemple courbe DET……..……. 31

Chapitre II

Figure II.1 Schéma de principe d'un système DSR basé sur la transmission des vecteurs des paramètres du côté client (front-end) au coté serveur (back-end) pour faire la reconnaissance……. 34
Figure II.2 Schéma de principe d'un système DSR basé sur la transmission de la parole codée (bit-stream) du coté client au coté serveur où il faut décoder le bit-stream et resynthétiser la parole. A partir de celle ci, l'extraction des paramètres est effectuée en vue de la reconnaissance……..... 34
Figure II.3 Interaction dans le modèle client-serveur…. 36
Figure II.4 Modèle de la communication via socket……..……. 36
Figure II.5 Schéma d'un socket……... 37
Figure II.6 Structure de l'entête IP basé sur 20 octets……. 38
Figure II.7 (a) : Modèle de référence OSI, (b) : Modèle TCP/IP (Internet)…… 39
Figure II.8 Structure d'un segment TCP……. 40
Figure II.9 Structure d'une trame UDP……... 41

Chapitre III

Figure III.1 Implémentation coté client……... 46
Figure III.2 Communication entre client serveur…… 47
Figure III.3 Implémentation coté serveur…… 47
Figure III.4 Performance d'un système VAL pour 128 gaussiennes de modèle GMM-UBM…… 49
Figure III.5 Performance d'un système VAL pour 256 gaussiennes de modèle GMM-UBM…… 49
Figure III.6 Performance d'un système VAL à base de GMM-UBM en utilisant 40 MFCC……..... 51
Figure III.7 Performance d'un système VAL à base de GMM-UBM en utilisant 60 MFCC……..... 51
Figure III.8 Performance d'un système RAL à base de 128 gaussiennes de modèle GMM-UBM en utilisant 19 paramètres MFCC avec leur delta et l'énergie……. 52

Figure III.9 Les performances d'un système VAL sur IP en utilisant le codec G.729 avec modélisation GMM-UBM…….... 54
Figure III.10 Les performances d'un système VAL normal et sur IP basé sur le codec G.729 en utilisant le modèle GMM-UBM…….... 55

Abréviations

VoIP	Voice Over IP
ToIP	Telephony Over IP
DSR	Distributed Speech and Speaker Recognition
UIT	Union Internationale des Télécommunications
RAL	Reconnaissance Automatique Du Locuteur
IAL	Identification Automatique Du Locuteur.
VAL	Vérification Automatique Du Locuteur
FFT	Fast Fourier Transform
MFCC	Mel-Frequency Cepstral Coefficient
DTW	Dynamic Time Warping
HMM	Hidden Markov Model
GMM	Gaussian Mixture Model
UBM	Universal Background Model
FA	Fausse Acceptation
FR	Faux Rejet
DET	Detection Error Tradeoff
OSI	Open System Interconnection
TCP/IP	Transmission Control Protocol / Internet Protocol
IP	Internet Protocol
TCP	Transmission Control Protocol
UDP	User Datagram Protocol
RTP	Real-Time Protocol
API	Application Programming Interface

Introduction générale

INTRODUCTION GENERALE

Depuis le début de la recherche dans le domaine du traitement du signal, les chercheurs ont toujours eu une attention particulière pour le signal de parole, car la parole est sans doute le moyen de communication le plus simple et le plus efficace chez les humains. Grâce aux développements des technologies de l'informations et de la communications, en particulier en traitement du signal et en informatique, le rêve de communiquer avec des machines est devenu de plus en plus réalisable. Les recherches actuelles proposent de nombreux systèmes de reconnaissance automatiques, parmi lesquels deux ont connu une progression considérable : La Reconnaissance Automatique de la Parole (RAP) qui consiste à reconnaître le message prononcé et la Reconnaissance Automatique du Locuteur (RAL) qui consiste à reconnaître (ou authentifier) l'identité du locuteur à l'origine du signal de parole présenté. Les ordinateurs et les logiciels qui se construisent actuellement, bien que capables de traiter énormément d'informations en un temps très court, n'ont pas encore la capacité de générer ou de comprendre les finesses de la parole humaine. Cependant, de nombreuses applications en reconnaissance de la parole sont déjà industrialisées. Allant de la dictée vocale à la commande d'opérations diverses dans les navettes spatiales. De plus en plus, les entreprises de télécommunications et de services (banques, assurances), désireuses d'améliorer leur service à la clientèle, tentent d'introduire des applications basées sur la reconnaissance de la parole. Le développement de la reconnaissance de la parole a booste les nombreuses applications de reconnaissance du locuteur, plus particulièrement dans les services demandant une reconnaissance de l'identité du locuteur comme l'accès aux boîtes vocales, à des services par abonnements, consultation de comptes en banques ou d'autres transactions bancaires à distance, achats par téléphone, le contrôle d'accès à distance de bases de données, les services d'information et de réservation à distance, etc. La tendance actuelle montre une évolution vers l'exécution de diverses transactions en utilisant la Voix sur IP (VoIP : Voice Over IP). Cette nouvelle technologie est en pleine expansion du fait de la valeur ajoutée qu'elle apporte aux utilisateurs par rapport à la téléphonie classique, telle que la mobilité, ou la possibilité de transmettre non seulement la voix, mais aussi des données et du contenu multimédia, ou encore le coût réduit des appels long distance. La VoIP propose donc des perspectives en termes de nouveaux services, la convergence du réseau de données et du réseau supportant la voix, aboutissant à l'émergence de nouveaux services voix-données.

Le principe de la voix sur réseau par paquets consiste, à partir d'une numérisation de la voix, à comprimer le signal, à le découper en paquets de données et à les transmettre sur le réseau IP. A l'arrivée, les paquets transmis sont réassemblés, le signal de données obtenu est décompressé puis converti en signal analogique pour restituer le signal sonore à l'utilisateur. Les caractéristiques de la VoIP, telles que la possibilité d'une plus grande compression du signal de parole, ou l'utilisation optimale du réseau grâce à la transmission de l'information par paquets, en font une application ambitieuse. Toutefois, un matériel de téléphonie IP ne pourra s'imposer que dans la mesure où la qualité de la parole reçue sera suffisante. Une considération importante dans toute compression de parole est la qualité du signal reconstruit. Les codecs audio sont basés sur différentes techniques de compression de la parole visant à éliminer la redondance dans le signal de parole pour obtenir une bonne compression et de réduire les coûts de stockage et de transmission.

Les nouvelles applications de la VoIP, comme par exemple, Skype, Google Talk, etc, utilisent de nombreux codecs audio. Les codecs vocaux typiques utilisés dans la VoIP comprennent ceux proposés par l'ITU-T telles que G.711, G.729 et G.723.1 ; par ETSI tels que AMR; les codecs open-source tels que les codecs iLBC et Speex ; et les propriétaires tels que le codec Silk de Skype. Ces codecs ont un débit variable dans la gamme de 6 à 40kbit/s et une fréquence d'échantillonnage variable sur une bande étroite à une bande large. Certains codecs ne peuvent fonctionner qu'à un débit binaire fixe, tandis que de nombreux codecs avancés peuvent avoir des débits binaires variables qui peuvent être utilisées pour l'adaptation afin d'améliorer la qualité de la voix.

Le développement de la VoIP, et par conséquent de la téléphonie sur IP (ToIP : Telephony Over IP), a ouvert de nouveaux horizons aux applications en reconnaissance vocale, d'où l'émergence de la Reconnaissance Distribuée (Distributed Speech and Speaker Recognition : DSR), touchant aussi bien la RAP que la RAL.

Ce travail a pour but de diagnostiquer les nouveaux défis posés à la reconnaissance du locuteur dans le contexte récent de la voix sur IP, et de proposer quelques solutions permettant d'y améliorer les performances d'un système de reconnaissance automatique sur VoIP.

Dans le cadre de ce travail nous nous intéressons au codec G.729 dédié à la VoIP. Il s'agit donc de mettre en œuvre le codec G.729, dans le but de construire le système de la reconnaissance du locuteur sur IP (DSR: Distributed Speech/Speaker Recognition) basé sur l'architecture client/serveur, qui offre la possibilité de répartir les tâches de reconnaissance automatique de locuteur entre les machines clientes et serveurs sur le réseau IP. Afin de

concevoir ce système de reconnaissance distribué, il convient : d'une part de comprendre en quoi le signal de parole est réellement complexe, c'est à dire connaitre l'objet ou l'observation d'entrée, d'autre part de définir correctement la tache de l'architecture client/serveur basé sur le G.729 codec de la parole, c'est à dire les contraintes imposées et les performances attendues. Ce travail s'appuyant sur l'implémentation de l'architecture client/serveur en intégrant le codec G.729, avec implémentation de l'encodeur coté client et le décodeur coté serveur. Ainsi, le travail a tout d'abord consiste à étudier le domaine de vérification du locuteur sur IP couvrant a la fois la compression du son jusqu'a sa restitution, en passant par l'architecture réseau choisie, le codage de compression audio sur IP et la modélisation de vérification utilisée. L'étude de toutes ces composantes a permis d'établir les différents axes de recherche étudies pendant ce travail.

Chapitre 1 détaille la reconnaissance automatique du locuteur (RAL). Il présente tout d'abord les mécanismes de production de la parole, ensuite les sources de variabilité du signal de la parole. Il souligne également les applications majeures associées a la RAL et les différentes taches liées a la RAL. Il expose quelques approches utilisées pour les systèmes de RAL et présente, enfin, les méthodes d'évaluation des performances des systèmes de RAL.

Chapitre 2, rappelle le contexte d'un système de reconnaissance distribuée (DSR), l'architecture client-serveur permettant d'allier la tache de la reconnaissance de locuteur sur IP, ainsi que les différents protocoles qui régissent la transmission les données via un réseau IP (Internet Protocol), en particulier TCP/IP et UDP.

A partir de l'architecture présentée au chapitre 2, nous reportons les résultats obtenus au chapitre 3. Ces résultats incluent l'étude des scores issus du système de vérification basé sur le paradigme GMM-UBM sur IP.

Nous concluons ce travail par une récapitulation des principales conclusions et discussions de ces travaux tout en donnant les perspectives des futurs travaux qui auront pour but d'utiliser d'autres type de paramètres extraites à partir du codec G.729, d'autres types de codecs ainsi que des travaux sur d'autres types de modèles.

Chapitre I

La reconnaissance automatique du locuteur

I.1 Introduction

La reconnaissance automatique du locuteur (RAL) est un terme générique regroupant les problèmes relatifs à l'identification ou à la vérification du locuteur sur la base de l'information contenue dans le signal acoustique : il s'agit de reconnaître une personne à partir de sa voix.

Dans ce chapitre, nous décrivons le processus de production de la parole à travers les mécanismes mis en jeu lors de la phonation. Ensuite, nous énonçons les variabilités interlocuteurs, ainsi que les paramètres servant à la discrimination des locuteurs. Nous terminons par une présentation de la structure générale d'un système de RAL, qui peut se subdiviser en trois étapes qui sont l'analyse acoustique du signal de parole, la modélisation du locuteur et une dernière étape de décision (décider de l'identité du locuteur).

I.2 Mécanismes de production de la parole

Pour développer un système de reconnaissance automatique du locuteur, il est nécessaire de connaitre les paramètres acoustiques caractérisant le locuteur. Pour cela, une bonne compréhension du processus de production de la parole est nécessaire. La parole est considérée parmi les principaux moyens de communication de l'être humain. Elle contient essentiellement le sens du message prononcée par le locuteur ainsi que des informations individuelles concernant l'identité et parfois l'émotion du locuteur.

I.2.1 Production de la parole

Le signal de la parole appartient à la classe des signaux acoustiques produits par des vibrations des couches d'air. Les variations de ce signal reflètent les fluctuations de la pression de l'air. Le processus de production de la parole est un mécanisme très complexe qui repose sur une interaction entre les systèmes neurologique et physiologique. La parole commence par une activité neurologique [01]. Après que soient survenues l'idée et la volonté de parler, le cerveau dirige les opérations relatives à la mise en action des organes phonatoires. Le fonctionnement de ces organes est bien, quant à lui, de nature physiologique [02].

Une grande quantité d'organes et de muscles entrent en jeu dans la production des sons des langues naturelles. Le fonctionnement de l'appareil phonatoire humain repose sur l'interaction entre trois entités: les poumons, le larynx et le conduit vocal. (Voir figure I.1)

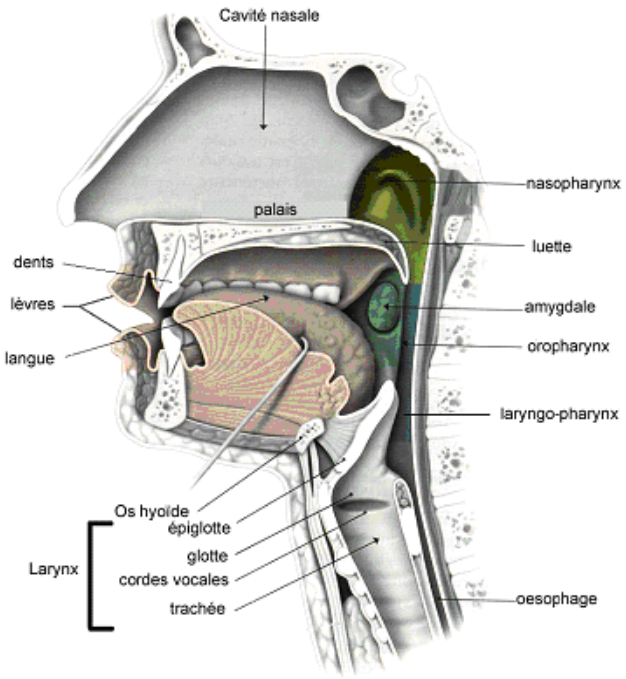


Figure I.1 : Organes de production de la parole [03].

Le larynx est une structure cartilagineuse qui a notamment comme fonction de réguler le débit d'air via le mouvement des cordes vocales. Le conduit vocal s'étend des cordes vocales jusqu'aux lèvres dans sa partie buccale et jusqu'aux narines dans sa partie nasale [04].

L'air des poumons est comprimé par l'action du diaphragme. Cet air sous pression arrive ensuite au niveau des cordes vocales. Si les cordes sont écartées, l'air passe librement et permet la production de bruit. Si elles sont fermées, la pression peut les mettre en vibration et l'on obtient un son quasi-périodique dont la fréquence fondamentale correspond généralement à la hauteur de la voix perçue. L'air mis ou non en vibration poursuit son chemin à travers le conduit vocal et se propage ensuite dans l'atmosphère. La forme de ce conduit, déterminée par la position des articulateurs tels que la langue, la mâchoire, les lèvres ou le voile du palais, détermine le timbre des différents sons de la parole. Le conduit vocal est ainsi considéré comme un filtre pour les différentes sources de production de la parole telles que les vibrations des cordes vocales ou les turbulences engendrées par le passage de l'air à travers les constrictions du conduit vocal [02].

Le son résultant peut être classé comme voisé ou non voisé selon que l'air émis a fait vibrer les cordes vocales ou non. Dans le cas des sons voisés, la fréquence de vibration des cordes vocales, dite fréquence fondamentale ou pitch, noté F_0, s'étend généralement entre 40

– 140 Hz pour les hommes, entre 180 – 300 Hz pour les femmes et entre 300 – 600 Hz pour les enfants [01] [04].

I.2.2 La perception de la parole sur la bande téléphonique

Il est nécessaire de rappeler que les fréquences audibles par une oreille humaine sont habituellement situées entre 20 Hz et 20 kHz : ce qui à première vu semble très loin de la bande étroite de la téléphonie. De plus, suite à de nombreuses utilisations, le cerveau humain, a créé une référence de la qualité « sonore » de la voix humaine, transmise à travers un système de téléphonie [05]. L'évaluation de la qualité est alors biaisée par la référence de la téléphonie fixe, fortement liée aux fréquences de la bande étroite (300-3400 Hz).

Par ailleurs, l'intelligibilité est fortement liée aux premiers formants. Dans Gleiss [06] il est montré que la sensation naturelle de la voix dépend fortement du premier formant et l'intelligibilité du second formant. En effet, les différents phonèmes sont perçus en fonction :

- Du rapport entre les formants;
- De leurs variations dans le temps.

De plus, le premier formant étant approximativement entre 270 et 730 Hz pour les hommes, et entre 310 et 850Hz pour les femmes, le choix de la fréquence de coupure basse à 300 Hz a été choisie judicieusement. La bande étroite permet un bon compromis entre intelligibilité et qualité du son.

Pour autant, celle-ci ne permet pas de transmettre la fréquence fondamentale F_0, qui dans Gleiss [06] semble être liée à la sensation naturelle de la voix. Celle-ci comprise entre 110 et 200 Hz pour un adulte, et montant jusque 300 Hz pour un enfant, permet de transmettre la prosodie [07], comme les intonations ou les émotions. La bande étroite introduit une dégradation de la sensation naturelle de la voix par :

- L'atténuation du premier formant;
- L'absence de transmission de F_0;
- L'absence de transmission des hautes fréquences.

En conclusion, la parole humaine produit des fréquences qui en partie ne sont pas comprises dans la bande étroite, et qui sont nécessaires pour obtenir une voix humaine naturelle. Ce qui complique le processus de reconnaissance du locuteur à travers un canal téléphonique ou un réseau IP.

I.3 Sources de variabilité du signal de la parole

Le signal de parole est très complexe où se mêlent informations linguistiques, informations caractéristiques du locuteur, informations relatives au matériel utilisé pour la transmission ou l'enregistrement du signal, etc. En outre, le signal de parole est très redondant. Cette caractéristique est d'ailleurs reconnue pour faciliter la communication entre deux personnes dans un environnement très bruyant. Par ces différents aspects, le signal de parole présente une très grande variabilité.

La capacité des systèmes de RAL à différencier plusieurs individus repose essentiellement sur la variabilité inter-locuteur la disposition du signal de parole à varier entre différents individus. Néanmoins, le signal de parole renferme d'autres types de variabilité qui rendent problématique la tâche de reconnaissance, telles que la variabilité intra-locuteur ou la variabilité due au matériel. Par ailleurs, les systèmes de RAL doivent faire face à d'autres difficultés liées davantage au domaine applicatif, comme l'utilisation des systèmes dans des conditions difficiles, les tentatives d'imposture, etc. [08]

I.3.1 Variabilité intra-locuteur

Si le signal de parole est variable entre deux individus, il varie également pour un même individu. Cette variabilité intra-locuteur est induite par l'évolution naturelle ou volontaire de la voix d'une personne. Cette évolution peut être :

- Ponctuelle ou à très court terme. L'état pathologique (fatigue, rhume, etc.) ou émotionnel (stress) d'une personne provoquent des altérations momentanées dans sa voix. Dans ce sens, la voix d'une personne peut évoluer entre le début et la fin de la journée (fatigue, irritation due à la pollution) [09, 10]. D'autre part, il est impossible pour un individu de répéter consécutivement deux phrases identiques et de produire un même signal de parole pour ces deux phrases. Une légère variation est toujours observée. Finalement, une personne a la possibilité de modifier volontairement sa voix.
- A moyen terme. En RAL, le comportement d'un individu se modifie au fur et à mesure de son utilisation du système. L'individu devient de plus en plus confiant et sa voix évolue dans ce sens.
- A long terme. La voix change au fur et à mesure du vieillissement d'une personne.

La variabilité intra-locuteur pose le problème de la représentativité des signaux de parole collectés lors des sessions d'entraînement (et des modèles des locuteurs correspondant)

au sein des systèmes de RAL. Des travaux ont montré que les performances d'un système sont très fortement corrélées au temps qui sépare les sessions d'entraînement et les tests [11, 12]. Plus ce temps augmente, plus les performances se dégradent.

I.3.2 Variabilité inter-locuteurs

Les signaux de parole véhiculent plusieurs types d'informations. Parmi eux, la signification du message prononcé est d'importance primordiale. Cependant, d'autres informations telles que le style d'élocution ou l'identité du locuteur jouent un rôle important dans la communication orale. Ecouter un interlocuteur permet d'avoir des indications concernant son sexe, son état émotionnel et bien souvent de l'identifier si on l'a déjà entendu. Dans notre vie quotidienne, ces informations, sont très utiles. Elles nous permettent, par exemple, de différencier les divers messages que nous entendons selon le locuteur et leur degré d'importance. Si toutes les voix étaient perçues de la même façon, il serait par exemple impossible de suivre une émission radio faisant participer des personnes différentes.

La grande variabilité entre les locuteurs est due, d'une part, à l'héritage linguistique et au milieu socioculturel de l'individu, et d'autre part aux différences physiologiques des organes responsables de la production vocale. L'expression acoustique de ces différences peut être traduite par une variation de la fréquence fondamentale, dans l'échelle des formants (plus haute chez les femmes et les enfants que chez les hommes) et dans le timbre de la voix (richesse en harmoniques due à la morphologie du locuteur et au mode de fermeture des cordes vocales).

I.3.3 Variabilité due au matériel

Le signal de parole est porteur d'informations caractérisant le matériel utilisé lors de sa capture (ex : microphone, combiné téléphonique), de sa transmission (ex : lignes téléphoniques) et de son enregistrement (ex : microphones, convertisseurs). Ces informations apparaissent le plus souvent sous la forme de déformations/dégradations du signal de parole. Ces déformations sont différentes selon le type de matériel utilisé.

I.3.4 Robustesse en environnements difficiles

Comme nous venons de l'évoquer, les environnements téléphoniques mettent à rude épreuve les systèmes de RAL. Néanmoins, d'autres environnements nécessitent de la part des systèmes de RAL une grande robustesse.

Le réseau GSM est considéré ici comme un environnement à part entière, en marge des environnements téléphoniques. Des travaux expérimentaux sur la comparaison du réseau téléphonique classique et d'un réseau GSM font état de différences significatives dans la qualité des signaux [08]. En effet, les signaux transmis par réseau GSM montrent un niveau de bruit bien supérieur (les appels par téléphones mobiles sont souvent effectués dans des endroits plus bruyants que ceux d'un téléphone fixe : voiture, gare, rue), un niveau de voix plus élevé, souvent proche de la saturation entraînant des distorsions au sein du signal ainsi que de potentielles dégradations des signaux dues au codage de la parole.

Finalement, les systèmes de RAL doivent renforcer leur robustesse face au bruit ambiant. En effet, d'une manière similaire à la variabilité intra-locuteur ou à la variabilité due aux changements de matériel, la variabilité du niveau de bruit entre apprentissage et test peut susciter une baisse de performances des systèmes de RAL.

I.3.5 Tentatives d'imposture - locuteurs non coopératifs

Selon l'application visée, un système de RAL peut faire l'objet d'attaques d'individus usurpant l'identité de quelqu'un d'autre. Ces attaques (ou tentatives d'imposture) peuvent, par exemple, avoir pour dessein des transactions frauduleuses sur le compte bancaire d'un client ou l'accès à des données confidentielles. Un système de RAL doit par conséquent être robuste face à de telles attaques.

Dans un contexte judiciaire, le système de RAL peut être soumis à des locuteurs non coopératifs des locuteurs qui ne désirent pas être reconnus par le système.

I.4 La reconnaissance vocale

L’expression vocale est une caractéristique propre d’un locuteur. La reconnaissance vocale est un terme générique regroupant les problèmes relatifs à la reconnaissance du locuteur sur la base de l’information contenue dans le signal acoustique de la parole [02], elle est définie comme étant un processus de prise de décision utilisant des caractéristiques de la parole, afin de déterminer si une personne en particulier est à l’origine d’une énonciation. Cette prise de décision porte sur une éventuelle familiarité entre la voix cible et les voix de référence. La reconnaissance vocale peut être divisée en deux grandes classes : l’identification et la vérification comme on peut le constater sur la Figure I.2.

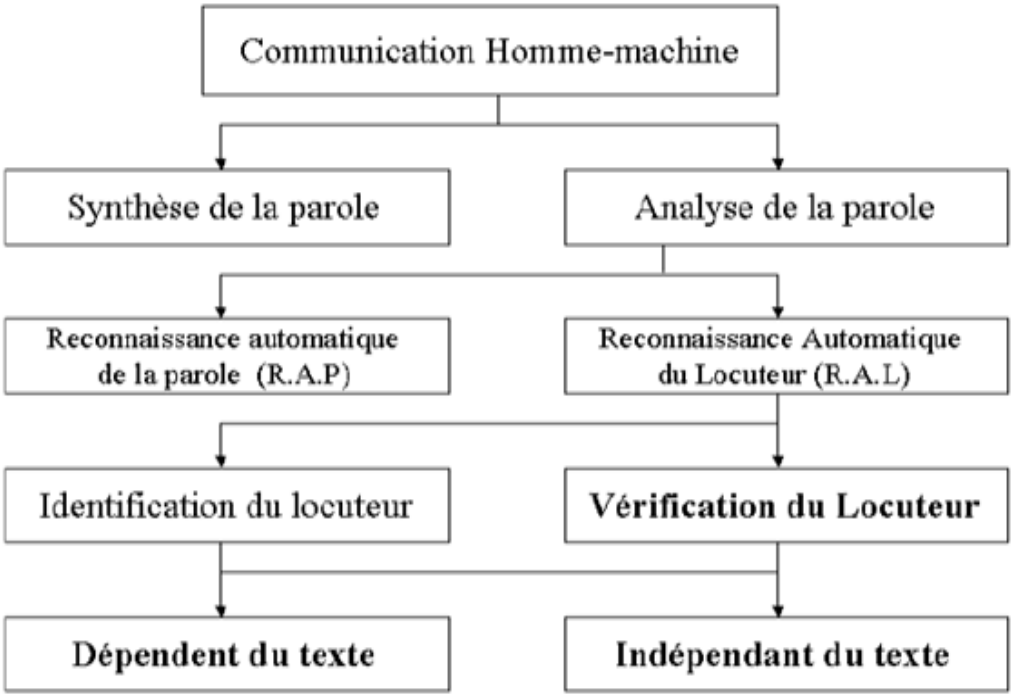


Figure I.2 : Principaux axes du traitement automatique de locuteur.

I.5 Différentes Tâches en RAL

L'Identification Automatique du Locuteur et la Vérification Automatique du Locuteur sont les tâches pionnières du domaine de la RAL. Plus récemment, les besoins applicatifs ont fait naître de nouvelles tâches comme l'Indexation par Locuteur de flux audio ou le Suivi de Locuteurs (Speaker Tracking) ou de nouvelles variantes telles que la détection de l'interaction d'un locuteur dans une conversation.

I.5.1 Identification Automatique du Locuteur

L'Identification Automatique du Locuteur (IAL) est le processus qui consiste à déterminer, parmi une population de locuteurs connus, la personne ayant prononcé un message donné. D'un point de vu schématique (voir figure I.3), une séquence de parole est donnée en entrée du système d'IAL. Pour chaque locuteur connu du système, la séquence de parole est " comparée " à une référence caractéristique du locuteur. L'identité du locuteur dont la référence est la plus " proche " de la séquence de parole est donnée en sortie du système d'IAL.

Deux modes sont proposés en IAL : l'identification en ensemble fermé pour lequel on suppose que la séquence de parole est effectivement prononcée par un locuteur connu du

système et l'identification en ensemble ouvert pour lequel le locuteur peut ne pas être connu. En mode "ensemble ouvert", le système d'IAL doit décider de la fiabilité de son jugement en acceptant ou rejetant l'identité qu'il a trouvée.

De par son principe - déterminer une identité parmi des identités potentielles - les performances des systèmes d'IAL se dégradent généralement au fur et à mesure que la population de locuteurs augmente [13, 14, 15].

En IAL, les applications sont peu nombreuses. On peut retenir, par exemple, l'utilisation d'un système d'IAL en vue de faciliter l'adaptation au locuteur des systèmes de RAP. Par ailleurs, il peut être intéressant pour des applications commerciales d'associer un même mot de passe pour une petite population de locuteurs (membres d'une famille, d'une société) [16]. Dans une telle situation, un système d'IAL en ensemble ouvert et dépendant du texte peut être utilisé pour contrôler l'accès à des données sensibles [08].

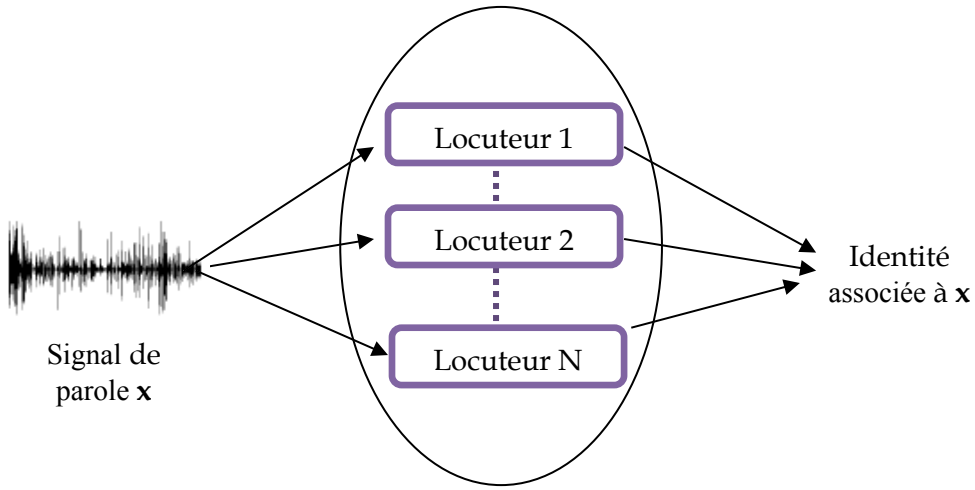


Figure I.3 : Identification automatique du locuteur.

I.5.2 Vérification Automatique du Locuteur

La Vérification Automatique du Locuteur (VAL) est le processus décisionnel permettant de déterminer, au moyen d'un message vocal, la véracité de l'identité revendiquée par un individu (figure I.4) [15]. L'identité ainsi que le message vocal constituent les deux entrées du système de VAL. L'identité, nécessairement connue du système, désigne automatiquement la référence caractéristique d'un locuteur. Une mesure de similarité est calculée entre cette référence et le message vocal puis comparée à un seuil de décision. Dans

le cas où la mesure de similarité est supérieure au seuil, l'individu est accepté. Dans le cas contraire, l'individu est considéré comme un imposteur et rejeté.

Les applications de VAL sont multiples et principalement commerciales :

- Serrures vocales pour le contrôle d'accès à des locaux ;
- Authentification pour l'accès à distance à des données sensibles ou à des services spécifiques à travers le réseau téléphonique (consultations ou transactions bancaires, consultations de bases de données à caractère confidentiel, consultations de boîtes vocales, télé-achat, etc.) ;
- Protection de matériel contre le vol (téléphones portables, voitures, etc.) ;
- Incarcération à domicile nécessitant une authentification régulière du prévenu [17,18, 19].

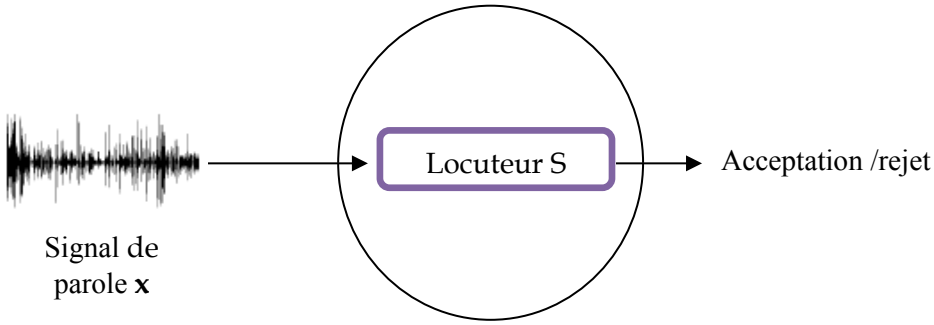


Figure I.4 : La vérification automatique du locuteur.

I.5.3 Détection de Locuteurs

La détection de locuteurs dans un flux audio est une variante de la VAL. Sa particularité est de considérer un flux audio composé de séquences de parole produites par plusieurs locuteurs (conversations, débats, conférences, etc.). Dans ce contexte, la tâche de détection consiste à déterminer si un locuteur donné intervient ou non dans le document audio. Dans le cas d'un flux audio mono-locuteur, la tâche de détection se résume à la tâche de vérification [13].

La tâche de détection est évidemment motivée par les instances militaires ou judiciaires. Néanmoins, elle demeure très intéressante dans le domaine de l'indexation de documents audio pour laquelle la détection d'un locuteur connu peut permettre de cibler plus

facilement un document audio particulier (séquence d'un journal télévisé ou d'une émission radio) [08].

I.5.4 Indexation par Locuteur et ses variantes

La tâche d'Indexation Automatique par Locuteur consiste à cibler les interventions des locuteurs dans un flux audio (figure I.5). En d'autres termes, indexer un document audio en locuteurs revient à indiquer à quel moment un individu prend la parole et qui est cet individu. La seule entrée d'un système d'indexation est le document audio à indexer.

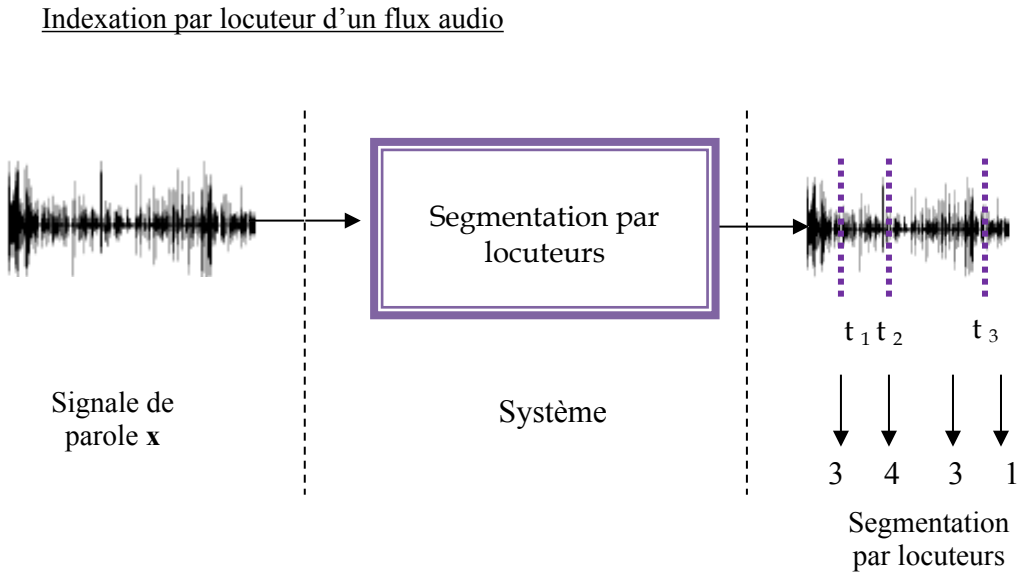


Figure I.5 : La tâche d'Indexation par Locuteur d'un flux audio.

Aucune information n'est donnée au système concernant le nombre de locuteurs présents dans le document ou leur identité. Contrairement aux systèmes d'IAL ou de VAL, les systèmes d'indexation ne détiennent pas de référence pour les locuteurs présents dans un document audio. Leur principe repose généralement sur une phase de segmentation " aveugle" en locuteurs suivie d'une phase de regroupement. Un système d'IAL permet finalement d'identifier les différents locuteurs présents dans le document. La sortie d'un système d'indexation ressemble généralement à la séquence suivante : le locuteur A est intervenu aux instants t1, t4, t6, le locuteur B aux instants t2, t5, le locuteur C à l'instant t3.

La tâche de suivi de locuteurs peut être considérée comme une version simplifiée de l'Indexation par Locuteur d'un flux audio (figure I.6). Le principe reste le même : déterminer les interventions d'un ou plusieurs locuteurs, appelés locuteurs cibles, dans un flux audio. La

simplification réside dans le fait que le système de suivi de locuteurs connaît nécessairement les locuteurs présents dans le document à indexer ou, du moins, ceux dont il doit détecter les interventions. Il possède une référence caractéristique pour chacun des locuteurs. Malgré cette simplification, le suivi de locuteurs reste une tâche très complexe. Trois grandes approches sont recensées dans la littérature :

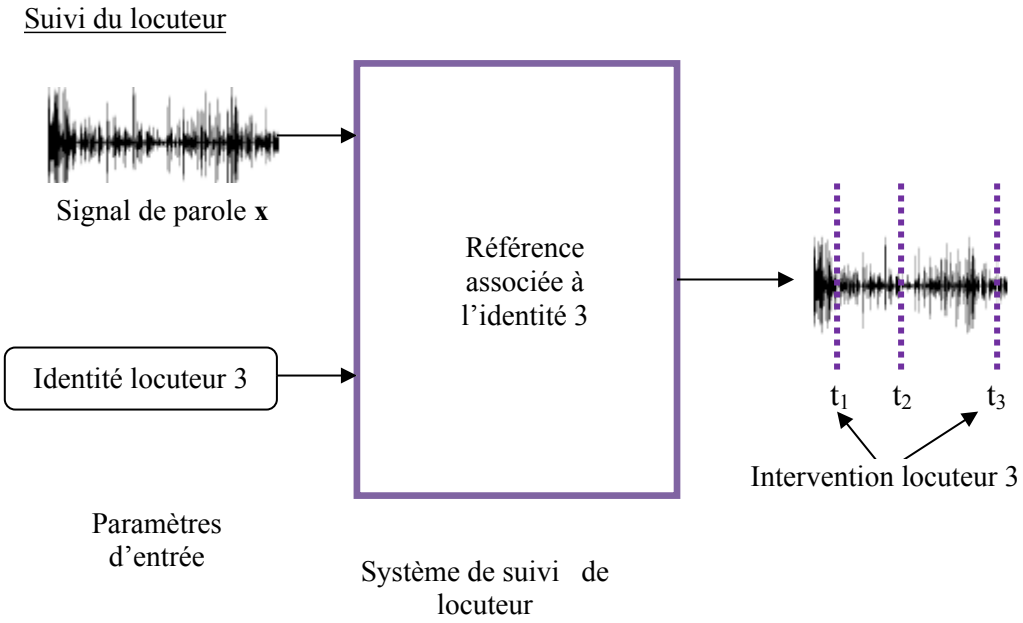


Figure I.6 : La tâche de suivi de locuteurs.

- Une segmentation"aveugle" en locuteurs, identique à celle employée pour l'Indexation par Locuteur d'un flux audio, est appliquée sur le signal de test. Les segments - résultat de la segmentation - sont soumis à un système de VAL classique afin de déterminer les segments appartenant effectivement au locuteur cible [08].
- Le signal de test est découpé en une suite de blocs de trames, de taille fixe, sur lesquels sont appliqués à un système de VAL. Un processus de décision, à base de seuils, permet en phase finale d'accepter ou de rejeter les blocs appartenant au locuteur cible [11].
- La troisième approche est similaire à la précédente excepté le processus de décision. Dans ce cas, la décision repose sur un HMM ergodique composé d'états

correspondant au locuteur cible, à un modèle générique de parole et à un modèle générique de non parole (silence, bruit...) [15, 19].

Les systèmes d'Indexation Automatique par Locuteur d'un flux audio sont principalement utilisés pour le traitement de bases de données audio (recherche de séquences d'émissions télévisées ou radiophoniques par le suivi du présentateur, estimation du temps de parole de chaque intervenant lors de débats, etc.). D'autres applications sont envisageables comme la recherche de messages par locuteur sur un répondeur téléphonique.

I.6 Structure d'un système de RAL

La reconnaissance automatique du locuteur peut être interprétée comme une tâche particulière de reconnaissance de formes. Différents modules sont présents dans ce système (voir figure I.7). Tout d'abord, le message vocal, capté par un microphone, est converti en signal numérique. Il est ensuite analysé dans un étage d'analyse acoustique. A l'issue de cette étape, le signal est représenté par des vecteurs de coefficients pertinents pour la modélisation du locuteur. Dans l'étape d'apprentissage, on crée un modèle du locuteur. A la reconnaissance, un module de reconnaissance va mesurer la similarité entre les paramètres acoustiques du signal prononcé et les modèles de locuteurs présents dans la base. En dernier lieu, un module de décision, basé sur une stratégie de décision donnée, fournit la réponse du système. On peut également introduire un module d'adaptation pour augmenter les performances du système de reconnaissance.

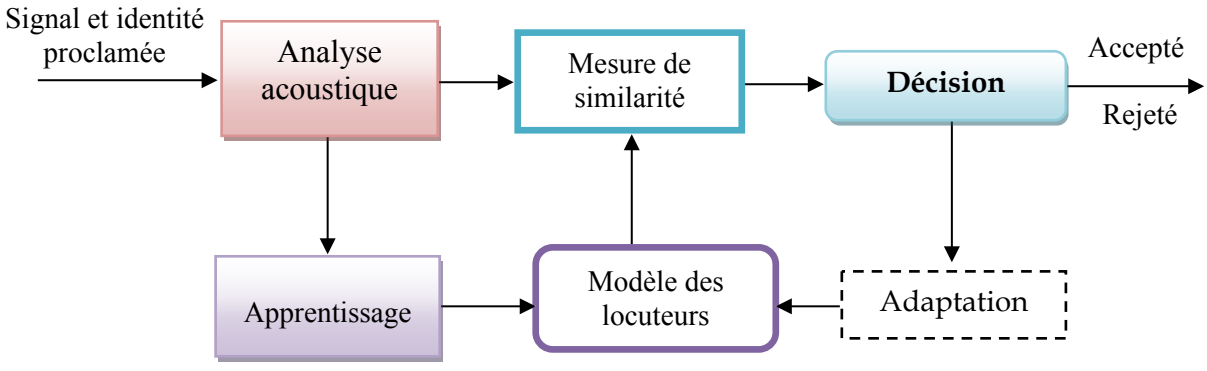


Figure I.7 : Schéma typique d'un système de reconnaissance automatique.

I.7 Analyse acoustique du signal de parole

Le signal de parole, de par sa complexité (multitudes d'informations et redondance), ne peut être exploité directement. Une représentation simplifiée de celui-ci est par conséquent nécessaire. Le processus de paramétrisation consiste à extraire du signal de parole les informations pertinentes en vue de la reconnaissance. Cette représentation repose généralement sur des vecteurs de paramètres acoustiques, calculés périodiquement sur le signal de parole. On considère généralement trois grandes classes de paramètres, qui sont les paramètres de l'analyse spectrale, les paramètres prosodiques et les paramètres dynamiques [08].

I.7.1 Paramètres prosodiques

Le terme "paramètres prosodiques" réunit l'énergie et la fréquence fondamentale (ou pitch). Ces paramètres s'avèrent cependant fragiles en pratique et ne permettent pas, à eux seuls, de discriminer les locuteurs. En conséquence, ils sont souvent associés à d'autres paramètres tels que ceux de l'analyse spectrale (surtout l'énergie).

I.7.1.1 Energie totale

L'amplitude du signal de la parole varie au cours du temps selon le type de son. En particulier, l'amplitude des segments non voisés est généralement plus faible que celle des segments voisés. L'énergie à court terme du signal de la parole fournit une représentation convenable qui reflète ces variations d'amplitude [20]. Elle est calculée à partir de la relation suivante :

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

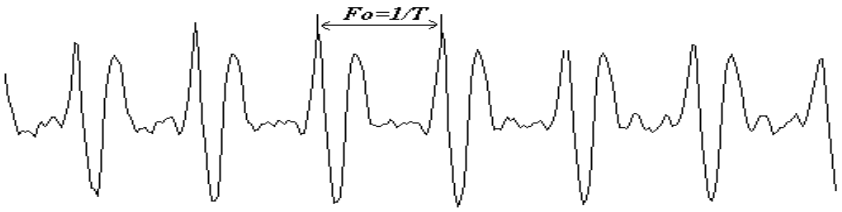
Avec E **:** La valeur à évaluer.

N**:** La largeur de la fenêtre d'analyse.

$x(k)$**:** Le signal numérique.

I.7.1.2 Fréquence fondamental

La période du fondamental est par définition la fréquence de vibration des cordes vocales, elle est appelée aussi le pitch figure I.8. C'est un paramètre très important dans les différentes applications de la parole.



.

Figure I.8 : La fréquence fondamentale.

L'extraction du pitch est une tâche particulièrement difficile pour trois raisons :

- ✓ La vibration des cordes vocales n'a pas nécessairement une périodicité complète.
- ✓ Il est difficile de séparer le pitch des effets du trait vocal.
- ✓ La plage de dynamique de la fréquence du fondamental est très grande. (Elle s'étend approximativement de : 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes et de 200 à 600 Hz chez les enfants) [21].

I.7.2 Analyse spectrale du signal de parole

L'analyse spectrale de parole présente des avantages au niveau de la perception, car l'oreille humaine effectue ce genre d'analyse. De plus, celle-ci fait apparaître des propriétés et des paramètres (formants) auxquelles on attache une grande importance. Pour cela on introduit en traitement du signal les outils suivants :

I.7.2.1 La transformée de Fourier discrète :

La transformée de Fourier $S(f)$ d'un signal réel continu $s(t)$ est donnée par :

$$S(f) = \int_{-\infty}^{+\infty} s(t) e^{-j2\pi ft} dt \qquad \text{(I.2)}$$

Par cette transformation, $s(t)$ est remplacé de façon biunivoque par son spectre complexe $S(f)$. Cette transformation est réversible et $s(t)$ peut être déterminée de façon unique par la transformée de Fourier inverse :

$$s(f) = \int_{-\infty}^{+\infty} S(f) e^{j2\pi ft} df \qquad \text{(I.3)}$$

La transformation de Fourier fournit ainsi deux descriptions duales, temporelle et fréquentielle, d'un signal. La transformation de Fourier permet ainsi de décrire un signal continu dans l'espace des fréquences. Ceci peut être étendu au cas des signaux discrets. A un tel signal $s(n)$ on associe sa Transformée de Fourier Discrète (TFD) définie par :

$$S(f) = \sum_{n=0}^{N-1} s(n) e^{-j2\pi fn} \quad \text{(I.4)}$$

Le calcul d'une TFD sur N point requiert de l'ordre de N^2 opérations complexes. De nombreux travaux ont été menés de façon à réduire la complexité de ce calcul. Le principe de base est en général de calculer un ensemble de TFD pour des valeurs inférieures à N puis à combiner les résultats, accélérant ainsi le calcule d'une TFD sur un processeur standard. Ces algorithmes de calcul sont regroupés sous le terme générique de transformation de Fourier rapide, FFT (Fast Fourier Transform en anglais) [22]. La FFT joue un rôle central pour l'analyse du signal parole comme pour beaucoup d'autres signaux et intervient dans les méthodes présentées dans ce chapitre.

I.7.2.2 La transformée de Fourier à courte terme

Le signal de parole étant par essence non stationnaire, la nécessité d'une analyse tempes-fréquence a été reconnue de langue date. La solution la plus couramment utilisée en traitement de signal de parole est de calculer des spectres de Fourier à court terme parfois dénommés spectres instantanés. Un spectre à court terme est le résultat d'une analyse de Fourier local sur une portion de signal de faible durée, limitée par une fenêtre temporelle h (par exemple fenêtre de Hamming) pendant laquelle le signal est quasi stationnaire, soit de 32 millisecondes pour le signal de parole considéré dans cette étude. Dans le cas d'un signal discret, le spectre à court terme peut s'écrire pour une fenêtre centrée sur m : (m représente le même échantillon du signal discrétisé)

$$S_N(f) = \sum_n s_m(n) \mathrm{h}(n - m) e^{-j2\pi fn} \quad \text{(I.5)}$$

La qualité de l'analyse dépend largement du nombre de points N du signal considérés dans la fenêtre h. Plus le nombre de points augmente, plus la résolution fréquentielle est grande. En faisant glisser la fenêtre d'observation et en concaténant les spectres à court terme successifs on forme un spectrogramme qui rend compte de l'évolution temps-fréquence du signal [22].

I.7.2.3 Les Coefficients MFCC (Mel Frequency Cepstral Coefficients)

La paramétrisation MFCC (*Mel Frequency Cepstral Coefficients)*, est la plus utilisée en reconnaissance automatique de la parole. Cette méthode d'extraction compresse les vecteurs acoustiques décrivant le signal, tout en conservant l'information utile à la

reconnaissance automatique. Cette méthode s'appuie sur le comportement de l'oreille humaine; en effet, si nous sommes capables de distinguer des sons de 100 Hz et de 150 Hz, cela n'est plus possible pour des sons de 4000 Hz et 4050 Hz : notre résolution fréquentielle n'est pas la même selon la fréquence considérée, donc l'oreille humaine ne répond pas également à toutes les fréquences. La figure I.9 présente le champ auditif humain, délimité par la courbe de seuil de l'audition. Sa limite supérieure en fréquence est d'environ 16000 Hz (variable selon les individus).

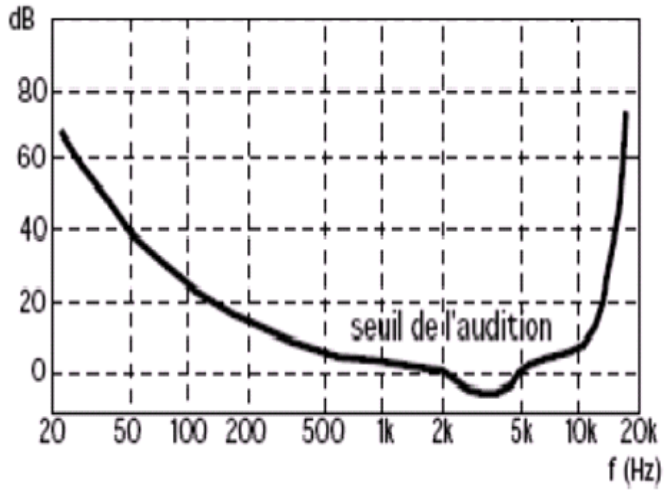


Figure I.9 : Le champ auditif humain.

A l'intérieur de son domaine d'audition, l'oreille ne présente pas une sensibilité identique à toutes les fréquences. La courbe révèle un maximum de sensibilité dans la plage [500 Hz, 10 kHz], en dehors de laquelle les sons doivent être plus intenses pour être perçus.

- **L'échelle Mel :**

L'échelle Mel est caractérisée par le fait que l'espacement sur l'axe des fréquences est linéaire pour les fréquences inférieures à 1KHz alors qu'il est logarithmique pour le reste des fréquences (supérieures à 1khz). Nous pouvons donc utiliser la formule approximative suivante afin de faire correspondre à chaque fréquence en Hz une fréquence sur l'échelle Mel [23]:

$$f_{mel} = 2595 \times \log_{10}(1 + \frac{FHz}{700}) \quad \text{(I.6)}$$

- **L'utilité de l'échelle Mel [23]**

Des études psychophysiques ont montré que la perception humaine ne suit pas une échelle linéaire dans le domaine fréquentiel. Comme il est montré sur la figure I.10. L'échelle Mel permet donc de modéliser une perception de l'oreille linéairement. On remarque qu'avant 1000 Hz, la courbe est à peu près droite, ce qui traduit bien l'équivalence entre Hz et Mels à ces fréquences.

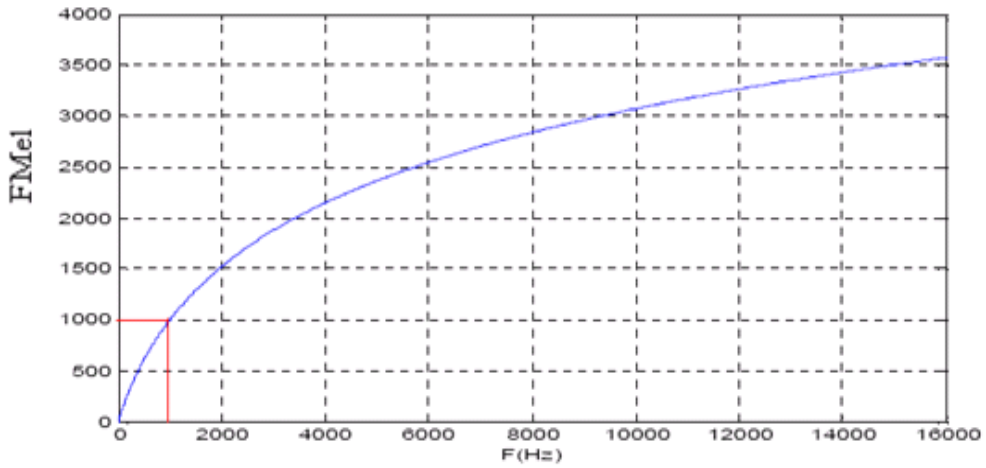


Figure I.10 : La transformation du Hz en Mel.

- **Calcul des coefficients cepstraux (MFCC) [24]**

Les différentes étapes pour calculer les coefficients MFCCs d'un signal parole sont : le découpage du signal en trames, le fenêtrage, le calcul de la FFT, le filtrage Mel, le calcul du logarithme de l'énergie et la transformée en cosinus discrète (figure I.11).

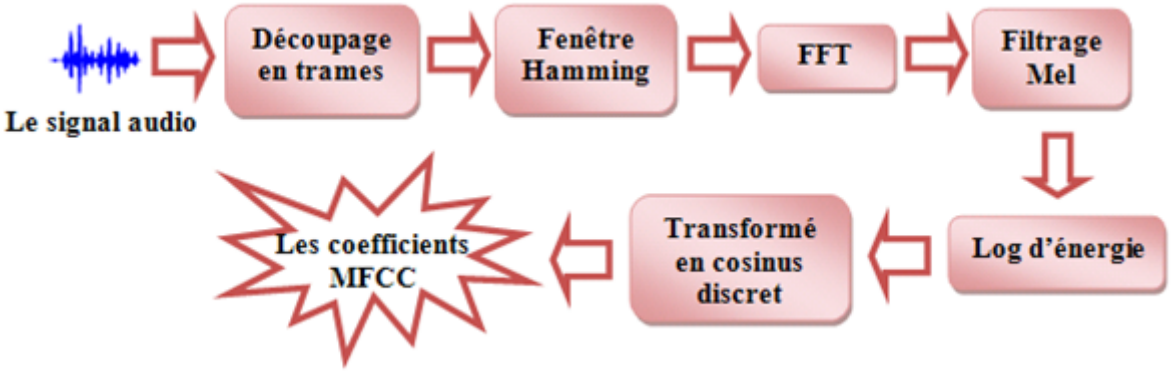


Figure I.11 : Calcul des coefficients MFCC avec une échelle Mel.

Après son découpage en trames, le signal subit une analyse par une fenêtre glissante de durée courte de 32 ms avec un recouvrement de 50% pendant lesquelles le signal parole

peut être considéré comme un signal quasi stationnaire. On utilise une fenêtre de Hamming plutôt qu'une fenêtre rectangulaire pour effiler le signal original des cotés et d'éviter la déformation du spectre liée aux effets de bord durant la transformation du domaine temporel au domaine fréquentiel.

Après le découpage en trames et le fenêtrage du signal parole, la transformée de Fourier est calculée pour chaque trame pour obtenir le spectre du signal. Le spectre présente beaucoup de fluctuations. L'intérêt est porté seulement sur l'enveloppe du spectre. Une autre raison de lisser le spectre est la réduction de la taille des vecteurs spectraux. Pour réaliser ceci, nous multiplions le spectre précédemment obtenu par un banc de filtres tenant compte de la réponse acoustique de l'oreille humaine. Un banc de filtre est une série de filtres, dont la forme est définie par la localisation des fréquences gauche, centrale et droite de chaque filtre. Les filtres utilisés sont triangulaires comme le montre la figure I.12. La localisation des fréquences centrales des filtres est donnée par [24]:

$$f_{mel} = 1000 \times \frac{\log(1 + f/1000)}{log2} \quad \text{(I.7)}$$

Où f est la fréquence en Hz.

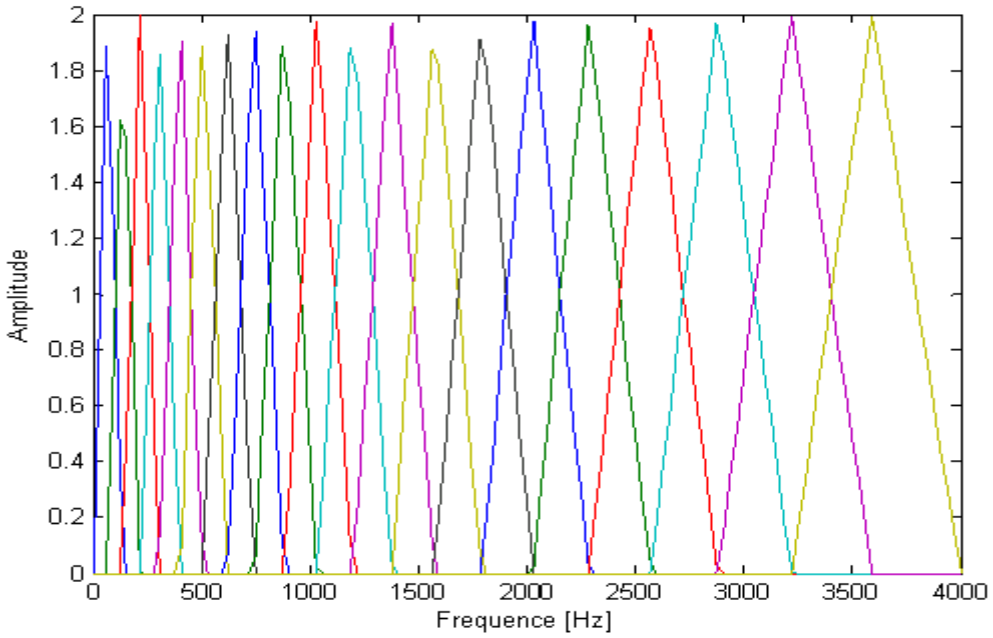


Figure I.12 : Banc de Filtres Triangulaires équidistance en échelle Mel.

Finalement, nous prenons le logarithme de cette enveloppe spectrale et nous multiplions chaque coefficient par 20 afin d'obtenir l'enveloppe spectrale en *dB*. Ensuite, les coefficients cepstraux sont obtenus par une transformée en cosinus discrète à partir des logarithmes des énergies issues du banc de filtres. L'avantage de la transformation cepstrale est de fournir des coefficients peu corrélés [25, 26], l'expression de ces coefficients est données par :

$$c_n = \sum_{k=1}^{K} S_k cos\left[n(k - \frac{1}{2})\frac{\pi}{K}\right], \quad n = 1,2\ldots,L \tag{I.8}$$

Où K est le nombre de coefficients spectraux calculés précédemment, S_k sont les coefficients spectraux, et L est le nombre de coefficients cepstraux que nous voulons calculer $(L \leq K)$ Finalement, nous obtenons des vecteurs cepstraux pour chaque fenêtre.

I.7.3 Paramètres exploitant la dynamique du signal de parole

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 sujette à des problèmes d'estimation des modèles lors de l'apprentissage. La seconde possibilité consiste à calculer les dérivées (Δ) des paramètres instantanés [20, 27].

Pour simplifier le calcul, une approximation des dérivées première et seconde est généralement obtenue à l'aide de fonctions polynomiales comme le montre l'équation II.7 pour le calcul des coefficients issus de la dérivée première (coefficients Delta). Cette même équation sera appliquée sur les coefficients Delta afin d'obtenir les coefficients issus de la dérivée seconde (coefficients Delta- Delta) [08, 37].

$$\frac{\partial c_m(t)}{\partial t} \approx \Delta c_m(t) = \frac{\sum_{n=-N}^{N} n c_m(t+n)}{\sum_{n=-N}^{N} n^2} \tag{I.9}$$

Où $c_m(t)$ représente le coefficient à dériver, $\Delta c_m(t)$ désigne le coefficient Delta et N est relatif à la taille de la fenêtre temporelle de longueur $2N + 1$ trames sur laquelle les coefficients dérivés sont calculés.

I.8 Modélisation des locuteurs

Ce paragraphe parcourt les techniques les plus couramment utilisées en reconnaissance du locuteur. Comme dans le cas de reconnaissance du la parole, le problème de reconnaissance du locuteur peut se ramener à un problème de classification. Différentes approches ont été développées, néanmoins on peut les classer en quatre grandes familles :

- L'approche vectorielle où le locuteur est représenté par un ensemble de vecteurs de paramètres (MFCC… etc.) dans l'espace acoustique. Ses principales techniques sont la reconnaissance à base de DTW et la reconnaissance par quantification vectorielle.
- L'approche statistique qui consiste à représenter chaque locuteur par une densité de probabilité dans l'espace des paramètres acoustiques. Elle couvre les techniques de modélisation par les modèles de Markov cachés, par les mélanges de gaussiennes et par des mesures statistiques du second ordre.
- L'approche connexionniste qui consiste, principalement, à modéliser les locuteurs par des réseaux de neurones.
- L'approche prédictive.
- L'approche discriminante

I.8.1 L'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 (MFCC, PLP… etc.) issus des signaux de test est calculée. L'approche vectorielle compte deux grandes techniques : la programmation dynamique et la quantification vectorielle.

- **Reconnaissance du locuteur à base de DTW**

La programmation dynamique (Dynamic Time Warping : DTW) consiste à aligner temporellement une séquence de vecteurs de paramètres (MFCC, PLP… etc.) 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 vecteurs de test et moyennée sur l'ensemble de la séquence.

De par son principe, la programmation dynamique est utilisée exclusivement en mode dépendant du texte. Très rapide et montrant des performances relativement bonnes, la programmation dynamique est toutefois très sensible à la qualité d'alignement et notamment au choix du point de départ [27, 28].

- **Quantification vectorielle**

Il s'agit de représenter l'espace acoustique par un nombre fini de vecteurs acoustiques. Cela consiste à faire un partitionnement de cet espace en régions, qui seront représentées par leur vecteur centroïde [29]. Pour déterminer la distance d'un vecteur acoustique à cet espace, on effectue une mesure de distance avec chacun des centroïdes des régions et on retient la distance minimale. Si le vecteur acoustique provient du même locuteur pour lequel on a établit le dictionnaire de quantification, la distance sera en général moins grande que si ce vecteur provient d'un autre locuteur. Ainsi, on va représenter un locuteur par son dictionnaire de quantification [30].

La quantification vectorielle s'applique en mode dépendant ou mode indépendant du texte. La rapidité et les performances de cette technique dépendent fortement de la taille du dictionnaire de quantification [08].

I.8.2 L'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 seul modèle des locuteurs [08]. 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.

- **Modèles de Markov cachés**

Les modèles de Markov (ou HMM pour *Hidden Markov Models*) ont été initialement introduits en reconnaissance de la parole. Puis, leur utilisation s'est étendue peu à peu au domaine de la reconnaissance du locuteur. Dans cette approche, où ne s'intéresse pas à mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme acoustique ait été engendrée par le modèle du locuteur. Le modèle d'un locuteur est constitué de l'association d'une chaîne de Markov, une succession d'états avec des

probabilités de transition d'un état à l'autre, et de lois de probabilités (probabilités d'observation d'un vecteur acoustique dans un état).

De par leur principe, les modèles de Markov cachés s'appliquent parfaitement au mode dépendant du texte [31].

- **Les mélanges de gaussiennes**

La reconnaissance du locuteur par mélanges de gaussiennes (ou GMM pour Gaussian Mixture Models) consiste à modéliser un locuteur par une somme pondérée de composantes gaussiennes. Ainsi une large gamme de distributions peut être parfaitement représentée. Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques. Les mélanges de gaussiennes est considéré comme un cas particulier des HMM et une extension de la quantification vectorielle [31, 32]. Nous détaillons cette méthode dans le chapitre trois.

- **Le model GMM-UBM**

Les approches génératives utilisées en reconnaissance du locuteur reposent essentiellement sur le paradigme GMM-UBM (figure I.13). Ce paradigme consiste à estimer le modèle GMM d'un locuteur en adaptant le modèle du monde UBM avec les données de ce locuteur. Différents critères d'adaptation existent dans la littérature. La méthode la plus utilisée en reconnaissance du locuteur est celle du Maximum a Posteriori (MAP) [33], [34]. Lors du test de vérification, le calcul de score fait intervenir l'UBM et le modèle correspondant à l'identité proclamée. La décision rejet ou accès est prise par rapport à ce score.

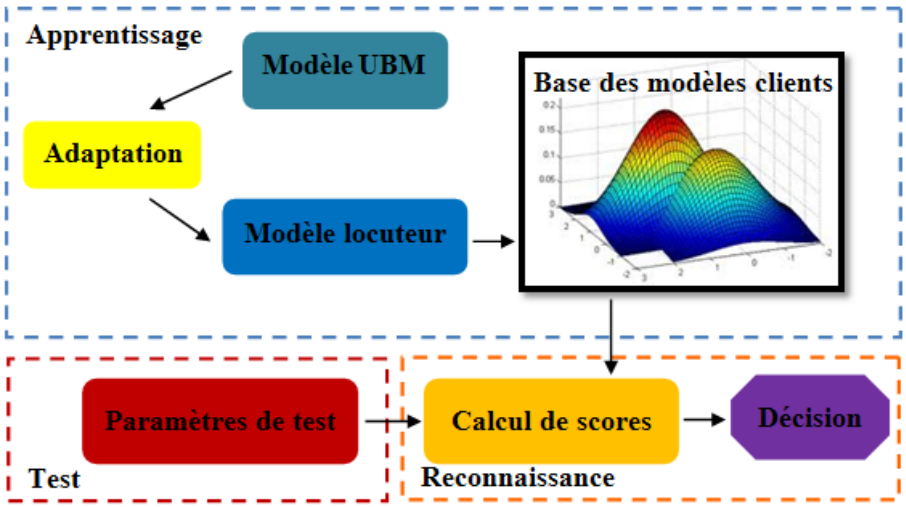


Figure I.13 : Structure du paradigme GMM-UBM en RAL.

I.8.3 L'approche connexionniste

Les réseaux de neurones ont été assez largement utilisés en reconnaissance du locuteur. Ils offrent en effet une bonne alternative au problème de la discrimination entre les locuteurs. Ces outils de classification permettent de séparer des classes, dans un espace de représentation donné, de façon non linéaire. L'inconvénient important de l'application de cette technique en identification du locuteur est le coût important lié à l'ajout d'un nouveau locuteur dans la base de référence (ce n'est pas le cas en vérification du locuteur). On peut aussi utiliser les réseaux de neurones en les couplant à d'autres techniques, comme par exemple les modèles de Markov cachés. On parle alors de méthodes hybrides [35].

I.8.4 L'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. L'erreur de prédiction moyenne constitue alors la mesure de similarité entre le signal de test et le modèle de locuteur (fonction de prédiction). Une autre

solution envisageable est d'estimer une fonction de prédiction sur la séquence de test et de la comparer, à l'aide d'une distance, à la fonction de prédiction estimée lors de l'apprentissage [36].

I.8.5 L'approche discriminante

La plus employée en RAL sont les Support Vector Machine (SVM) [37]. 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. Cette approche donne aujourd'hui des performances similaires μa l'approche GMM. Ces deux méthodes sont aussi combinées dans un nouveau formalisme, le GMM/SVM Super-Vecteur [38] qui profite des capacités génératives du GMM et discriminantes du SVM.

I.9 Prise de décision

Au delà de la structure commune entre les différentes tâches d'un système de RAL, la stratégie de décision est différente selon la tâche choisie. Le plus souvent, cette stratégie se formalise dans un cadre bayésien que nous présentons ci-dessous pour les tâches de vérification et d'identification.

I.9.1 Décision en identification

En identification, un signal de test est comparé à toutes les références des locuteurs connus du système, résultant en un ensemble de mesure de similarité (ou un ensemble de mesure de distance) à l'entrée du processus de décision. Aussi, la règle de décision consiste à choisir le locuteur dont la mesure de similarité est maximale (ou minimale dans le cas de mesure de distance). Pour l'évaluation des performances du système d'identification du locuteur, le taux de classification correcte est souvent utilisé. Ce taux est le rapport entre le nombre des segments correctement identifies et le nombre total des segments de test.

$$Taux\ d'identification\ correct\ \% = \frac{Tests\ \ correctement\ identifiés}{Tests\ \ totales} \qquad (I.10)$$

I.9.2 Décision en vérification

En vérification, le processus de décision consiste à comparer la mesure de similarité entre le signal de test et le modèle du locuteur proclamé à un seuil de décision. Celui ci,

accepte l'identité proclamée si la mesure est supérieure au seuil de décision et la rejette au cas contraire. Pour la mesure des performances, l'erreur de fausse acceptation (ou le système accepte le locuteur test alors qu'il s'agit d'un imposteur) et l'erreur de faux rejet (le système rejette le locuteur test alors qu'il s'agit bien du locuteur proclamé) sont souvent utilisées. Ces deux grandeurs sont données par les formulations suivantes :

$$p(FA) = \frac{Tests\ ayant\ amen\acute{e}\ \grave{a}\ une\ fausse\ acceptation}{Tests\ imposteurs} \quad \text{(I.11)}$$

$$p(FR) = \frac{Tests\ ayant\ amen\acute{e}\ \grave{a}\ un\ faux\ rejet}{Tests\ clients} \quad \text{(I.12)}$$

I.10 Mesures de performances

Les performances d'un système de VAL s'évaluent en fonction de deux taux d'erreurs. La probabilité de faux rejets (FR) ou de rejet du client à l'identité proclamée et la probabilité de fausses acceptations (FA) ou d'acceptations d'impostures. Ces taux sont étroitement liés. Au point de fonctionnement, pour un certain seuil de vérification, ces deux taux sont définis. En fonction du type d'application souhaitée, le seuil de vérification peut être choisi pour minimiser le taux de fausses acceptations : application de sécurité, ou minimiser le taux de faux rejets pour augmenter l'ergonomie d'utilisation. Il n'est pas possible de minimiser conjointement ces deux taux (figure I.14).

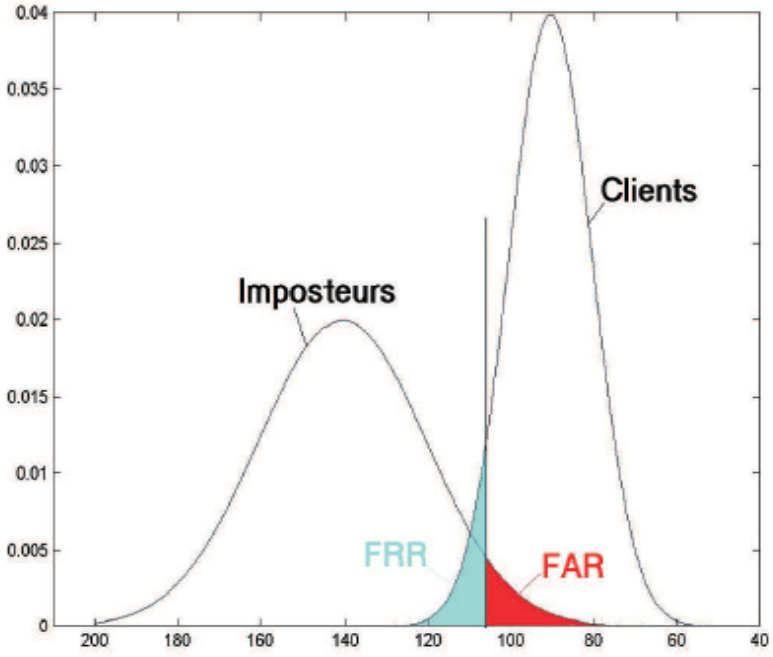


Figure I.14 : Types d'erreurs dans un système RAL.

I.11 Les courbes DET (Detection Error Tradeoff)

La représentation la plus communément utilisée pour évaluer la pertinence du seuil de décision en fonction de ces deux taux d'erreurs est la courbe DET (Detection Error Tradeoff [39]) figure I.15. Les échelles des axes suivent la répartition d'une loi normale. L'échelle logarithmique est utilisée pour rendre la courbe DET linéaire quand les scores des systèmes suivent une distribution Gaussienne. La courbe DET permet d'évaluer, pour chaque seuil de vérification, les valeurs du couple (FA, FR). La figure I.15 illustre un exemple de courbe DET.

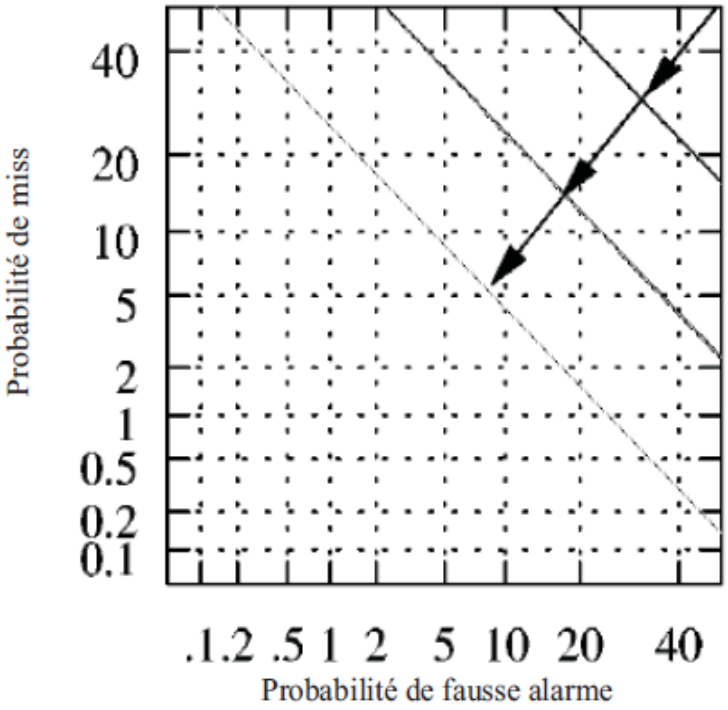


Figure I.15 : Exemple courbe DET.

I.12 Conclusion

Ce chapitre est une introduction au domaine de la reconnaissance automatique du locuteur. Il présente les différentes tâches liées à la RAL telles que l'Identification, la Vérification Automatique du Locuteur et les tâches plus récentes comme le suivi de locuteur où l'Indexation par Locuteur de flux audio. Les diverses applications et les problèmes liés à l'exploitation de la RAL sont aussi exposés, comme la variabilité intra locuteur où la variabilité due au matériel

Un système de la reconnaissance automatique du locuteur, quelle que soit la tâche considérée, se résume à trois étapes principales qui sont :

- l'analyse acoustique du signal de parole ;
- la modélisation du locuteur ;
- la décision.

Dans ce chapitre, nous avons décri les différentes classes des paramètres de l'analyse acoustique (les paramètres prosodiques, les paramètres d'analyse spectrale et les paramètres exploitant la dynamique du signal de parole). Ces paramètres, une fois calculés, seront utilisés dans les deux autres étapes.

Chapitre II

La reconnaissance automatique du locuteur sur IP

II.1 Introduction

Le développement de la VoIP, et par conséquent de la téléphonie sur IP, a ouvert de nouveaux horizons aux applications en reconnaissance vocale, comme l'intégration du système de la reconnaissance sur le réseau IP. Un tel système exige une reconnaissance vocale fiable pour les différents systèmes de reconnaissance qui sont distribués à travers le réseau. Néanmoins, les technologies de la reconnaissance vocale qui sont maintenant disponibles doivent être améliorées, parce qu'ils ne sont pas tout à fait capables de fonctionner dans des conditions difficiles imposées par les réseaux IP. L'introduction de la RAL dans de tels systèmes nous amène à la reconnaissance distribuée (Distributed Speech and Speaker Recognition : DSR).

Ce chapitre est consacré à la présentation de la reconnaissance automatique du locuteur distribuée sur IP (DSR) et à la description de l'architecteur client-serveur.

II.2 Système distribué

Il existe moult définitions d'un system sur IP (Distribué). Selon Tanenbaum [40]: « A distributed system is a collection of independent computers that appears to its users as a single coherent system ». Un système distribué (réparti) est un ensemble d'ordinateurs (ou processus) indépendants qui apparaît à un utilisateur comme un seul système cohérent. Une autre définition est proposée par Coulouris et ses collègues [41]: « We define a distributed system as one in which hardware or software components located at networked computers communicate and coordinate their actions only by passing messages ». Cette définition introduit la notion générique de composant qui peut représenter aussi bien des éléments logiciels que des éléments matériels. La collaboration entre ces différents éléments apparaît comme le résultat de communications et de coordinations basées sur l'unique principe d'échange de messages. Cette définition précise que les composants, logiciels ou matériels, appartiennent à un même réseau informatique. Du fait que l'ensemble des ordinateurs forment un système en entier, la défaillance d'un ordinateur peut avoir un impact négatif le fonctionnement du système et introduire des incohérences.

De manière générale, on peut dire que le système distribué est l'ensemble d'ordinateurs indépendants connectés en réseau et communiquant via ce réseau Cet ensemble apparaît du point de vue de l'utilisateur comme une unique entité. Les principaux objectifs des systèmes distribués sont de faire coopérer plusieurs ressources dans l'optique de partager des tâches, de faire des traitements parallèles, etc. Ainsi, un système distribué peut être vu comme

une application qui coordonne les tâches de plusieurs équipements informatiques (ordinateurs, téléphones mobile, PDA, capteurs...). Cette coordination se fait le plus souvent par envoi de messages via un réseau de communication qui peut être un LAN (Local Area Network), WAN (Wide Area Network), Internet, etc.

II.3 La reconnaissance automatique de locuteur distribuée sur IP (DSR)

La RAL distribuée (DSR : Distributed Speech/Speaker Recognition) est un système qui offre la possibilité de diviser les tâches de reconnaissance automatique de locuteur sur IP entre les machines clientes et serveurs. Les travaux du groupe STQ Aurora (Speech Processing, Transmission and Quality Aspects) de l'ETSI (European Télécommunications Standards Institute) ont donné lieu au premier standard DSR ES 201 108 [42], publié par ETSI en Février 2000, suivie par DSR ES 202 050 [43] en Octobre 2002. La deuxième norme est une version améliorée de la première et concerne aussi la réduction du bruit. Cette normalisation Frond-end était non seulement nécessaire pour des raisons d'efficacité et de robustesse, mais aussi pour permettre aux serveurs de réseaux de fournir un support de reconnaissance vocale indépendamment du type de client qui demande le service. Dans un tel système, on distingue deux concepts :

- Au niveau du premier concept, le bloc d'extraction de paramètres acoustiques (au niveau du client) est séparé du reste de bloc de RAL et il est installé au client (front-end). Ensuite, les données sont envoyées vers le coté serveur (back-end) [44], [45] pour gérer le processus de la reconnaissance automatique de locuteur comme illustré dans la Figure II.1. Ce système découple entièrement l'étage de traitement et de paramétrisation acoustique du reste de l'unité de reconnaissance automatique de locuteur, en utilisant une architecture client-serveur sur un réseau de communication. Cette architecture divise la reconnaissance automatique de locuteur en deux étapes. La première étape est effectuée coté client (front-end) où la sortie est représentée par les vecteurs de paramètres acoustiques. La reconnaissance se déroule coté serveur (back-end), après avoir reçu les données transmises par le client.

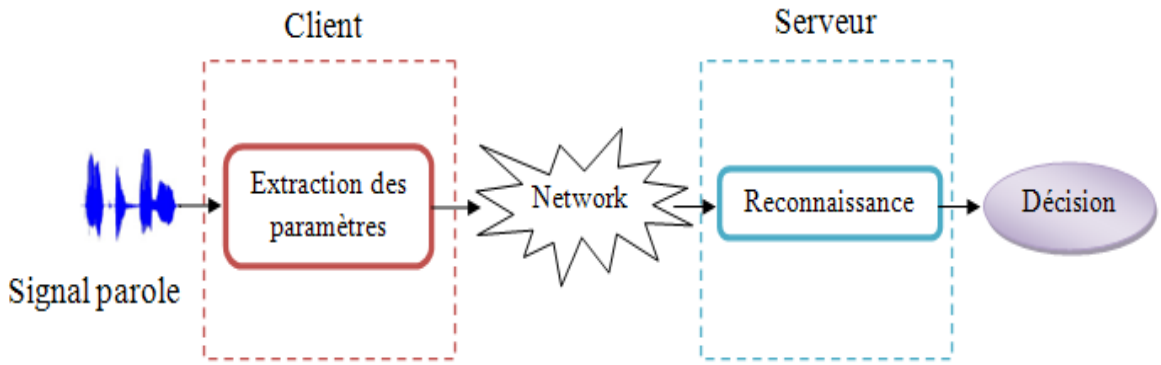


Figure II.1 : Schéma de principe d'un système DSR basé sur la transmission des vecteurs des paramètres du côté client (front-end) au coté serveur (back-end) pour faire la reconnaissance.

- Le deuxième concept de la reconnaissance distribuée est basé sur la parole resynthétisée par un codec de la voix. Le codeur hébergé sur le client encode la parole puis envoie le bit-stream au décodeur hébergé sur le serveur, via un réseau IP. Le bit-stream reçu est décodé par le décodeur. Cette opération est suivie de l'extraction des paramètres à partir des éléments décompressés, puis le processus de la reconnaissance automatique du locuteur est appliqué comme indiqué sur la Figure II.2.

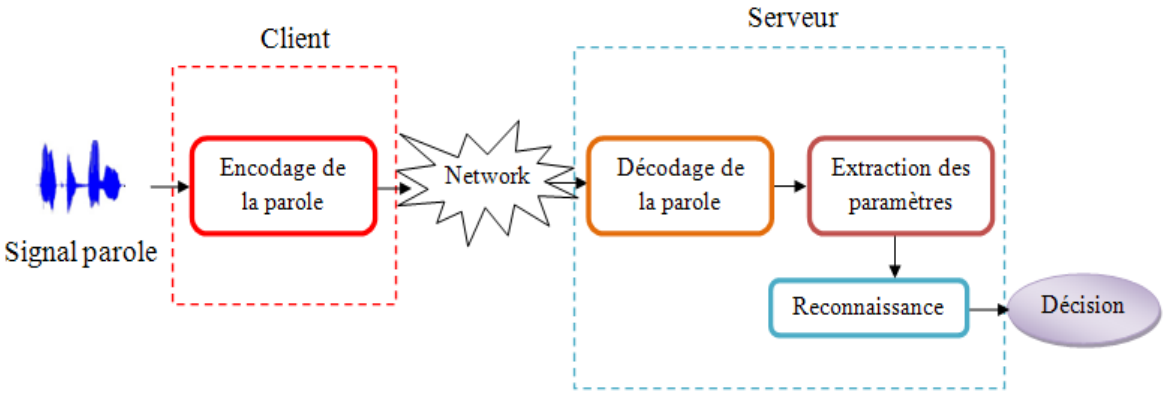


Figure II.2 : Schéma de principe d'un système DSR basé sur la transmission de la parole codée (bit-stream) du coté client au coté serveur où il faut décoder le bit-stream et resynthétiser la parole. A partir de celle ci, l'extraction des paramètres est effectuée en vue de la reconnaissance.

Une grande variété de techniques de codage ont été utilisées afin de minimiser la quantité de données envoyées à travers un réseau, tout en préservant la qualité de la parole [46]. Néanmoins, il a été constaté que les performances de reconnaissance peuvent être significativement dégradées lorsque la parole est reçue par ces réseaux [47]. Cela est dû principalement à deux sources de dégradation ; les distorsions occasionnées par le codage binaire à bas débit de la parole, mais aussi à l'erreur de transmission du canal.

La création d'environnements distribués nécessite, à un moment ou à un autre, d'implanter le modèle de communication qui permettra aux différentes machines et applications présentes, d'échanger des informations. Dans le paragraphe suivant, nous allons décrit le principe de fonctionnement de l'architecture client-serveur.

II.4 Architecture client-serveur

Le modèle le plus simple d'application distribué est le modèle client-serveur (voir figure II.3). Un serveur est un processus qui fonctionne en continu et en attente d'être contacté par un processus client. Un processus client initie le contact avec le serveur en se connectant à un port spécifié. Une architecture client-serveur se présente comme un ensemble de programmes clients et serveurs, situés le plus souvent à distance les uns des autres, et communiquant à travers un réseau. Il en résulte que dans la majorité des cas, les communications entre client et serveur suivent le modèle suivant : le client envoie une requête à destination du serveur et celui-ci répond à cette requête. Dans ce travail, l'envoi et la réception des requêtes reposant sur le mécanisme des Sockets [48]. Cette solution permet la transmission de données entre deux machines distinctes d'un réseau, à l'aide de primitives de bas-niveau. Elle offre l'avantage d'être supportée par la quasi-totalité des systèmes d'exploitation et langages de programmation.

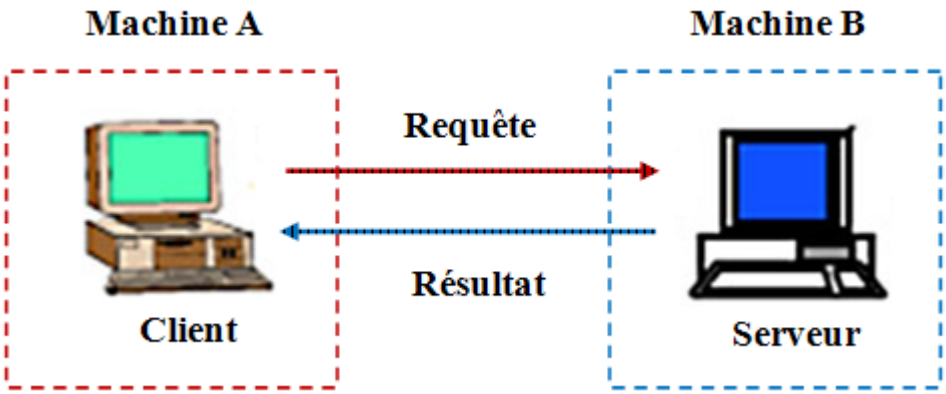


Figure II.3 : Interaction dans le modèle client-serveur.

II.5 Sockets

Les sockets sont une interface de programmation entre les applications et les couches réseau (figures II.4). Il s'agit d'une interface souple, d'assez bas niveau (couches 4 et 3 selon la norme modèle OSI, Annexe A). Le terme socket désigne à la fois une bibliothèque d'interface avec le réseau et l'extrémité d'un canal de communication bidirectionnel via lequel un processus peut émettre et recevoir des données.

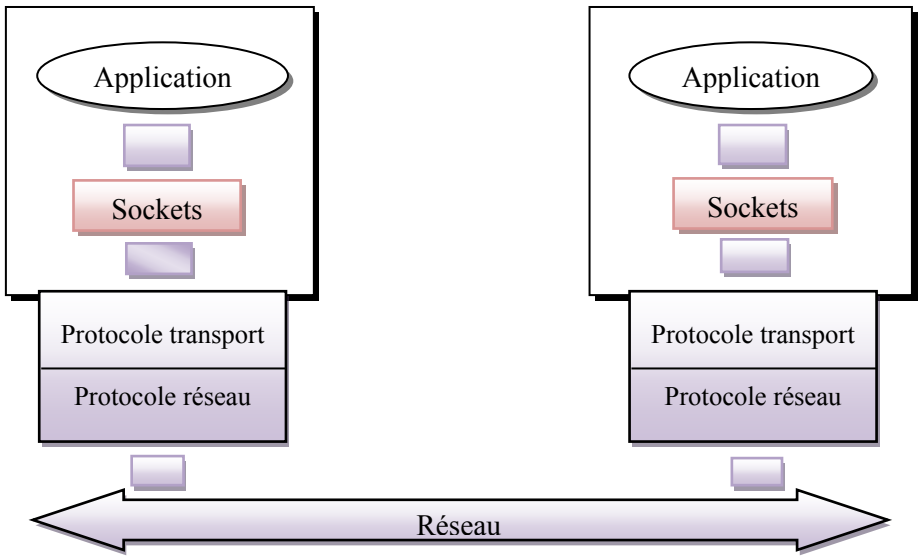


Figure II.4 : Modèle de la communication via socket.

Depuis que les interfaces réseau ont été standardisées autour de la couche IP, les sockets sont la brique de base utilisée par toutes les applications réparties sur un réseau. Ils consistent en une connexion plus ou moins explicite entre deux applications : l'une est le serveur, offrant une connexion disponible à l'autre, le client, qui s'y connecte en s'adressant à la bonne adresse IP et au bon port (figure II.5). L'adresse IP est une caractéristique ou un numéro qui permet d'identifier de manière unique un ordinateur sur le réseau Internet. Le numéro de port est un entier inférieur ou égal à 65536 qui correspond au processus d'application ou au service réseau.

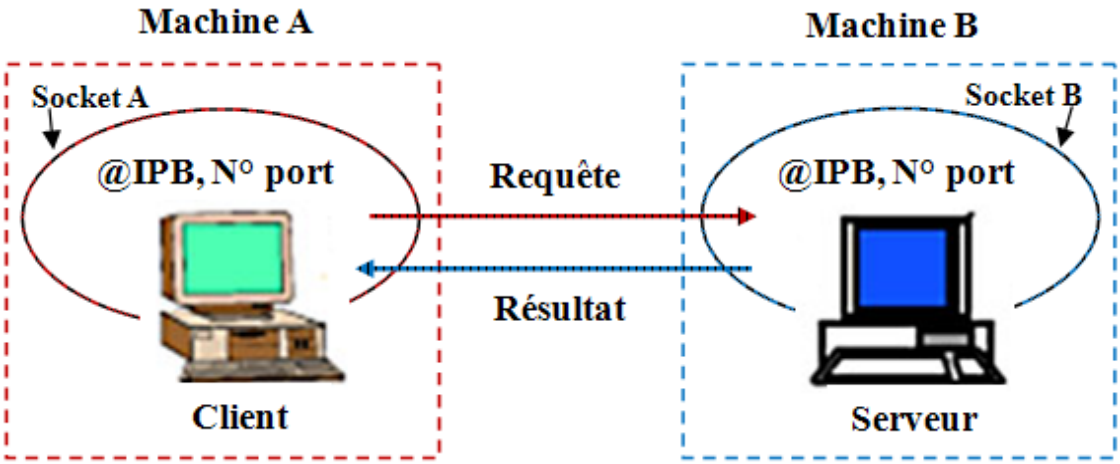


Figure II.5 : Schéma d'un socket.

Les connexions socket sont donc des connexions de type client-serveur, obligeant le client à connaitre l'adresse du serveur, et le port accessible. Les deux principaux modes de communications via une connexion socket sont TCP (appelé alors TCP/IP) et UDP [49]. TCP (pour Transmission Control Protocol) est un mode de communication connecté (flux de données). Le client ouvre une session avec le serveur, puis échange des données fiables (avec contrôle d'erreur), et la session prend fin soit de la volonté de l'un ou l'autre, soit au bout d'un temps limite (time out). UDP est dit non-connecté (ou datagramme) car il n'existe aucun concept de session. Le client envoie des paquets de données au serveur, qui peut les recevoir (ou non) dans l'ordre (ou non). Bien qu'étant moins fiable que TCP, UDP présente cependant le grand intérêt de fournir une rapidité de transmission supérieure. Ce type de communication est donc recommandé pour des applications où la rapidité de transfert importe plus que l'exhaustivité des données (comme du streaming par exemple).

II.6 Protocoles réseau et transport

On présente ici les principaux protocoles de la couche transport d'Internet et de la couche réseau que sont les protocoles TCP (Transmission Control Protocol), UDP (User Datagram Protocol) et IP. Tous les deux utilisent IP comme couche réseau, mais TCP procure une couche de transport fiable (alors même qu'IP ne l'est pas), tandis qu'UDP ne fait que transporter de manière non fiable des datagrammes.

II.6.1 Le protocole IP

"Internet Protocol", que certains appellent Interworking, est le protocole réseau le plus répandu dans le monde, il est utilisé dans la majorité des réseaux et surtout dans le grand réseau Internet. Il permet de découper l'information à transmettre en paquets, de les adresser et de les transporter indépendamment les uns des autres via le réseau pour recomposer ensuite le message initial une fois arrivé à destination. Donc, IP est un protocole qui permet l'adressage des machines et le routage des paquets de données [50]. Il correspond à la couche 3 (réseau) de la hiérarchie des couches ISO [51] [52]. Son rôle est d'établir des communications sans connexion de bout en bout entre des réseaux, de délivrer des trames de données (datagramme), et de réaliser la fragmentation et le réassemblage des trames pour supporter des liaisons n'ayant pas la même MTU (Maximum Transmission Unit, c'est-à-dire la taille maximum d'un paquet de donnée). Un paquet IP est composé d'un entête et de données. La figure II.6 représente la structure de l'entête IP basé sur 20 octets.

<table>
<tr><td>Version</td><td>IHL</td><td>Type de service</td><td colspan="2">Longueur totale du datagramme</td></tr>
<tr><td colspan="3">Identification</td><td>Drapeaux</td><td>Décalage fragment</td></tr>
<tr><td colspan="2">Durée de vie</td><td>Protocole</td><td colspan="2">Somme de contrôle entête</td></tr>
<tr><td colspan="5">Adresse IP source</td></tr>
<tr><td colspan="5">Adresse IP destination</td></tr>
<tr><td colspan="5">Données IP</td></tr>
</table>

Figure II.6 : Structure de l'entête IP basé sur 20 octets [52].

Le protocole IP est souvent associé au protocole de contrôle de transmission de données TCP, on parle ainsi du protocole TCP/IP. En ce qui concerne son architecture, comme le présente la figure II.7, TCP/IP suit un modèle en couches légèrement différent du modèle OSI à 7 couches.

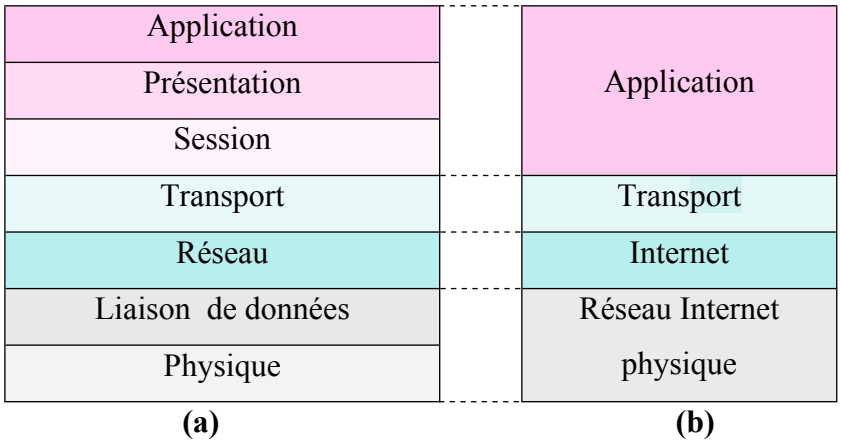


Figure II.7 : (a) : Modèle de référence OSI, (b) : Modèle TCP/IP (Internet) [53].

II.6.2 Le protocole TCP

TCP (Transmission Control Protocol) est un protocole de transport utilisé à grande échelle sur Internet. C'est un protocole de niveau 4 (transport) qui assure un transfert bidirectionnel de données, de façon fiable et sans erreur, avec contrôle et retransmission des données effectués aux extrémités de la liaison.

La principale caractéristique de TCP est qu'il est un protocole dit en mode connecté. Cela signifie qu'avant tout échange de données, une connexion entre les deux extrémités de la liaison doit être établie. Une fois réalisée, cette connexion, qualifiée de virtuelle, demeure existante jusqu'à terminaison explicite, et peut-être considérée comme un tube particulier offrant une communication sûre reliant les ports respectifs des deux extrémités. Cela s'oppose aux protocoles de transport sans connexion comme UDP.

Un segment TCP est encapsulé dans un paquet IP à la source, et n'est décapsulé puis analysé que lors de l'arrivée du paquet IP dans le nœud de destination : le contrôle se fait de bout-en-bout. Une fois le paquet reçu, la couche IP extrait le segment TCP (voir figure II.8) et le transfère à la couche transport (TCP) où l'entête est analysé.

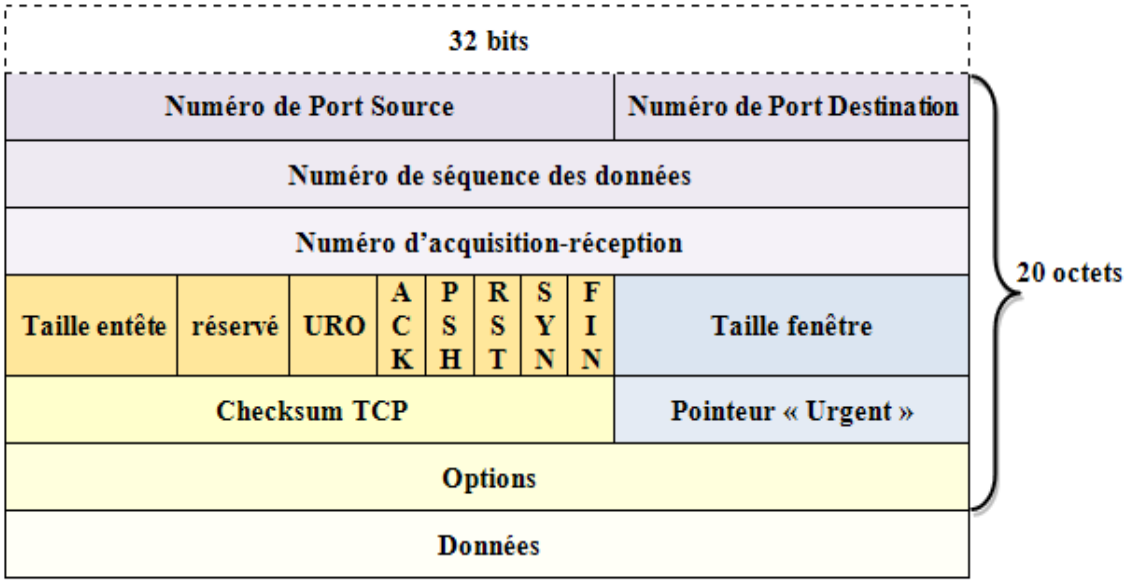


Figure II.8 : Structure d'un segment TCP [52].

Il existe plusieurs types de transfert de données. Par exemple du point de vue de la couche applicative, ils peuvent être interactifs ou non. Une connexion de type terminal (telnet) est interactive, tandis qu'une connexion de type transfert de données brutes (ftp) ne l'est pas. Des études ont montré qu'environ 10% du volume en octets, correspondant à la moitié du trafic en terme de segments TCP est du trafic interactif [50].

II.6.3 Le protocole UDP

UDP (User Datagram Protocol) est un protocole de communication sans connexion : cela signifie qu'il n'y a aucune garantie qu'une trame UDP émise arrive à destination. Les trames UDP sont encapsulées par la couche réseau IP sous-jacente. Elles peuvent être de longueur quelconque, la couche IP se charge de leur fragmentation et de leur réassemblage de façon transparente.

La qualité de service offerte par UDP est la même que celle fournie par IP. UDP ajoute cependant quelques informations importantes à l'entête, outre la longueur des données utiles et un checksum : ce sont les numéros de port source et de port destination (figure II.9). Ces numéros de port permettent de différencier plusieurs connexions ou services différents

entre deux extrémités identiques dans le réseau IP, afin de pouvoir multiplexer les connexions entre deux mêmes machines.

UDP est principalement utilisé dans les réseaux locaux (dans lesquels la probabilité de perte d'un paquet IP est moindre) ce qui justifie la non-nécessité de connexion virtuelle (par opposition à TCP). Les services l'utilisant sont généralement NFS (Network File System), et NIS (Network Information Service), respectivement pour le partage de fichiers et la centralisation d'informations relatives aux réseaux. Ils sont en effet conçus pour fonctionner en mode déconnectés. De fait, ces services doivent détecter et éventuellement corriger eux-mêmes les erreurs de transmission. Le protocole UDP sert aussi à transporter des données autorisant des pertes mais ayant des contraintes fortes sur les délais. En effet, un protocole en mode connecté procède à des retransmissions en cas de perte, ce qui peut considérablement augmenter le délai de bout en bout.

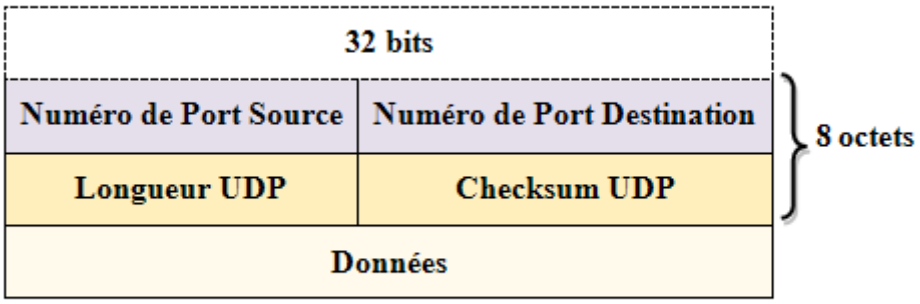


Figure II.9 : Structure d'une trame UDP [52].

II.7 Conclusion

Dans ce chapitre, nous avons décrit l'architecture d'un système de reconnaissance sur IP. Nous avons abordés aussi les principes de modèle client serveur, puis l'implémentation des sockets. Ensuite nous avons également cité le protocole IP et les deux protocoles TCP et UDP de la couche Transport utilisés par ce système pour la transmission des données. Le protocole de transport TCP est capable de répondre aux exigences de fiabilité et d'ordre requises par les applications de données, telles que la navigation web, le transfert de fichiers, le courrier électronique, etc. Le protocole UDP est un protocole de transport minimaliste. Sa simplicité est requise par deux classes d'applications :

- Celles qui effectuent de nombreux échanges requête-réponse de petite taille. Ces applications privilégient la simplicité de délivrance des paquets. On peut citer les applications DNS (Domain Name System), SNMP (Simple Network Management Protocol), DHCP (Dynamic Host Configuration Protocol), RIP (Routing Information Protocol). Leur très faible volume en taille de paquet aura peu d'impact sur le réseau.
- Celles qui ont des contraintes temporelles et d'interactivité. Il s'agit des applications de streaming média tel que IPTV (Internet Protocol Television), VoD (Video on Demand), Voix sur IP (VoIP). Des pertes occasionnelles de paquets sont encore tolérables.

Chapitre III

Evaluation expérimentale

III.1 Introduction

Ce chapitre traite de la mise en œuvre d'un système de vérification de locuteur sur IP, basé sur l'architecture client/serveur, en exploitant le codec G.729 à 8kbit/s. Le client transmet la parole produite par un locuteur et codée à l'aide du codeur G.729, en utilisant le protocole UDP, au serveur sur lequel le locuteur doit être reconnu.

Nous décrivons dans un premier temps la base de données utilisée, la base de données transcodée G.729, et dans un second temps, l'analyse acoustique appliquée. Nous présenterons également les performances de la vérification du locuteur, basée sur la méthode GMM-UBM. Ces performances sont évaluées en fonction de l'ordre des modèles, et du nombre des coefficients.

III.2 Outils de programmation utilisés

III.2.1 MATALB

Le langage Matlab (Contraction du terme anglais Matrix Laboratory), a été conçu par Cleve Moler à la fin des années 1970 à partir des bibliothèques Fortran. Matlab a ensuite évolué, en intégrant par exemple la bibliothèque LAPACK en 2000, en se dotant de nombreuses boites à outils (Toolbox) et en incluant les possibilités données par d'autres langages de programmation comme C++ ou Java. Les m-files de Matlab sont utilisés pour générer les programmes du codec G729 et pour générer et extraire les coefficients LPCC et MFCC à partir des LSP basés sur le G.729 bit-stream.

III.2.2 C++

Le C++ est actuellement le langage le plus utilisé au monde. C'est un langage de programmation permettant la programmation sous de multiples paradigmes comme la programmation procédurale, la programmation orientée objet et la programmation générique. Pour notre application, on a utilisé C++ pour générer le client et le serveur.

III.3 Description des bases de données

III.3.1 La base de données ARADIGIT

La base de données parole exploitée dans ce travail est la base de données ARADIGIT. Cette base a été conçue au laboratoire LCPTS de la faculté d'Electronique et d'Informatique de l'USTHB [54]. Elle est constituée de prononciations des 10 chiffres de la langue Arabe, de zéro jusqu'à neuf, prononcés par 110 locuteurs (hommes et femmes) avec

trois répétitions pour chaque chiffre. Cette base a été enregistrée par des locuteurs algériens de différentes régions âgés entre 18 et 50 ans dans un environnement calme, avec un niveau de bruit ambiant inférieur à 35 dB, sous le format WAV, avec une fréquence d'échantillonnage égale à 22,050 kHz puis sous échantillonnée à 16 KHz puis à 8kHz.

(صفر, واحد, اثنان, ثلاثة, أربعة, خمسة, ستة, سبعة, ثمانية, تسعة)

III.3.2 Bases de données extraites d'ARADIGIT

La base de données ARADIGIT contient des fichiers audio de très court durée, et la plateforme ALIZE demande des données d'apprentissage de longue durée. Dans ce but, les chiffres de zéro jusqu'à sept ont été concaténés pour construire des fichiers audio de 4 secondes de parole, utilisés comme des données d'apprentissage. Les deux chiffres ; huit et neuf sont concaténés et exploités pour les tests. Cette nouvelle base de données, nommée ARADIGIT8K, est constituée de 270 prononciations, 60 sont utilisées pour construire le modèle du monde, et les 210 sont exploités pour l'apprentissage et le test.

La base de données ARADIGIT8K doit être encodée par le G.729 au niveau du client, puis les éléments encodés sont envoyés via un réseau LAN (Local Area Network) basé sur le protocole UDP au serveur, où le bit-stream doit être décodé par le G.729, pour restituer la base de données G.729-ARADIGIT8K,

Dans ce travail nous exploitons deux bases de données :

1. ARADIGIT8K: La base de données originale et son codage;
2. G.729-ARADIGIT8K: La base de données transcodée au niveau du serveur par le G.729 via un réseau LAN en utilisant le protocole UDP;

III.4 Architecture client/serveur avec le codec G.729

III.4.1 Côté client

Pour transmettre la parole avec la base de données codée par le codeur G.729 au niveau du client, chaque fichier audio codé est alors mis sous forme de paquets IP et transmis, en utilisant le protocole UDP vers le serveur distant. Le serveur attendra les fichiers codés envoyés par le client en utilisant le protocole UDP.

La figure suivante montre la mise en œuvre du codeur G.729 coté client. Le fichier audio est codé par le G.729 et les trames issues du codeur seront mises dans des paquets UDP pour être prêts pour la transmission dans un réseau LAN.

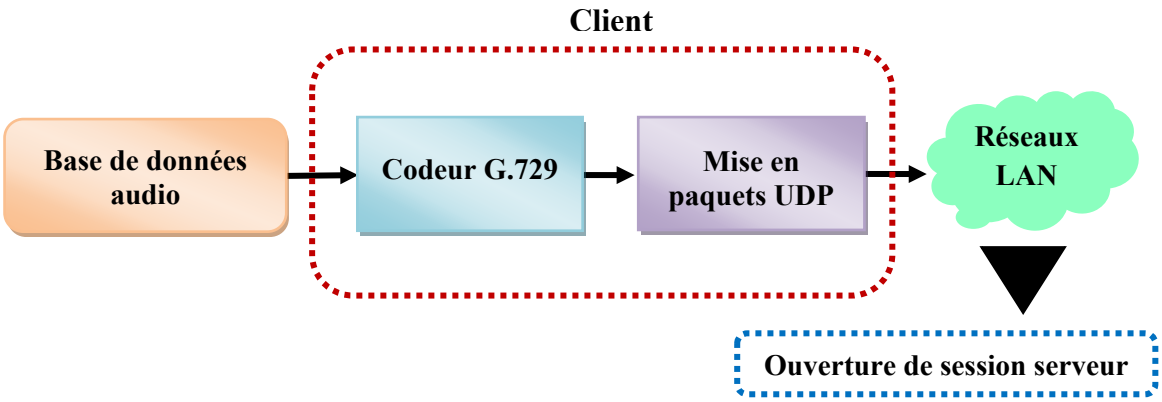


Figure III.1 : Implémentation coté client.

III.4.2 Les étapes de transmission

Après initialisation et création du socket des deux cotes client et serveur il y a plusieurs étapes qui suivent pour connecter entre eux le client et le serveur et transmettre des signaux de parole issus de la base de données codée. Le client demande l'adresse IP du serveur. Le client doit commencer l'envoi des signaux de parole en suivant les étapes ci-après :

1. Ouvrir le fichier en lecture
2. Envoyer la longueur du nom du fichier
3. Envoyer le nom du fichier
4. Envoyer la taille du fichier
5. Envoyer le fichier en paquet (la taille de chaque paquet reste au choix).

En recevant ces fichiers au niveau du serveur, ce dernier, ouvre le port correspondant, accepte la demande de connexion au client, reçoit la taille du nom du fichier puis le nom du fichier et enfin reçoit tous le fichier envoyé. La figure ci-dessous résume les étapes de transmission des fichiers.

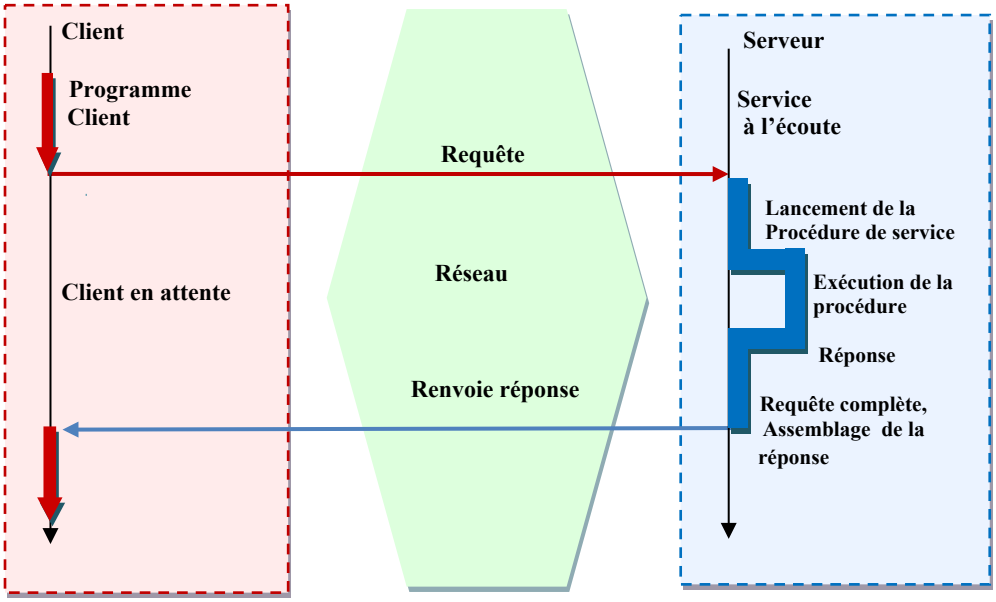


Figure III.2 : Communication entre client serveur.

III.4.3 Côté serveur

Le côté serveur est l'unité où les paquets reçus du client sont pris dans le reste du processus de transmission. Dans un premier temps, le serveur reçoit une requête UDP depuis le client et envoie une confirmation de réponse d'acceptation d'établissement d'une connexion. Une fois la connexion établie, le serveur sera en attente pour recevoir les paquets UDP. Chaque paquet reçu est stocké pour être ensuite écrit dans un fichier avec un format spécifique requis par le décodeur G.729, comme illustré dans figure ci-dessous.

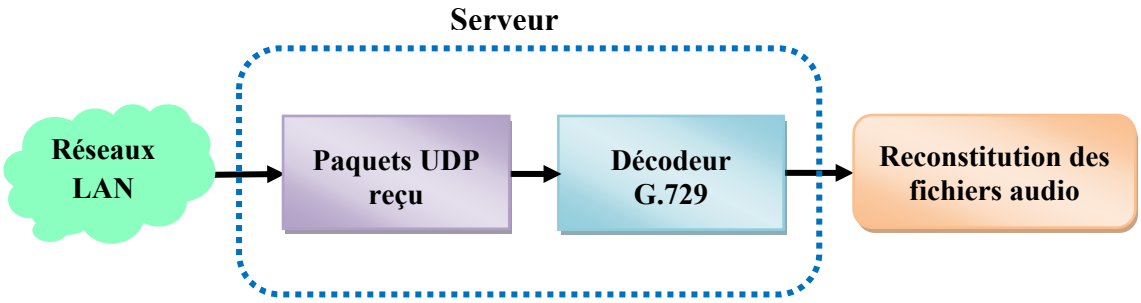


Figure III.3 : Implémentation coté serveur.

III.5 Extraction des caractéristiques

L'extraction/sélection des caractéristiques permet de réduire la dimensionnalité des données. Les méthodes de réduction de dimension sont nombreuses et ont pour objectif de conserver le maximum d'information possible dans un espace de dimension inférieure. Dans ce travail les paramètres sont extraites à partir de deux bases de données ; ARADIGIT8K et G.729-ARADIGIT8K (ARADIGIT8K transcodée G.729). Notre travail est focalisé sur l'extraction des MFCC, ces coefficients, une fois calculés, seront utilisés dans l'étape de modalisation par le GMM-UBM.

III.6 Evaluation des performances

Avant de procéder à la phase d'évaluation des performances sur IP nous examinons d'abord le chemin optimal qui donne les meilleures performances de la vérification du locuteur sans codec.

III.6.1 Influence de l'ordre de modèle sur ARADIGIT8K

Dans cette expérience, nous étudions les performances du système GMM-UBM, en utilisant 40 paramètres MFCC (20 paramètres+ 20 delta-paramètres) extraites à partir de la base ARADIGIT8K. Les figures ci-dessous illustrent les courbe DET et représentent l'influence de l'ordre des modèles (ou nombre de gaussiennes) sur les performances de la vérification du locuteur dans le cas où on a très peu de données d'apprentissage (4 à 5 secondes de parole).

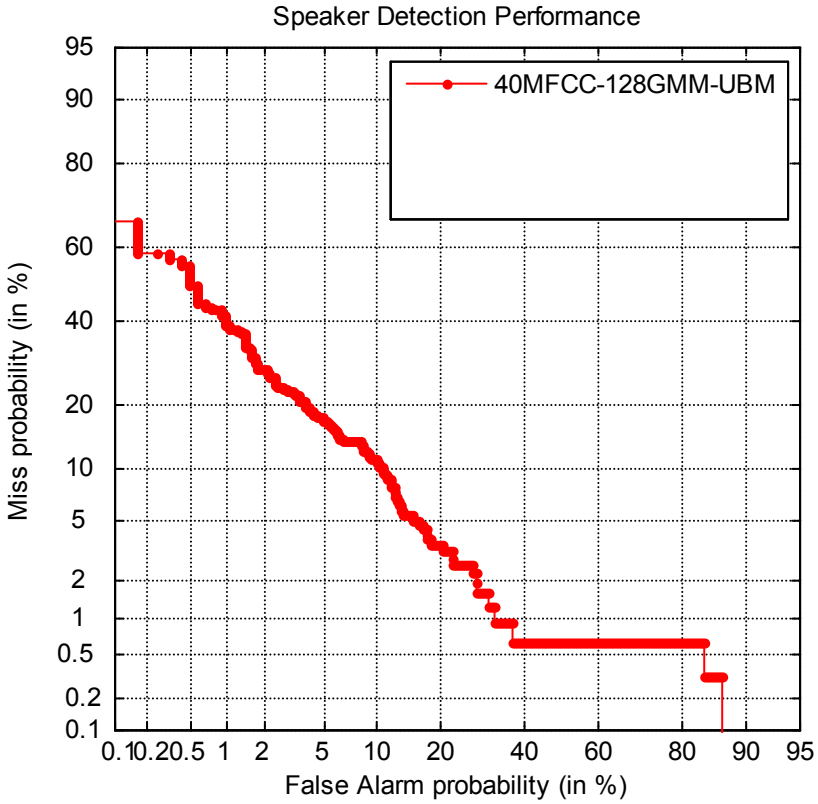


Figure III.4 : Performance d'un système VAL pour 128 gaussiennes de modèle GMM-UBM.

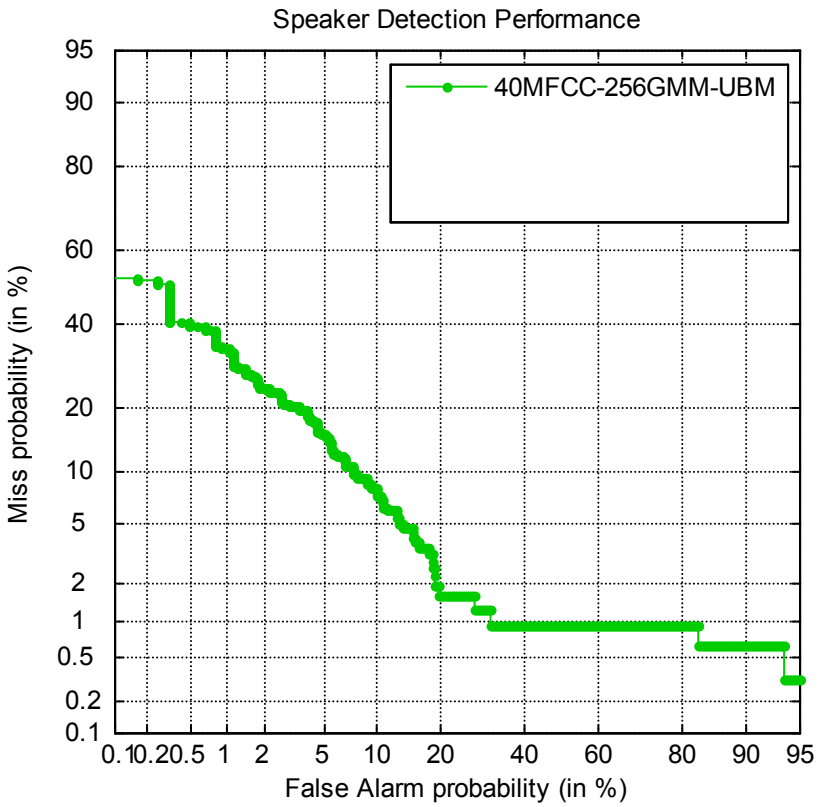


Figure III.5 : Performance d'un système VAL pour 256 gaussiennes de modèle GMM-UBM.

Les figures III.4 et III.5 montrent que l'augmentation du nombre de gaussiennes de 128 à 256 n'apporte pas une amélioration des performances, pour le système de vérification GMM-UBM basé sur les coefficients MFCC, au contraire les performances se dégradent avec l'augmentation du nombre de gaussiennes.

Le choix de l'ordre du modèle GMM-UBM dépend de sa finesse et de la quantité de données d'apprentissage. Choisir un ordre trop peu élevé va nuire à la précision du modèle. Choisir trop de composantes engendrera une charge de calcul plus importante. En général, 128 composantes suffisent pour représenter un locuteur disposant de très peu de données d'apprentissage (5 secondes de parole).

D'après ces résultats, la meilleure configuration que nous devrons utiliser dans la suite des expériences est l'utilisation un nombre de GMM égale à 128.

V.6.2 Influence de nombre des paramètres sur ARADIGIT8K

Cette partie consiste à étudier l'influence de nombre des coefficients MFCC sur les performances de vérification en exploitent la base ARADIGIT8K. Pour cela on fixe l'ordre du model GMM-UBM à 128.

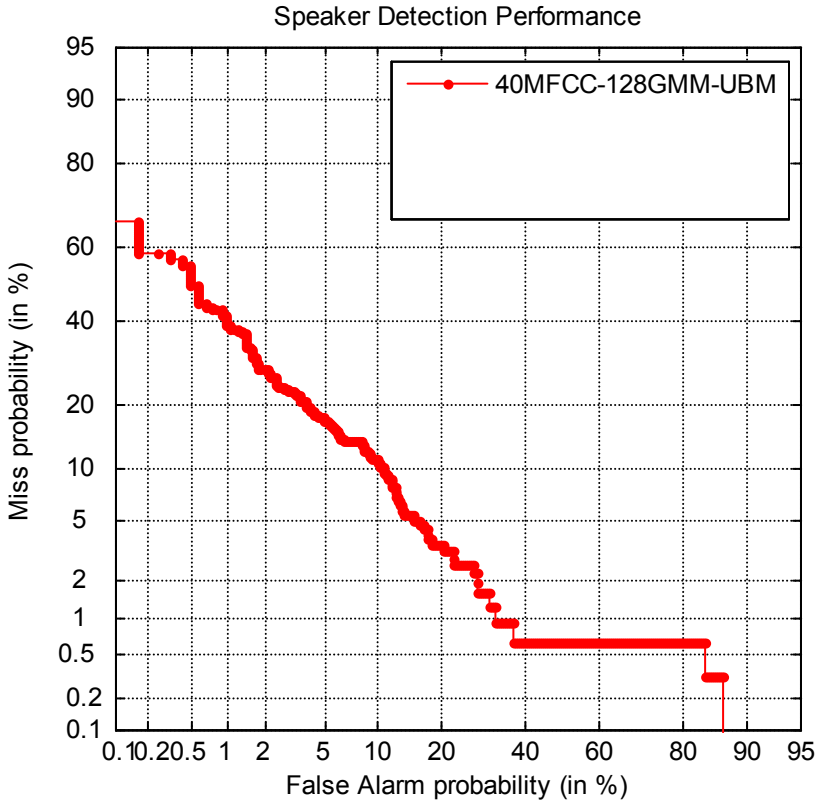


Figure III.6 : Performance d'un système VAL à base de GMM-UBM en utilisant 40 MFCC.

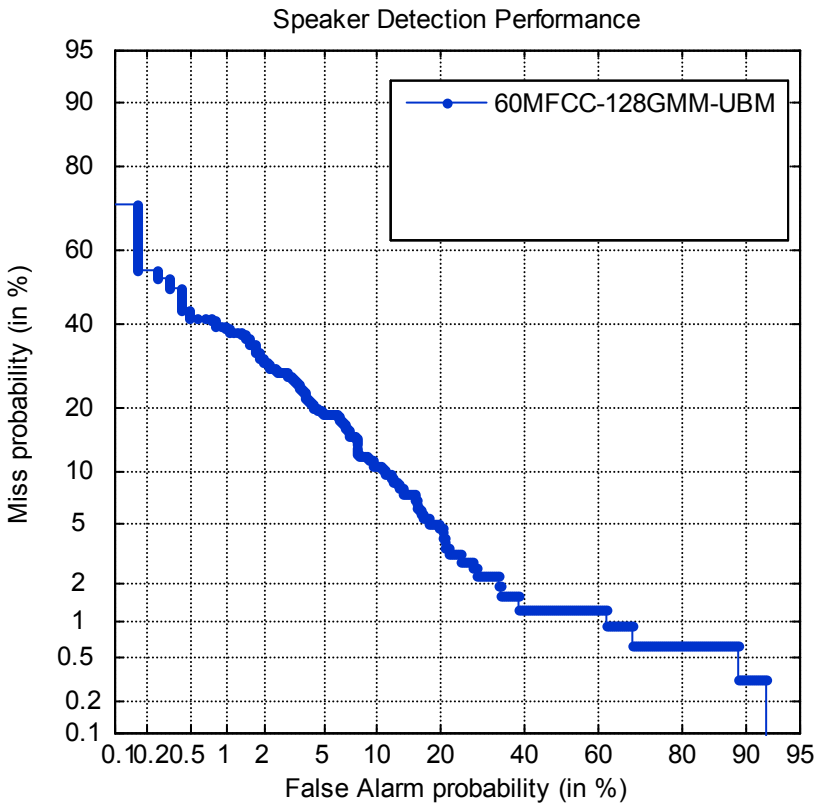


Figure III.7 : Performance d'un système VAL à base de GMM-UBM en utilisant 60 MFCC.

Les figures III.6 et III.7 présentent l'évaluation des performances de vérification du modèle GMM-UBM pour 40 et 60 coefficients MFCC extraites directement de la base ARADIGIT8K. On remarque qu'avec 40 coefficients MFCC on obtient le taux de reconnaissance correct de 91%. On remarque bien que lorsque le nombre des paramètres MFCC augmente les performances tendance à rester stable ou diminuer. Il est donc inutile de prendre un nombre des paramètres supérieur à 40.

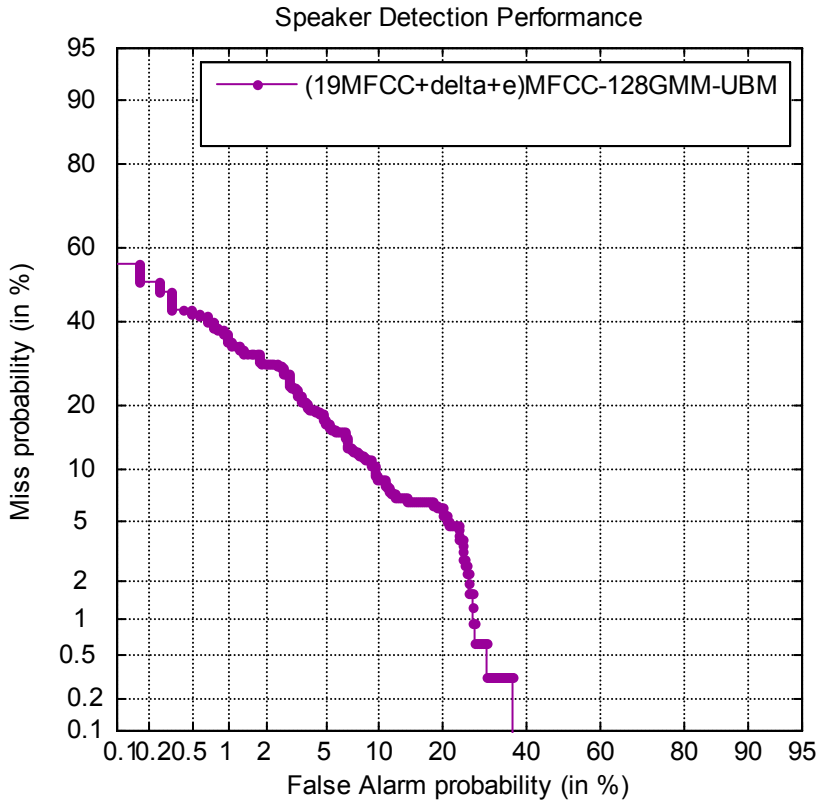


Figure III.8 : Performance d'un système RAL à base de 128 gaussiennes de modèle GMM-UBM en utilisant 19 paramètres MFCC avec leur delta et l'énergie.

La figure III.8 montre que les performances obtenus par les paramètres 40 (20MFCC+20delta) et 40(20LPCC+20delta) sont meilleurs que les performances obtenues par les paramètres 40(19MFCC+19delta+2energie). Ces résultats permettent de conclure que dans notre étude, le rajout de paramètres supplémentaires n'ajoute aucune amélioration significative sur les performances de vérification du locuteur par l'approche GMM-UBM.

D'après ces résultats, et pour la meilleure configuration que nous devrons utiliser dans la suite des expériences nous 'utilisons et 40 coefficients (20coefficients+ 20delta) basé sur 128 gaussiennes.

V.6.3 Influence du G.729 sur ARADIGIT8K

Le codage de la parole est très utilisé sur les réseaux de communication. Il permet de d'optimiser l'utilisation de la capacité du canal (ou la bande passante) nécessaire au transfert de la parole. Une considération importante dans tout codage de parole est la qualité du signal reconstruit. Les recherches sur les différents types de codage essaient toujours de trouver un bon compromis entre la qualité du signal de parole restitué et le débit de transmission.

Dans cette expérience, nous allons utiliser le codeur G.729 implémenté sur l'architecture client serveur. Ce codec comprend deux parties: le codeur et le décodeur. Le codeur analyse le signal et extrait un nombre réduit de paramètres pertinents qui sont représentés par un nombre restreint de bits pour archivage sur le client et transmission via le réseau IP, en utilisant toujours le protocole UDP où chaque paquet contient 80 bits de la sortie de codeur. Le serveur récupère les paramètres comme bit-stream envoyé par le client et le décodeur utilise ces paramètres pour reconstruire un signal de parole synthétique puis on applique le système de reconnaissance.

L'objectif de cette expérience est d'étudier l'influence du codec G.729 sur la qualité de la voix et la vérification du locuteur sur IP. Dans ce but, on applique le G.729 à la base de données ARADIGIT8K pour obtenir la base transcodée G.729 (G729-ARADIGIT8K). Les résultats obtenus sont illustrés dans la figure III.9 :

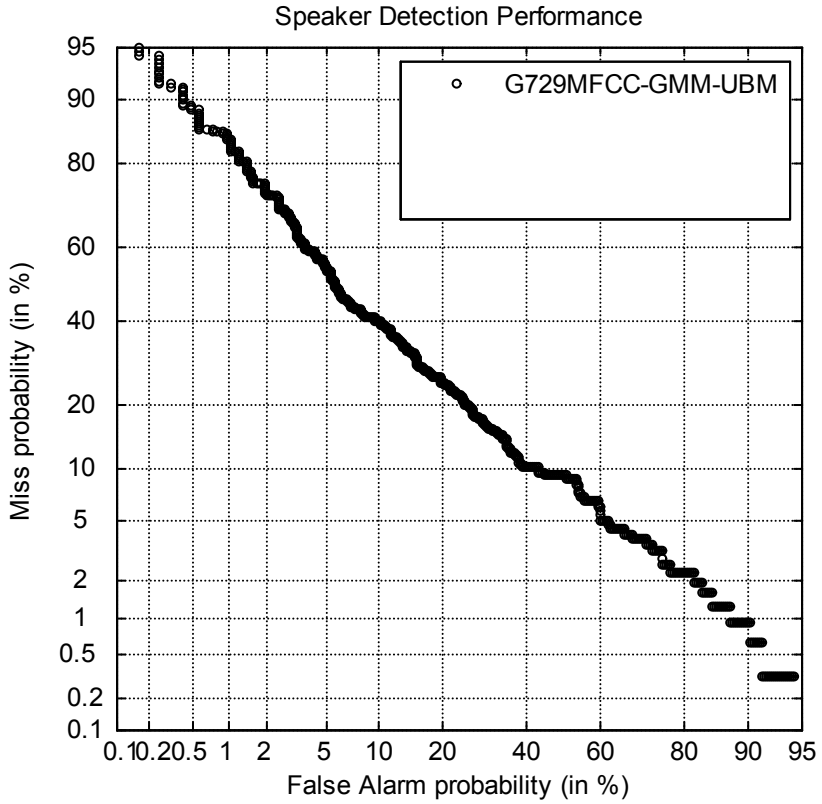


Figure III.9 : Les performances d'un système VAL sur IP en utilisant le codec G.729 avec modélisation GMM-UBM.

Les résultats obtenus montrent que les performances de la reconnaissance diminuent à cause des distorsions apportées par le codec G.729. Cela est dû à la dégradation de la qualité du signal introduite par ce codec. Les performances se dégradent et tombent à 78% avec l'utilisation des MFCC basé GMM-UBM,

La figure III.10 illustre la comparaison entres les performances de la vérification du locuteur basée sur la base de données G.729-ARADIGIT8K sur IP et celles obtenues en utilisant la base ARADIGIT8K sans codage.

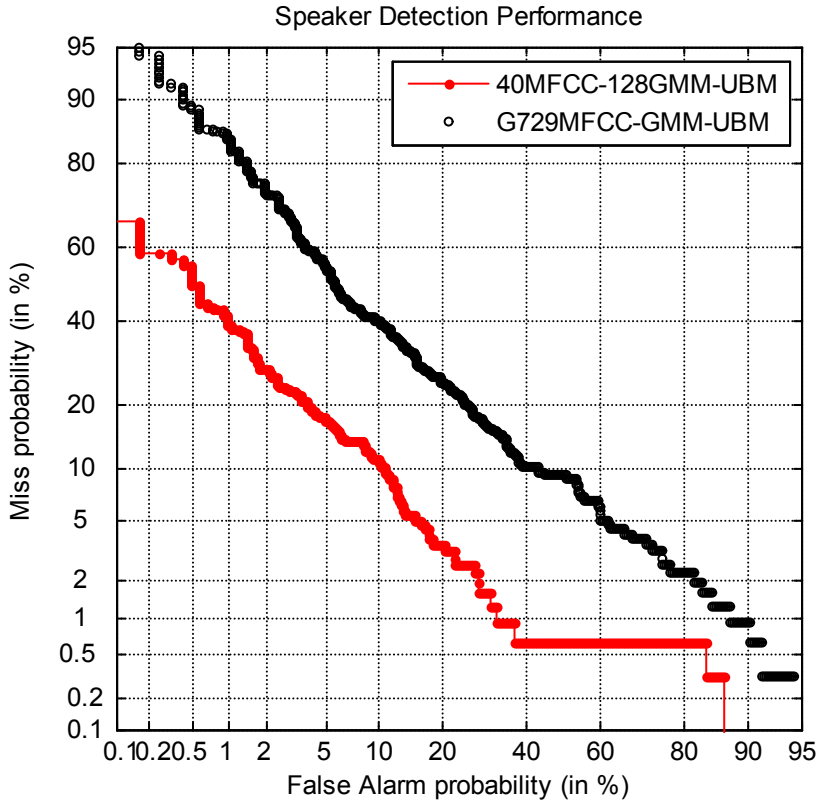


Figure III.10 : Les performances d'un système VAL normal et sur IP basé sur le codec G.729 en utilisant le modèle GMM-UBM.

Les courbes DET montrent une dégradation significative des taux de reconnaissance. Elle s'élève à plus de 10 % pour les deux systèmes GMM-UBM en utilisant les MFCC extraite à partir de la base G.729-ARADIGIT8K par rapport aux résultats basés sur ARADIGIT8K.

Dans cette expérience, nous avons comparé nos résultats de la vérification du locuteur avec une vérification sur IP utilisant la reconstruction du signal (parole synthétisée) au niveau du serveur. Pour cela, nous avons utilisé le codec G.729, dédié en particulier à la VoIP. La partie codeur agit au niveau du client et la partie décodeur au niveau du serveur devant effectuer la vérification après reconstruction du signal. Les résultats obtenus montrent clairement que les taux de vérification obtenus par les systèmes de reconnaissance appliqués après la reconstruction du signal décroissent significativement par rapport à ceux obtenus avec une base de données sans codage, car le calcul des paramètres cepstraux à partir du signal resynthétisé introduit des distorsions dues à la resynthèse (décodage) du signal.

Le taux de la reconnaissance est associé à la possibilité de séparer les distributions client et imposteur de chaque modèle. Cette caractéristique est directement liée à la performance de reconnaissance. La difficulté de décision en vérification se rapporte à chacun des éléments de la chaîne de traitement de système VAL sur IP. Elle peut être due à la parole en elle-même qui est plus ou moins fiable (5 secondes), aux conditions d'acquisition des données, au G.729 codec de parole choisi et à la fiabilité du système de reconnaissance à travers IP.

V.7 Conclusion

Ce chapitre a porté sur le développement et la mise en œuvre complète d'un système de VAL sur IP basé sur l'architecture client/serveur. Une mise en œuvre du côté du client avec son unité d'encodeur G.729 a été réalisée, ainsi qu'une mise en œuvre adaptée du côté serveur d'un décodeur et d'un système de reconnaissance automatique de locuteur basé sur le model GMM-UBM. Différents types de communication réseau ont été employés où le bit-stream issu du codeur G.729 est envoyé comme des paquets UDP vers le serveur, pour la reconstruction du signal puis appliqué au système de reconnaissance. Les performances de la reconnaissance distribuée basées sur la base G.729-ARADIGIT8K transcodée sont faibles par rapport à la base ARADIGIT8K, à cause des contraintes liées au codeur.

CONCLUSION GENERALE ET PERSPECTIVES

Ce travail s'inscrit dans le cadre d'un axe émergeant à savoir la vérification du locuteur, utilisant les réseaux de communications sur IP. Dans ce travail, nous avons exploité des techniques de traitements vocaux dédiés à vérification de locuteurs dans un système distribuée (DSR) sur IP basé sur l'architecture client/serveur, en utilisant le codec de la parole G.729.

La reconnaissance automatique du locuteur distribué (DSR), dans sa version vérification, consiste à confirmer ou infirmer l'identité proclamée d'un individu par sa voix a travers un réseau IP. Notre travail a porté sur le développement et la mise en œuvre complète d'un système de reconnaissance automatique du locuteur distribué (DSR) basé sur l'architecture client- serveur en utilisant C++, ainsi que le model GMM-UBM. Une mise en œuvre du côté client avec son encodeur G.729 a été réalisée, ainsi qu'une mise en œuvre du G.729 décodeur au côté serveur. Différents types de communication réseau ont été élaborées et testées où le G.729 bit-stream, issu de l'encodeur au niveau du client, est envoyé comme des paquets UDP vers le serveur, le G.729 décodeur va reconstruire le signal de parole (resynthesized speech) puis faire la vérification de locuteur au niveau du serveur.

Dans cette thèse, nous avons adressé quelques axes importants du domaine de la vérification du locuteur avec les réseaux IP, à savoir : Exploitation du codec G.729, investigation du modèle GMM-UBM et l'architecture client serveur. Néanmoins, il reste d'autres axes qui peuvent faire l'objet de futurs travaux, parmi lesquels :

- Le premier objectif de ces travaux futurs sera focalisé sur l'utilisation des réseaux WAN (Wide Area Network) et l'inclusion de la QoS (Qualité de Service) ;
- Investiguer d'autres méthodes de reconnaissance de locuteur, en particulier celle basée sur les I-vectors
- Utilisation d'autre type de codecs de parole.

BIBLIOGRAPHIE

Bibliographie

[1] A. Leman, 2011. *Diagnostic et évaluation automatique de la qualité vocale à partir d'indicateurs hybrides (Modèle DESQHI)*. Thèse de doctorat, Institut National des Sciences Appliquées de Lyon.

[2] S. Ouni, 2001. *Modélisation de l'espace articulatoire par un codebook hypercubique pour l'inversion acoustico-articulatoire*. Thèse doctorat, Université de Henri Poincaré-Nancy 1.

[3] A. Larcher, 2009. *Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée*. Thèse doctorat, Université d'Avignon et des Pays de Vaucluse en collaboration avec l'Université de Swansea.

[4] G. Fuchs, 2007. *Codage audio hiérarchique à faibles débits*. Thèse doctorat, Université de Sherbrooke (Québec), Canada.

[5] D. Meuwly, 2001. *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique.* Thèse de doctorat, Université de Lausanne.

[6] N. Gleiss, 1989. *Desirable Sending Frequency Response of Telephone Sets*. TELE (edition anglaise), Swedish Telecommunications Administration, pp. 18–23, SU-Stockholm.

[7] S. Greenberg, 2004. *Temporal properties of spoken language*, International Congres on Acoustics, Kyoto (Japan).

[8] I. Karlsson, T. Banziger, J. Dankovicov_a , T. Johnstone, J. Lindberg, H. Melin, F. Nolan, K. Scherer, 1998. *Speaker verification with elicited speaking-styles in the Verivox project*. Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pp. 207–210, Avignon (France).

[9] T. Banziger, G. Klasmeyer, T. Johnstone, T. Kamceva, K. R. Scherer, 2000. *Améliorer les systémes de vérification automatique du locuteur en intégrant la variabilité émotionnelle: Méthodes et premières données*. XXIIIème Journées d'Etudes sur la Parole (JEP), pp. 341–344, Aussois (France).

[10] A. Setlur et T. Jacobs, 1994. *Results of a speaker verification service trials using HMM models.* Workshop on Automatic Speaker Recognition, Identification, Verification, pp. 639–642, Martigny (Suisse).

[11] A. E. Rosenberg, 1976. *Automatic speaker verification.* IEEE Proceedings, vol. 64, n. 4, pp. 475–487.

[12] L. Heck, Y. Konig, K. Sonmez et M. Weintraub, 2000. *Robustness to telephone handset distorsion in speaker recognition by discriminative feature design*. Speech Communication, pp.181–192.

[13] J. Pelecanos et S. Sridharan, 2001. *Feature warping for robust speaker*

verification. A Speaker Odyssey, The Speaker Recognition Workshop, pp. 213–218, Crête (Grèce).

[14] R. Dunn, T. Quatieri, D. Reynolds et J. Campbell, 2001. Speaker recognition from coded speech in matched and mismatched conditions. A Speaker Odyssey, The Speaker Recognition Workshop, pp. 115–120, Crête (Grèce).

[15] L. Besacier, A. M. Ariyaeeinia, J. S. Mason, J. F. Bonastre, P. Mayorga, C. Fredouille, S. Meignier, J. Siau, N. Evans, R. Auckenthaler et R. Stapert, 2004. *Voice biometrics over the Internet in the framework of COST action 275*, EURASIP Journal of Applied Signal Processing, Special issue on biometric signal processing, pp. 466–479.

[16] G.R. Doddington, 1985. *Speaker recognition. Identifying people by their voices*. IEEE transactions, vol. 73, n. 11, pp. 1651–1664.

[17] J. Naik, 1994. *Speaker verification over the telephone: databases, algorithms and performance assessment*. Workshop on Automatic Speaker Recognition, Identification, Verification, pp. 31–38, Martigny (Suisse).

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

[19] C. Fredouille, 2000. *Approche Statistique pour la reconnaissance automatique du locuteur: Informations dynamiques et normalisation bayesienne des vraisemblances*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse (France).

[20] A.E. Rosenberg, 1976. *Automatic speaker verification.* IEEE Proceedings, volume 64(4), pp. 475–487.

[21] A.E. Rosenberg, I. Magrin-Chagnolleau, S. Parthasarathy et Q.Huang, 1998. *Speaker detection in broadcast speech databases.* International Conference on Spoken Language Processing (ICSLP), pp.1339–1342, Sydney (Australie).

[22] M. A. Przybocki et A. F. Martin, 1999. *Two-channel telephone data for speaker detection and speaker tracking*. European Conference on Speech Communication and Technology (Eurospeech), pp 2215–2218, Budapest (Hongrie).

[23] H. Ezzaidi, 2002. *Discrimination parole/musique et étude de nouveaux paramètres et modèles pour un système d'identification du locuteur dans le contexte de conférences téléphoniques*. Thèse de doctorat, Université de Québec.

[24] R. Boite, H. Bourlard, et T. Dutoit, 2000. *Traitement de la parole*. PPUR presses polytechniques.

[25] D. Matrouf, 1997. *Adaptation des modèles Acoustiques pour la Reconnaissance de la Parole*. Thèse de doctorat, Université de Paris-Sud.

[26] L. Rabiner et B.H. Juang, 1993. *Fundamentals of speech recognition*. signal processing. Prentice Hall, Englewood Cliffs, NJ.

[27] L. R. Rabiner, A. E. Rosenberg, et S. E. Levinson, 1978. *Considerations in dynamic time warping algorithms for discrete word recognition*. IEEE Transactions on Acoustics, Speech and Signal Processing 26(6), pp. 575–582.

[28] C. Myers, L. R. Rabiner, et A. E. Rosenberg, 1980. *Performance tradeoffs in dynamic time warping algorithms for isolated word recognition*. IEEE Transactions on Acoustics, Speech and Signal Processing ASSP-28(6), pp. 623–635.

[29] J. S. Mason, J. Oglesby et L. Xu, 1989. *Codebooks to optimise speaker recognition*. European Conference on Speech Communication and Technology (Eurospeech), pp. 267–270, Paris (France).

[30] F. K. Soong, A. E. Rosenberg, L. R. Rabiner et B. H. Juang, 1992. *A vector quantization approach to speaker recognition*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 387–390, Tampa (Etats Unis),

[31] A. E. Rosenberg et F. K. Soong, 1992. *Advances in Speech Signal Processing, Chapter Recent Research in Automatic Speaker Recognition*. pp. 701-738. Marcel Dekker.

[32] D. Reynolds, 1992. *A Gaussian mixture modeling approach to text independent speaker identification*. Thèse de doctorat, Georgia Institute of Technology, (Etats Unis).

[33] J.-L. Gauvain et C.-H. Lee, 1994. *Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains*. IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2, pp. 291–298.

[34] D. A. Reynolds, T. F. Quatieri, et R. B. Dunn, 2000. *Speaker verification using adapted gaussian mixture odels. Digital Signal Processing 10*. pp.19–41.

[35] J. Oglesby et J. S. Mason, 1990. *Optimisation of neural models for speaker identification*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 261–264.

[36] Y., Grenier, 1980. *Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur*. dans IXe`mes Journées d'Etudes sur la Parole (JEP), pp. 163–171, Strasbourg (France).

[37] Y. Gu et T. Thomas, 2001. *A text-independent speaker verification system using support vector machines classifier*. European Conference on Speech Communication and Technology (Eurospeech), pp. 1765–1769, Aalborg (Danemark).

[38] X. Dong, W. Zhaohui et Y. Yingchun, 2002. *Exploiting support vector machines in hidden Markov models for speaker verification*. International Conference on

Spoken Language Processing (ICSLP), pp. 1329–1332, Denver (Etats Unis).

[39] A. F. Martin et M. A. Przybocki, 1997. *The DET curve in assessment of detection task performance*. Proceedings of European Conference on Speech Communication and Technology (Eurospeech 97), pp. 1895–1898.

[40] J. M. Bardin, 2012. *RoSe : Un framework pour la conception et l'exécution d'applications distribuées dynamiques et hétérogènes*. Thèse Doctorat, Université de Grenoble.

[41] G. Coulouris, J. Dollimore, T. Kindberg et G. Blair, 2011. *Distributed systems : concepts and design*, 5thed. Addison-Wesley Publishing Company.

[42] ETSI, 2000. *Distributed speech recognition; front-end feature extraction algorithm: compression algorithms*. Speech processing, transmission and quality aspects (STQ).

[43] ETSI, 2002. *Distributed speech recognition; Advanced front-end feature extraction algorithm: Compression algorithms*. Speech processing, transmission and quality aspects (STQ).

[44] N. Srinivasamurthy et S. Narayanan, 2003. *Efficient Scalable Encoding for Distributed Speech Recognition, Integrated Media Systems Center*. Thèse doctorat, Université de the Southern California.

[45] S. Grassi, M. Ansorge, F. Pellandini, P.A. Farine, 2003. *Distributed speaker recognition using the ETSI aurora standard*.

[46] T. Turunen et D. Vlaj, 2001. A study of speech coding parameters in speech recognition. Proc. Eurospeech 2001, Aalborg, Denmark.

[47] S. Euler et J. Zinke, 1994. *The influence of speech coding algorithms on automatic speech recognition*. Proc. ICASSP, pp. 621–624.

[48] M. Rabah, 2000. *Gestion de la qualité de service et contrôle de topologie dans les réseaux ad hoc*. Thèse doctorat, Ecole nationale supérieure des télécommunications.

[49] L. Besaw, 1987. *Berkeley UNIX system calls and interprocess communication*, BSD Socket Reference.

[50] A.TRAD, 2006. *Déploiement à grande echelle de la voix sur IP dans des environnements hétérogènes*. Thèse doctorat, Université de Nice-Sophia Antipolis

[51] ISO/IEC, 1994. *Information technology – Open Systems Interconnection – Basic reference model: the basic model*. ISO/IEC7498-1.

[52] W. Stevens, 1994. *TCP/IP Illustrated*, Volume 1: The protocols. Chapter 17, 19, 20, 21 Addison Wesley.

[53] B. Abderrahim, 2013. *Intelligent mobile health monitoring systems*. Thèse doctorat. Université de Tlemcen.

[54] D. Yessad et A. Amrouche, 2013. *Robust regression fusion of GMM-UBM and GMM-SVM normalized scores using G729 bit-stream for speaker recognition over IP*. International Journal of Speech Technology (IJST). 31 July 2013 .ISSN : 1381-2416 Springer.

Résumé

Le développement de la VoIP, et par conséquent de la téléphonie sur IP (ToIP : Telephony Over IP), a ouvert de nouveaux horizons aux applications en reconnaissance vocale, d'où l'émergence de la Reconnaissance Vocale Distribuée (DSR : Distributed Speech and Speaker Recognition), touchant aussi bien la RAP que la RAL. En effet, l'identification et la vérification de locuteur deviennent une nécessité dans plusieurs domaines, notamment avec l'introduction des moyens de communications modernes. Domaine émergent à fort potentiel, la reconnaissance automatique du locuteur sur IP constitue donc un véritable chalenge pour le développement des technologies de communications du futur.

Dans ce travail, le développement du système DSR est basé sur l'architecture client/ serveur en exploitant le codec G.729. Le client transmet la voix codée issu du codeur G.729, en utilisant le protocole UDP, au serveur où s'effectue la reconnaissance basée sur le model GMM-UBM.

Mots clés : Le développement de la VoIP, Reconnaissance Vocale Distribuée, VOIP, le codec G.729

Abstract

The recent development of the VoIP (Voice over IP) technology, and consequently the ToIP (Telephony over IP), opened new challenge for speech the recognition applications, especially in distributed speaker recognition (DSR: Distributed Speech and Speaker Recognition). In fact, speaker identification and verification are becoming a necessity in several areas, especially with the introduction of modern means of communications. Automatic speaker recognition over IP became a great challenge for the growing development of the next communications technologies.

In this work, the DSR system performed is based on the client server architecture using G.729 codec. The client transmits the encoded voice issued from the codec G.729 using the UDP protocol, to the server where carried out the recognition based GMM-UBM model.

Keywords: The recent development of the VoIP ,distributed speaker recognition(DSR) , VoIP, the codec G.729