N°: 04 /2014-D/ELC

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieure de la Recherche Scientifique

Université des Sciences et de la Technologie Houari BOUMEDIENE
Faculté d'Electronique et d'Informatique

THÈSE

Présentée pour l'obtention du diplôme de DOCTORAT
En : ELECTRONIQUE
Spécialité : **Communication Parlée**

Par **: YESSAD Dalila**

THÈME

Reconnaissance automatique du locuteur dans les réseaux de communications VoIP

Soutenu publiquement le 24/06/2014 devant le jury composé de :

Mme	**M. GUERTI**	**Professeur, à ENP Alger**	**Présidente**
Mr	**A. AMROUCHE**	**Professeur, à l'USTHB**	**Directeur de Thèse**
Mme	**F. MERAZKA**	**Professeur, à l'USTHB**	**Examinatrice**
Mme	**L. HAMAMI**	**Professeur, à ENP Alger**	**Examinatrice**
Mme	**L. FALEK**	**Maître de conférence/A, à l'USTHB**	**Examinatrice**
Mr	**A. ANOU**	**Maître de conférence/A, à Blida**	**Examinateur**

Table des matières

Chapitre II: Les réseaux NGN et VoIP

Chapitre III: La reconnaissance automatique du locuteur distribuée

Chapitre IV: Approche de reconnaissance proposée basée sur la fusion des modèles et des scores à base de GMM-UBM et GMM-SVM

Table des figures

Chapitre V

Liste des tableaux

Chapitre IV

Abréviations

VoIP	Voice Over IP
ToIP	Telephony Over IP
DSR	Distributed Speech and Speaker Recognition
ITU	Union Internationale des Télécommunications
ETSI	European Telecommunications Standards Institute
RAL	Reconnaissance Automatique du Locuteur
IAL	Identification Automatique du Locuteur
VAL	Vérification Automatique du Locuteur
LPC	Linear Predictive Coding
MFCC	Mel-Frequency Cepstral Coefficient
LPCC	Linear Prediction Cepstral Coefficient
PLP	Perceptual Linear Predictive
RASTA	RelAtive SpecTrAl
DTW	Dynamic Time Warping
MSSO	Méthodes Statistiques du Second Ordre
HMM	Hidden Markov Model
GMM	Gaussian Mixture Model.
SVM	Support Vector Machine.
FA	Fausse Acceptation.
FR	Faux Rejet
DET	Detection Error Tradeoff
EER	Equal Error Rate
ROC	Receiving Operating Characteristic
NGN	réseaux de nouvelle génération
UNI	User-Network Interface
ANI	Application-Network Interface
NNI	Network-to-Network Interface
MG	Media Gateway
MGC	Media Gateway Controller
SG	Signalling Gateway

TDM	Time Division Multiplexing
MGCP	Media Gateway Control Protocol
SIP	Session Initiation Protocol
IETF	Internet Engineering Task Force
GLP	Gateway Location Protocol
MCM	multipoint control unit
RAS	Registration, Admission and Status
UA	User Agent
UAC	User Agent Client
GSM-EFR	Global System for Mobile Communications-Enhanced Full Rate
CS-ACELP	Conjugate-Structure Algebraic-Code-Excited Linear-Prediction
AMR	Adaptive Multi-Rate
VSELP	Vector-Sum Excited Linear Prediction
HR	Half Rate
RPE-LTP	Regular Pulse Excitation – Long Term Prediction
MP-MLQ	Multi-Pulse Maximum Likelihood Quantization
MOS	Mean Opinion Score
BI-LPC	Block-Independent Linear Predictive Coding
iLBC	Internet Low Bitrate Codec
CNG	Comfort Noise Generator
PLC	Packet Loss Concealment
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
UBM	Universal Background Model
EM	Expectation Maximisation
MAP	Maximum a Posteriori
NSRRF	Normalised Score based Robust Regression Fusion

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.

Ce travail de thèse est une contribution au développement des systèmes de reconnaissance automatique du locuteur distribuée (DSR) proposant une nouvelle approche dans le domaine de la reconnaissance du locuteur distribuée et la VoIP.

Dans cette thèse, le développement du système DSR est basé sur l'architecture client/ serveur en exploitant la plateforme ALIZE. Le client transmet la voix codée issu du codeur G.729, en utilisant le protocole UDP, au serveur où s'effectue la reconnaissance. Pour faire l'économie d'une reconnaissance à base du signal de parole reconstituée (reconstructed or resynthesized speech) très couteuse en temps machine nous exploitons directement les LSP issus du bit-stream (LSP-G.729-bit-stream). Dans le but d'améliorer les performances du système distribué basé sur LSP-G.729-bit-stream, nous avons proposé une méthode de fusion originale basée sur la régression robuste permettant d'augmenter les performances et d'apporter un gain significatif, jusqu'à 99% de taux vérification correct. Cette méthode, appelée NSRRF (Normalized Score following Robust Regression Fusion), combinant la fusion des scores normalisée et des modèles GMM-UBM et GMM-SVM, donne des résultats prometteurs, comparativement avec ceux obtenus avec DSR utilisant le codec G.729 sans fusion ou les systèmes DSR conventionnels avec reconstruction du signal (signal resynthétisé). Les résultats satisfaisants obtenus, ont permis de confirmer l'efficacité et la pertinence de l'approche dédiée à la reconnaissance de locuteur en VoIP proposée dans cette thèse.

***Mots-clé*:** reconnaissance du locuteur sur IP, Modèle de mélange de gaussiennes-Universal Background (GMM-UBM), Supervecteur de mélange de gaussiennes (GMM-SVM), Plateforme ALIZE, G.729 bit-stream, Normalisation Min-Max, Fusion par la Régression Robuste.

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 thesis, new approach for developing distributed automatic speaker recognition system (DSR) over IP networks is proposed.

In this work, the DSR system performed is based on the client server architecture using ALIZE platform. The client transmits the encoded voice issued from the codec G.729 using the UDP protocol, to the server where carried out the recognition. We used directly the LSP coefficients of G.729 bit-stream (LSP-G.729-bit-stream) without reconstructing the transmitted speech. In order to improve the performance of the distributed system based on the LSP-G.729-bit-stream, we have proposed an original fusion method based on robust regression able to increase performance and providing a significant recognition rate up to 99%, This method, so called NSRRF (Normalized Score following Robust Regression Fusion), combining the normalized scores issued from GMM-UBM and GMM-SVM models, shows promising results, compared with those obtained by DSR system using the G.729 codec without fusion, or conventional DSR systems using the reconstructed signal (synthesized signal). The obtained results are significant and confirm the effectiveness and the accuracy of the approach proposed in this thesis dedicated to distributed speaker recognition over IP.

***Keywords*:** Speaker recognition over IP, Gaussian Mixture Model -Universal background (GMM-UBM), Gaussian supervector (GMM-SVM), ALIZE platform, G.729 bit-stream, Min-Max Normalization, Robust regression fusion.

Dédicaces

A ma très chère mère

REMERCIEMENTS

Remerciements

Ces remerciements s'adressent tout d'abord à mon directeur de thèse, le Professeur Abderrahmane AMROUCHE de l'équipe Reconnaissance Vocale et Applications du Laboratoire de Communication Parlée et Traitement du Signal, Faculté d'Electronique et d'Informatique de l'USTHB. J'ai bénéficié tout au long de ces quatre années de sa grande connaissance du domaine et de ses conseils avisés qui m'ont permis de concrétiser mon travail à travers les contributions proposées dans cette thèse. J'ai été très heureuse de travailler sous sa direction.

Mes remerciements s'adressent aussi aux membres de mon jury qui m'ont fait l'honneur de juger ce travail.

Je tiens particulièrement à remercier Mme Mhania GUERTI, Professeur à L'Ecole Nationale Polytechnique d'Alger qui a bien voulu me faire l'honneur de présider ce jury.

Je tiens a exprimer toute ma gratitude à Madame Fatiha MERAZKA, Professeur à la Faculté d'Electronique et d'Informatique de l'USTHB, Madame Latifa HAMAMI, Professeur à L'Ecole Nationale Polytechnique d'Alger, Madame Leila FALEK, Maitre de Conférences à la Faculté d'Electronique et d'Informatique de l'USTHB, et Monsieur Abderrahmane ANOU, Maitre de Conférences à l'Université Saad dahlab de Blida, qui ont accepté d'examiner ce travail. Merci pour votre lecture constructive et vos suggestions.

Je tiens aussi à remercier tout le personnel de la Faculté d'Electronique et d'Informatique de l'USTHB.

Et puis pour faire du bon travail il faut une bonne ambiance, alors merci à tout ceux qui font régner la joie et la bonne humeur et ils sont nombreux, merci à tous.

Je garde le meilleur pour la fin, je tiens à remercier et à dédier cette thèse à ma mère et à mon mari. Merci pour ce merveilleux soutien et pour avoir toujours fait en sorte que tout se passe dans les meilleures conditions possibles. Merci enfin à ma sœur pour n'avoir cessée de m'assurer son soutien tout au long de ma thèse, en me poussant à continuer dans les moments difficiles. Merci Sabrina.

INTRODUCTION GENERALE

INTRODUCTION

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. La plupart de ces codecs fonctionnent dans la gamme de 4,8 kbit/s à 16 kbit/s et ont une qualité vocale et un taux de compression raisonnable. En général, plus le débit binaire de la parole est grand, plus la qualité de la parole est bonne et plus l'application est gourmande en bande passante et en stockage. Dans la pratique, c'est toujours un compromis entre l'utilisation de la bande passante et la qualité vocale. 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 de thèse 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. L'objectif de nos travaux a donc consisté à diagnostiquer le plus

précisément possible les problèmes dûs au codage de la parole, durant la tache de la reconnaissance automatique du locuteur sur IP. A l'issu du diagnostic, nous avons constaté une dégradation plus importante due à la compression et décompression sur le signal parole. La contribution de cette thèse correspond donc à la proposition d'une technique de fusion des scores normalisés afin d'améliorer les performances de système de reconnaissance soumis à des conditions de pertes pendant la compression. Dans le cadre de ce travail nous nous intéressons au codec G.729 dédié à la VoIP. Il s'agit donc d'étudier et de mettre en œuvre le codec G.729, dans le but de construire le système de la reconnaissance du locuteur distribuée (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 divers langages de programmation, divers axes de recherches et de développement ont dus être définis :

Contribution au développement de la reconnaissance distribuée et la synthèse vocale :

- Conception et implémentation de l'architecture client/serveur en intégrant le codec G.729, avec implémentation de l'encodeur coté client et le décodeur coté serveur ;
- Elaboration de la base de données transcodée G.729 ;
- Utilisation du protocole UDP pour intégrer la VoIP;
- Développement et implémentation du système de reconnaissance du locuteur à base GMM-UBM ;
- Développement et implémentation du système de reconnaissance du locuteur hybride GMM-SVM;
- Normalisation Min-max des scores.

Contribution à l'amélioration des performances des systèmes de RAL :

- Fusion des scores normalisés issus des deux systèmes GMM-UBM et GMM-SVM basés sur la régression robuste proposée comme méthode de fusion des

scores dans le but d'améliorer les performances de système de reconnaissance distribué.

En rapport avec les langages de programmation sous les systèmes d'exploitation Windows et Linux, nous avons développé des algorithmes en :

- C++ pour générer les exécutables de chaque module de système RAL inspire à partir de la plateforme de reconnaissance locuteur ALIZE ;
- MATLAB pour créer les paramètres acoustiques basés sur le G729 bit-stream.

Ainsi, le travail de cette thèse a tout d'abord consiste à étudier le domaine de reconnaissance du locuteur en VoIP 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, la modélisation, la normalisation, la fusion proposée et d'éventuels traitements d'amélioration. L'étude de toutes ces composantes a permis d'établir les différents axes de recherche étudies pendant cette thèse.

Le 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.

Le chapitre 2 rappelle le contexte de l'évolution des architectures de télécommunications traditionnelles vers les réseaux de nouvelle génération, avant d'évoquer les principales caractéristiques de ces nouveaux réseaux. La deuxième partie de ce chapitre est consacrée a la présentation de notion VoIP et ses contraintes.

Dans ce contexte, la première orientation de nos travaux de recherche, décrite au chapitre 3, consacré à la description 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.

La seconde orientation de nos travaux, exposée au chapitre 4, est réservée à l'étude des modèles exploités et la méthode de fusion contribuée. Dans un premier temps, nous détaillons la modélisation générative du locuteur à base de modèles de mélange de gaussiennes-Universal Background (GMM-UBM), nous présentons ensuite le modèle hybride GMM-SVM

issu de la combinaison des méthodes génératives et discriminantes basées sur les machines à vecteurs supports (SVM). Dans un second temps, nous justifierons nos choix de technique de normalisation des scores et de notre proposition originelle de fusion des scores par la régression robuste.

A partir de l'architecture présentée au chapitre 3, ainsi que les modèles choisis au chapitre 4, nous reportons les résultats obtenus au chapitre 5. Ces résultats incluent l'étude des scores issus du paradigme GMM-UBM et le modèle hybride GMM-SVM ainsi que de leur normalisation et fusion utilisant la méthode de régression robuste, proposée comme méthode de fusion des scores normalisés.

Nous concluons cette thèse par une récapitulation des principales conclusions et discussions de ces travaux de recherches.

CHAPITRE I

Introduction à la a 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'extraction des paramètres acoustiques du signal de parole, la modélisation du locuteur et une dernière étape de décision (ex: décider de l'identité du locuteur).

I.2 Mécanismes de production de la parole et perception des sons

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 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 [1]. 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 [2].

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).

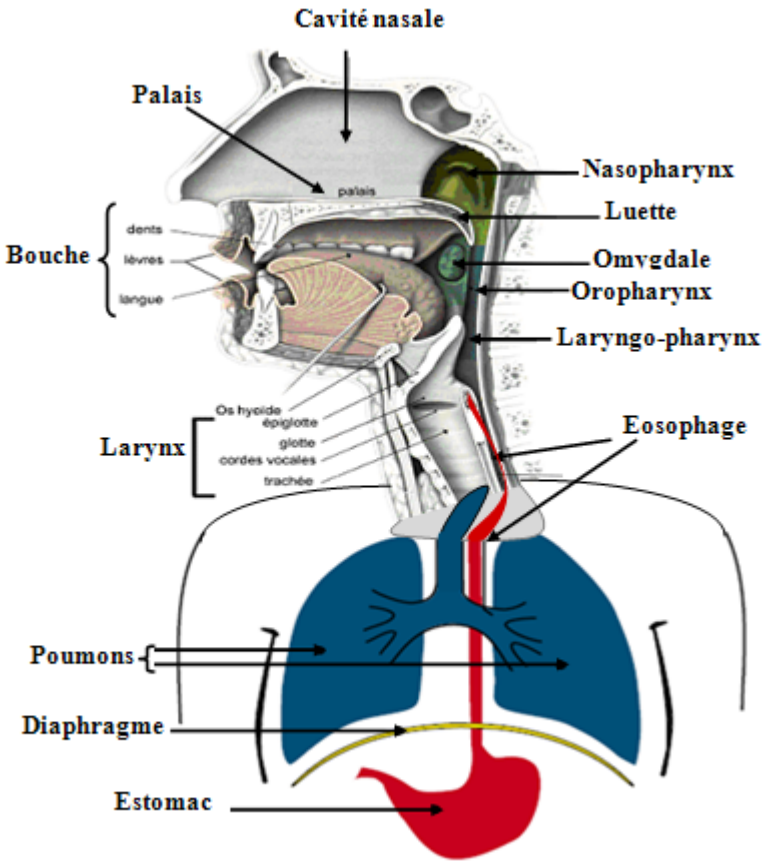


Figure I.1: Organes de production de la parole [3].

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 [2].

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 les particularités 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 [2], [3]. Le son résultant peut être

classé comme voisé ou non voisé, selon que l'air émis a fait vibrer les cordes vocales ou non [2], [3], [4].

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

L'oreille humaine ne peut percevoir que certains sons, le niveau d'intensité acoustique moyenne produit par un son de parole, mesuré à 1 mètre, est compris entre 30 et 110 dB, ce qui correspond respectivement à une voix chuchotée et à une voix criée [1]. L'étendue spectrale de la parole humaine est comprise en général entre 80 et 8000 Hz (certaines cantatrices peuvent atteindre 15 kHz dans le chant d'opéra), la puissance sonore de la parole dans une conversation normale se situe de 60 à 70 dB et le niveau d'intensité acoustique est maximal lorsque la fréquence se situe aux alentours de 500Hz [5]. La figure I.2 donne une représentation du domaine audible pour un être humain. On remarque tout d'abord que le niveau de perception dépend grandement de la plage de fréquences considérée ainsi que du niveau sonore. On définit alors deux courbes dans le plan fréquence-intensité : un seuil d'audibilité et un seuil de confort. La zone ainsi définie est le domaine dans lequel les sons peuvent être perçus. Tout signal en dehors de cette plage est inaudible, gênant ou même dangereux.

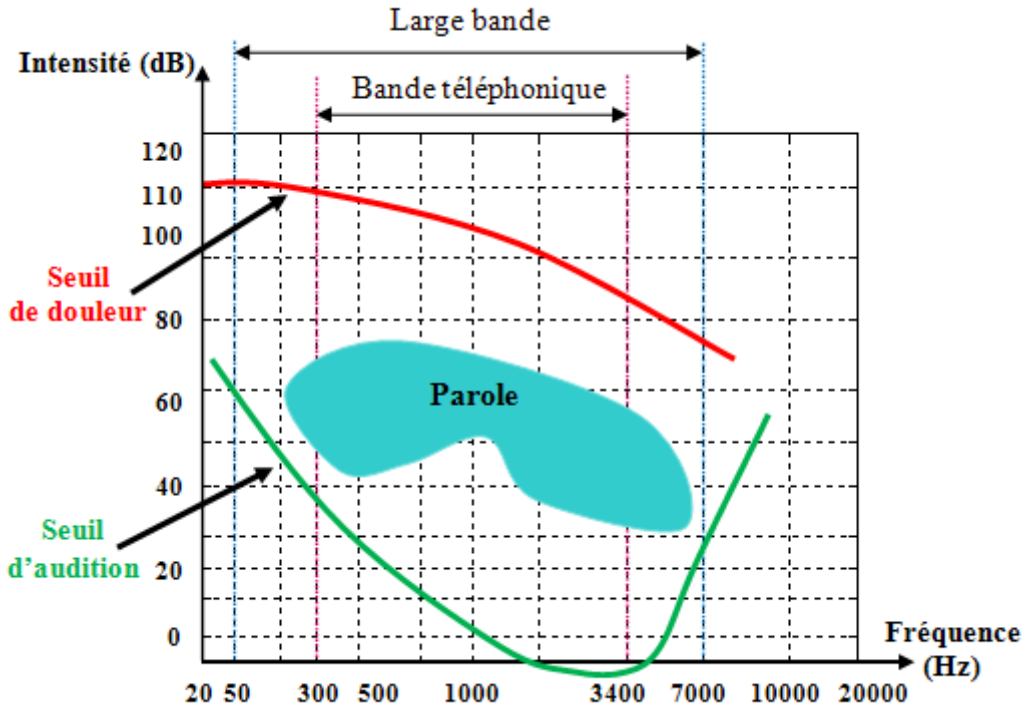


Figure I.2: Perception auditif.

La bande d'audition est composée des fréquences audibles par une oreille humaine situées entre 20 Hz et 20 kHz : ce qui à première vu semble très loin de la bande étroite de la téléphonie correspond à la plage de fréquence 300-3400 Hz, cette bande téléphonique est juste suffisante pour conserver l'intelligibilité du langage ainsi que les paramètres propres au locuteur (voix, émotion, etc.). En pratique la plage de 100-4000 Hz permet de positionner correctement les premiers formants (trois ou quatre) et de contenir toutes les premières harmoniques du pitch (voir § I.7.1.2), qui sont les informations utiles à la bonne compréhension de la parole humaine [1]. Cependant, l'intelligibilité de parole est intégralement contenue dans la bande téléphonique, sauf peut être pour quelques harmoniques où la fréquence de pitch est relativement faible (allant de 50 Hz pour les hommes à 600 Hz pour les voix les plus aiguës telle une voix d'enfant) [4]. Ainsi l'addition des fréquences inférieures à 300 Hz donne une meilleure représentation des premières harmoniques (phonèmes voisé). On remarque surtout cela pour un locuteur masculin pour lequel la fréquence de pitch est assez faible. La plage des hautes fréquences, supérieures à 3400 Hz n'a de l'importance que pour les phonèmes complexes.

La plage de la bande téléphonique, où le système auditif est le plus sensible, justifie le taux d'échantillonnage de 8 kHz pour les codecs de parole qui traitent en général le signal sur la bande étroite. Bien que la bande téléphonique soit suffisante pour la bonne compréhension du message, l'élargissement de la bande passante des codecs de 50 à 7000 Hz (large bande) permet de rendre la parole reconstituée plus naturelle avec une qualité se rapprochant d'une communication face à face [1], [6]. Les fréquences inférieures à 50 Hz et supérieures à 7000 Hz n'apportent pas plus d'information à la perception de la parole, surtout sur les phonèmes, l'unité principale constituant le langage parlé [7].

En conclusion, la parole humaine produit des fréquences qui, en partie, ne sont pas comprises dans la bande téléphonique, 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 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 dûe 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, etc.

I.3.1 Variabilité intra-locuteur

La variabilité intra-locuteur est une variabilité propre au locuteur qui ne peut pas reproduire exactement le même signal. Cette variabilité intra-locuteur est dépendante de état physique et psychologique pour un même individu, ainsi les facteurs de variabilités sont multiples, on cite:

- L'état pathologique : Des variations peuvent être induites involontairement sur la voix d'une personne, de type fatigue, rhume et stress, etc., ou les variations émotionnels. Ces facteurs provoquent des altérations momentanées dans la voix. Dans ce sens, la voix peut changer entre le début et la fin de la journée [8], [9]. Plus généralement, il est impossible pour une personne de répéter à l'identique le même signal de parole deux fois de suite. Une légère variation est toujours observée.
- Dans le cas d'une interaction volontaire et consciente avec un système de reconnaissance du locuteur, comme par exemple dans le cadre d'un accès sécurise, 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 ainsi sa voix évolue dans ce sens et s'en trouve modifiée.
- Enfin, à plus long terme, la voix change au fur et à mesure du vieillissement d'une personne.

L'influence de ces divers facteurs varie selon l'application visée. Des travaux ont montré l'importance des variations à long terme et que les performances d'un système se dégradent en augmentant le temps qui sépare les sessions des références et les tests [10], [11].

Plus ce temps augmente, plus les performances se dégradent. Néanmoins, même les variations à court terme (émotion, état pathologique) peuvent être très préjudiciables aux systèmes de RAL.

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

La transmission du signal de parole du locuteur au système de RAL chargé de l’analyser nécessite plusieurs étapes et emprunte divers types de supports. A chacune de ces étapes, le media utilisé (ex : microphone, combiné téléphonique) pour transporter ce signal y imprime sa marque. Ces empreintes 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

Si la bande téléphonique est reconnue pour dégrader les performances des systèmes de RAL, elle n’est pas la seule responsable. En effet, les dégradations de performances peuvent

être formulées par des canaux sur lesquels le signal de parole sera transmis jusqu' au système de RAL, notamment en termes de d'ajout de bruit. De très nombreux travaux ont été consacres à tenter de compenser la contribution du canal de transmission afin de s'affranchir de ces problèmes de variabilité [12], [13]. Néanmoins, ce problème voit peu à peu son importance s'amoindrir dans de nombreuses applications où la transmission de la parole, autrefois analogique, se fait de plus en plus sous forme numérique, surtout dans les cas de la téléphonie mobile ou de la transmission de la voix sur IP. La question de la contribution du canal se voit dans ce cas remplacée par celle des dégradations induites par le codage utilisé. De plus en plus d'études portent donc sur la RAL à partir des standards actuels de codage de la parole, s'intéressant plus particulièrement aux conséquences de la compression et des pertes de paquets lors de la transmission sur IP [14], [15].

I.3.5 Tentatives d'imposture et locuteurs non coopératifs

Selon l'application visée, un système de RAL peut faire l'objet d'attaques d'individus piquant l'identité de quelqu'un d'autre. Donc ce système doit être robuste face à de telles tentatives d'imposture, comme par exemple, les attaques par dessein des transactions frauduleuses sur le compte bancaire d'un client ou l'accès à des données confidentielles.

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 et ses applications

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. 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 quatre grandes classes : l'identification, la vérification, la détection et l'indexation comme on peut le constater sur la figure I.3.

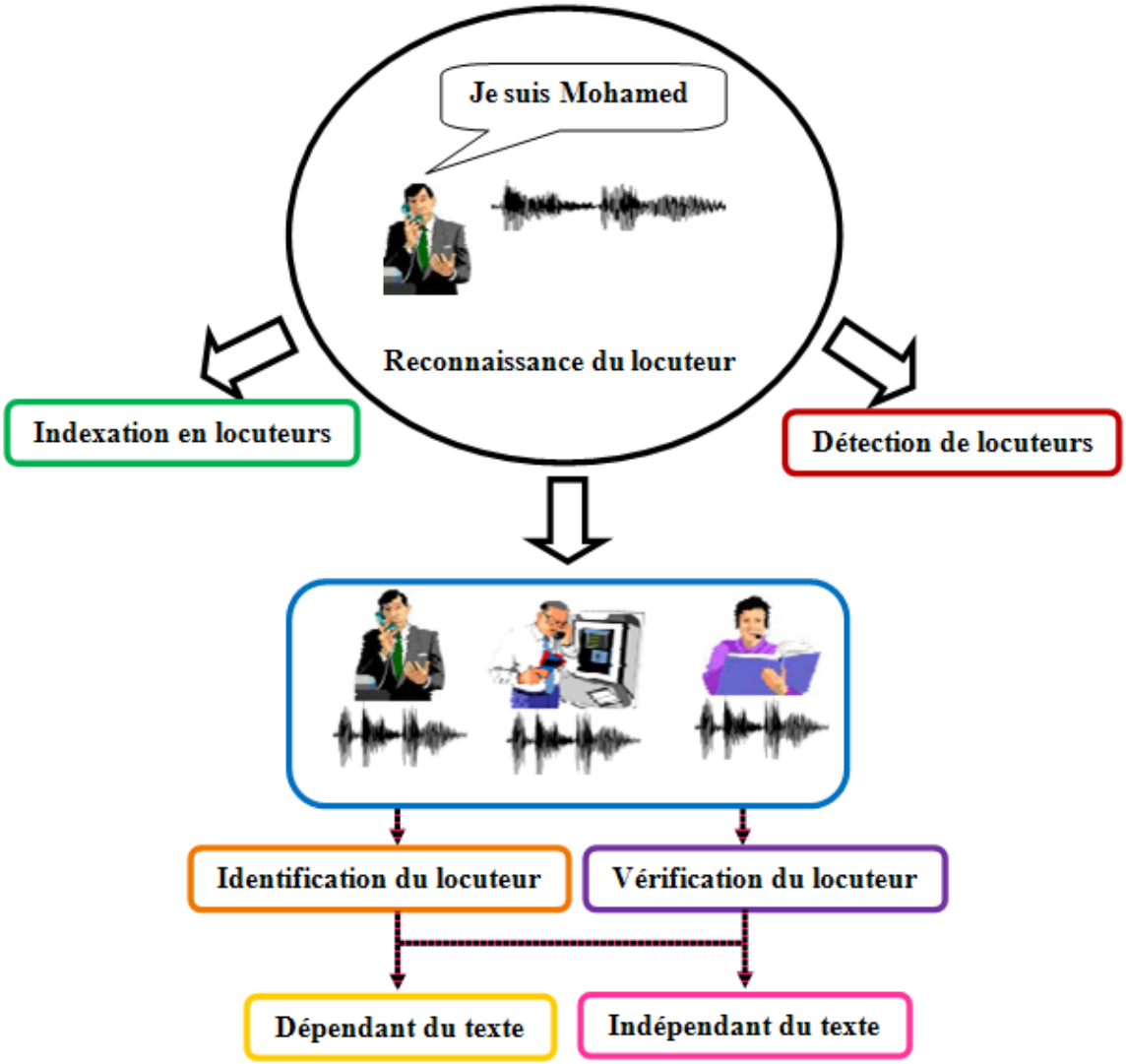


Figure I.3 : Principales applications liées à la reconnaissance automatique de locuteur.

Les applications de la reconnaissance du locuteur sont nombreuses. Elles couvrent à peu prés tous les secteurs dans lesquels il est souhaitable de sécuriser des actions, des transactions ou toutes sortes d'interactions, en identifiant où en vérifiant l'individu qui effectue l'opération. Le secteur bancaire ou celui des télécommunications sont à l'heure actuelle les plus demandeurs de cette technologie. Compte tenu du fait qu'elle n'offre pas un niveau de sécurité absolu, la reconnaissance du locuteur doit être utilisée de manière complémentaire à d'autres outils de sécurisation ou d'investigation.

On distingue plusieurs profils d'applications de la reconnaissance automatique du locuteur : le contrôle d'accès physique, la sécurisation de transactions à distance, l'organisation de l'information sonore et enfin les applications criminalistiques. Toutes ces

applications se distinguent par les contraintes de nature diverse qu'elles imposent au système de RAL, et se décrivent par les tâches associes. Diverses applications reposant sur une même tâche peuvent se différencier entre autres par leur degré d'indépendance où dépendance au texte. Les systèmes de RAL dits indépendants du texte ne tiennent aucun compte du contenu linguistique du signal de parole. `A l'oppose, les systèmes dits dépendants du texte utilisent la connaissance de tout ou partie de ce contenu linguistique pour affiner la reconnaissance du locuteur

I.4.1 Contrôle d'accès physique

Les applications de types contrôles d'accès physiques sont les applications nécessitant la présence effective de l'utilisateur devant le système pour réaliser l'opération souhaitée, celle-ci nécessitant une interaction matérielle en un endroit précis. A titre d'exemple on peut citer : l'accès à un lieu (domicile, véhicule), à des objets (coffre-fort), à de l'argent (distributeur automatique) ou à un terminal spécifique (poste de travail).

La reconnaissance automatique du locuteur peut être considérée comme un dispositif d'identification pour ces applications. Elle peut se faire sur site ou à distance quand la qualité de la transmission peut être contrôlée.

I.4.2 Applications dans le domaine criminalistique

L'utilisation de la reconnaissance automatique du locuteur dans les domaines judiciaires ou criminalistique peut aller jusqu'à l'orientation d'une enquête, la recherche de suspects ou la constitution d'éléments de preuves. Ceci doit faire face à des situations dont la difficulté et la diversité sont considérables. La voix anonyme a souvent été enregistrée dans des mauvaises conditions techniques, dans des situations de stress qui ont une influence sur la voix. Par contre, l'enregistrement de la voix du suspect est généralement obtenu dans des conditions différentes (en général bien meilleures). Par ailleurs, le suspect peut être plus ou moins coopératif lors de la procédure d'enregistrement. Dans ce contexte, il est important de souligner que la voix est très souvent assimilée, à tort, à une empreinte vocale au même titre que les empreintes digitales ou génétiques et peut constituer une preuve dans une procédure pénale. Le terme d'empreinte vocale est une aberration sachant que la voix ne possède pas de caractéristiques qui peuvent la rendre unique, elle peut être considérée comme une signature vocale.

I.5 Différentes tâches en RAL

Toutes les applications de la RAL reposent sur des principes communs et peuvent être définies comme variantes de quelques tâches de base, dont les principales sont l'identification et la vérification considérés comme des tâches pionnières de la RAL [16], [17]. 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) consiste à retrouver l'identité de locuteur ayant prononcé un message donné, parmi une population de locuteurs connus [18]. La figure I.4 illustre le principe du system d'IAL, à l'entrée de ce système un signal de test donné est comparé aux signaux des locuteurs références. L'identité du locuteur dont la référence est la plus proche du signal test est donnée en sortie du système.

Deux modes sont présentés en IAL : l'identification en ensemble fermé suppose que le locuteur à identifier est forcement un des locuteurs connus du système. L'identité du locuteur le plus probable parmi les locuteurs références est retournée en sortie du système. L'identification en ensemble ouvert, le locuteur peut ne pas être connu du système. Dans ce cas le système associe au locuteur le plus probable une étape de test de la fiabilité, en acceptant ou rejetant l'identité du locuteur test.

En général, les performances de système d'IAL se dégradent au fur et à mesure que le nombre des locuteurs références augmente, ainsi les applications de ce système sont peu nombreuses [19]. On peut citer, par exemple, pour les systèmes qui reposent sur le principe de l'identification en ensemble ouvert, la sécurité par authentification vocale surtout pour des applications commerciales associant un même mot de passe pour une petite population de locuteurs (membres d'une famille, d'une société). En ensemble fermé, l'intégration de l'IAL au sein d'un système de reconnaissance de la parole est faite pour assurer une adaptation automatique à l'utilisateur.

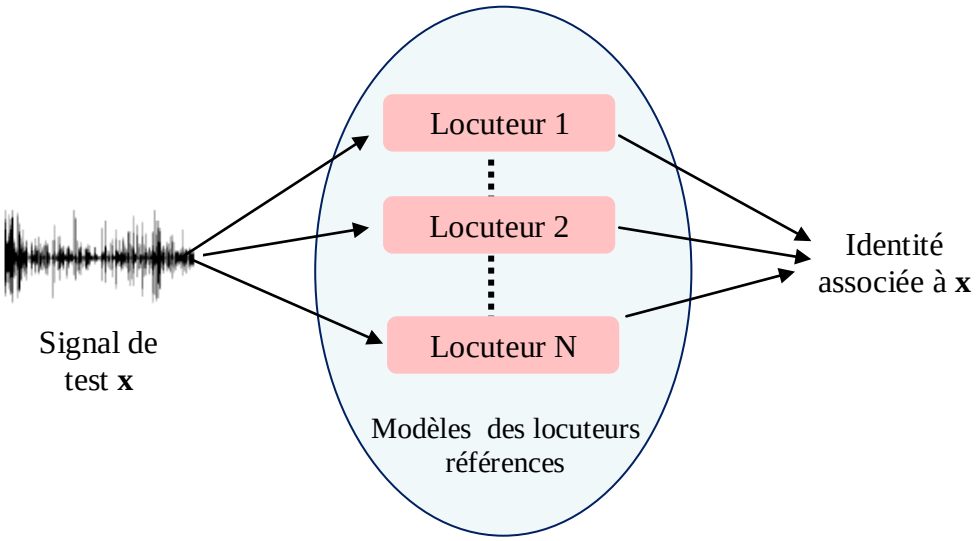


Figure I.4: Principe de l'identification automatique du locuteur.

I.5.2 Vérification automatique du locuteur

La Vérification Automatique du Locuteur (VAL) consiste à décider si l'identité proclamée d'un message vocal de test correspond à la véritable identité du locuteur, par un jugement binaire ; 1 pour l'acceptation ou 0 pour le rejet (figure I.5) [16], [17], [18], [20]. La tâche de VAL nécessite donc un signal test d'identité proclamée et la référence associée à l'identité proclamée. Une mesure de similarité entre le signal à vérifier et cette référence est calculée, Ensuite, cette mesure est comparée à un seuil de vérification. Le locuteur test est accepté si la mesure de similarité est supérieure au seuil, si non l'individu est considéré comme un imposteur et rejeté. Ce principe simple trouve son utilité dans un grand nombre d'applications commerciales [19]:

- 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.

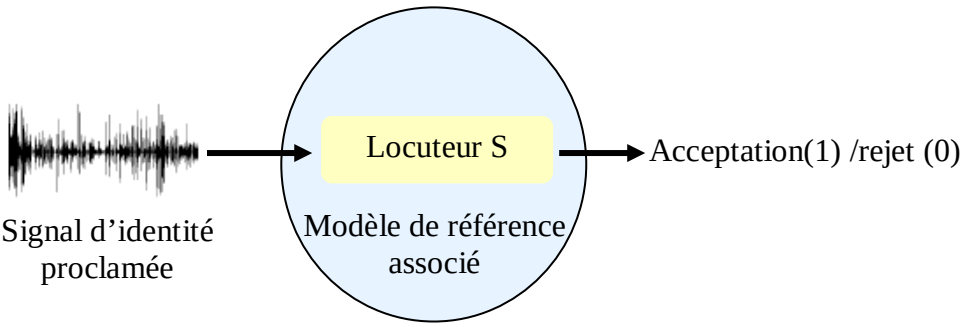


Figure I.5: Principe de la vérification automatique du locuteur.

I.5.3 Détection de locuteurs

Le principe de la détection consiste à déterminer si un locuteur donné intervient ou non dans un flux audio composé de séquences de parole produites par plusieurs locuteurs (conversations, débats, conférences, etc.) [21], [22]. Dans le cas d'un flux audio mono-locuteur, la tâche de détection se résume à la tâche de vérification. Les applications de cette tâche sont souvent motivées par les instances sécuritaires 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).

I.5.4 Indexation par locuteurs

La tâche d'indexation par Locuteurs permet de déterminer les temps de parole et les interventions des locuteurs dans un document audio. En d'autres termes, indexer un document audio en locuteurs en indiquant à quel moment une personne prend la parole et qui est cette personne. Le document à indexer est la seule entrée d'un système d'indexation. Ce système ignore le nombre de locuteurs présents dans le document ou leur identité, aussi le système ne possède pas de référence pour ces locuteurs. Un d'apprentissage aveugle et adaptatif est alors mis en place [23]. Il est possible de segmenter un flux audio par prise de parole des intervenants, étiqueter des données audio pour identifier le nombre de locuteurs présents dans le flux audio à indexer. Sur la figure I.6, la sortie du système d'indexation ressemble la séquence suivante : le locuteur A est intervenu aux instants t_1, t_3, le locuteur B aux instants t_2, t_5, le locuteur C à l'instant t_4.

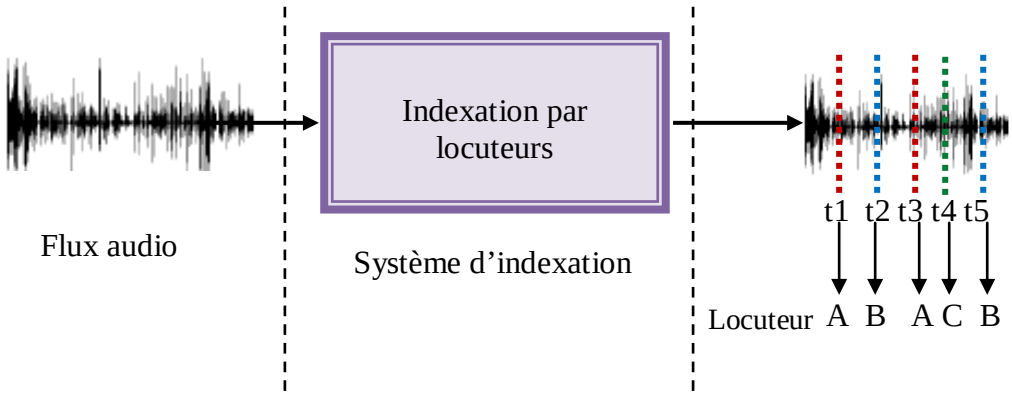


Figure I.6: Principe d'indexation par Locuteur dans un flux audio.

I.5.4.1 Suivi de locuteur

La tâche de suivi de locuteurs est une extension naturelle et simplifiée de l'indexation en locuteur d'un flux audio (figure I.7), à ceci près que les locuteurs présents dans le flux audio sont connus par le système de RAL. Il s'agit donc d'une simplification de la tâche d'indexation en locuteur. Elle consiste à trouver les frontières des interventions du locuteur recherché au sein du document multi-locuteurs, et ainsi de déterminer si ce locuteur intervient et si oui, quand [24].

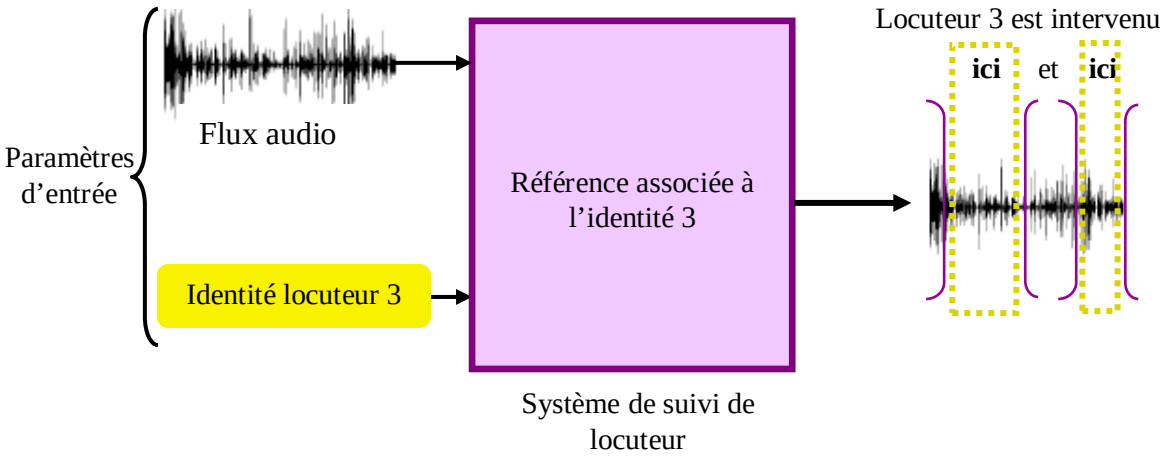


Figure I.7: Principe de base du suivi de locuteurs.

I.5.4.2 Segmentation en locuteurs

La segmentation en locuteurs consiste à déterminer le nombre de locuteurs présents dans un flux audio tout en délimitant leurs interventions (figure 8). La segmentation traite un flux audio pour lequel peu ou pas d'informations sont connues a priori. Elle doit extraire le nombre des locuteurs intervenant dans le flux et leur identité. Malgré la complexité de la segmentation. le champ d'application de la segmentation en locuteurs s'est étendu et cette variante se retrouve intégrée dans un cadre plus vaste de l'indexation en locuteurs de bases de données de documents multimédia [23], [25].

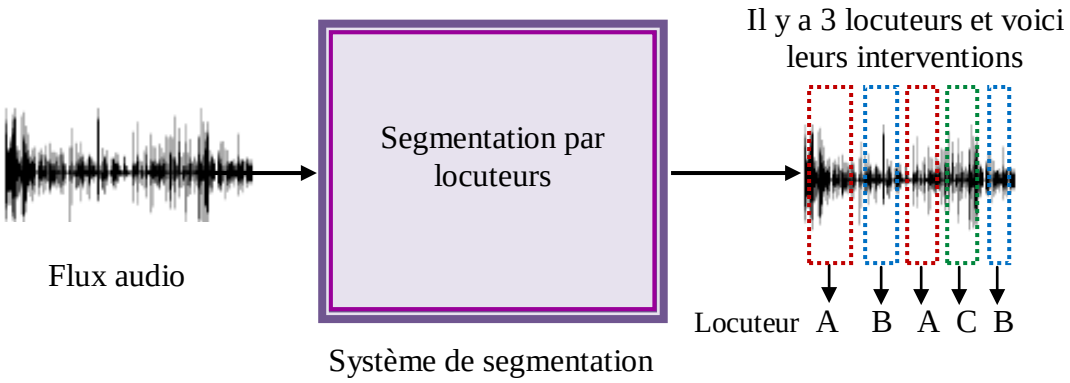


Figure I.8 : Principe de base de la segmentation en locuteurs.

I.6 Structure d'un système de RAL

La reconnaissance automatique du locuteur peut être interprétée comme une tâche particulière de reconnaissance de formes. Elle repose sur trois modules principaux à savoir ; l'analyse acoustique du signal parole, la modélisation du locuteur, et comme dernière étape la décision. La figure I.9 montre les différents constituants d'un système de vérification automatique de locuteur. Tout d'abord, les signaux vocaux, issus de différents locuteurs, captés par un microphone, sont convertis, numérisés puis enregistrés sur une base de donnés d'apprentissage. Ces signaux sont ensuite analysés et traités dans un étage d'analyse acoustique. A l'issu de ce traitement dans ce module, les signaux d'apprentissage sont représentés par des vecteurs de coefficients pertinents pour la modélisation des locuteurs au niveau de la phase d'apprentissage. A la phase de test, un module de reconnaissance va mesurer la similarité entre les paramètres acoustiques du signal prononcé et les modèles de

locuteurs présents dans la base des modèles. En dernier lieu, un étage de décision, basé sur une stratégie de décision donnée, fournit la réponse du système.

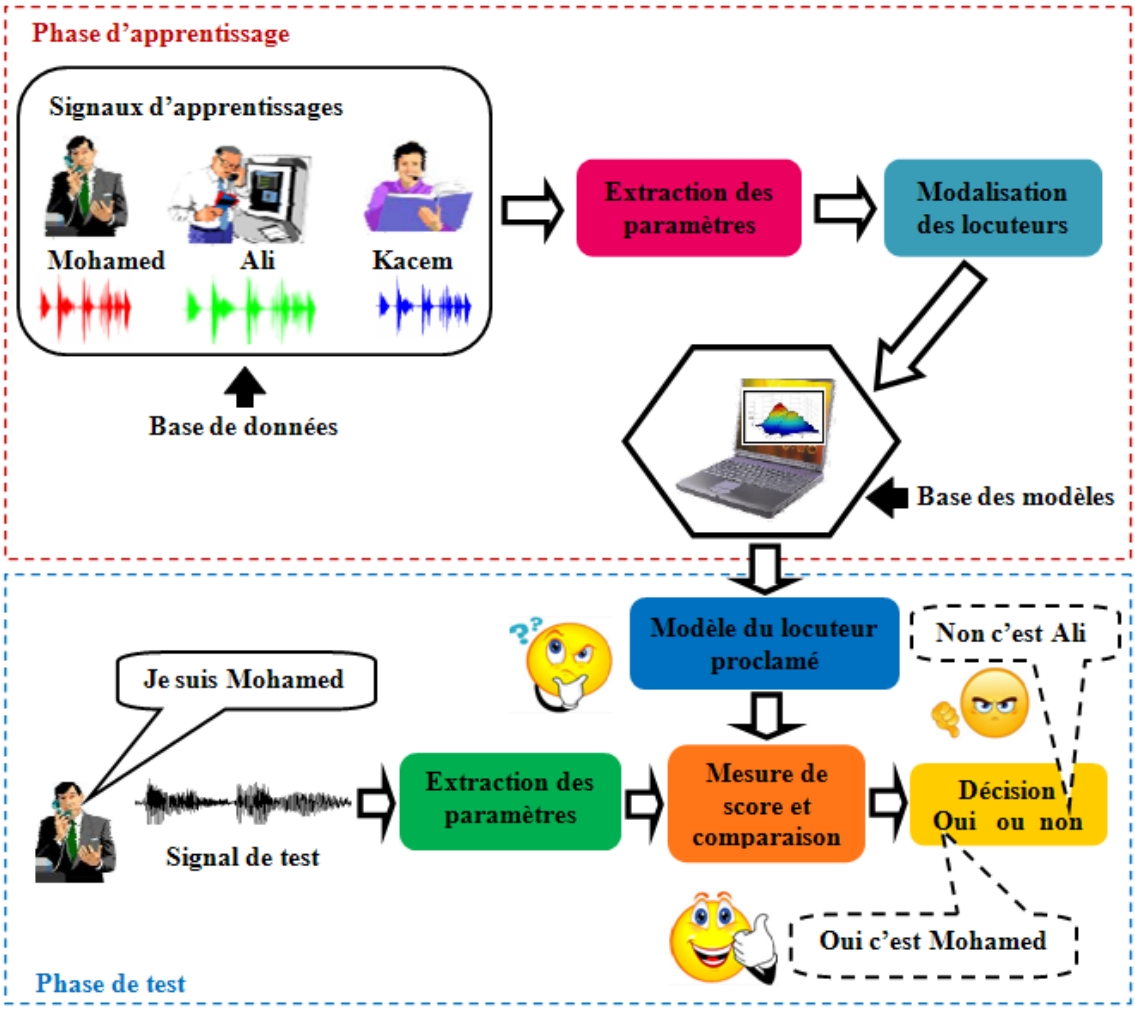


Figure I.9: Bloc diagramme d'un système de vérification automatique de locuteur.

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 prosodiques, les paramètres de l'analyse spectrale et les paramètres dynamiques.

I.7.1 Paramètres prosodiques

Le terme paramètres prosodiques, réunit la mélodie (en rapport avec la variation de la fréquence fondamentale), le stress ou l'accentuation (en relation avec l'énergie du signal) et le rythme (en relation avec la variation de durée des phonèmes et des pauses). L'énergie et la fréquence fondamentale (ou pitch) sont les plus utilisés. 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. 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 de l'énergie à évaluer.

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

$x(k)$**:** Le signal numérique.

I.7.1.2 Fréquence fondamentale

La période du fondamental est par définition la fréquence de vibration des cordes vocales. Elle est appelée aussi le pitch. C'est une caractéristique particulière de la parole et constitue un paramètre très important dans les différentes applications de la reconnaissance du locuteur [26]. 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 la dynamique de la fréquence du fondamental est très grande. Elle s'étend approximativement de : 80 à 200 Hz chez les hommes (voix grave), de 150 à 450 Hz chez les femmes et de 200 à 600 Hz chez les enfants (voix aiguës) [27].

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 (en particulier les formants) auxquelles on attache une grande importance. Pour cela on introduit en traitement du signal les outils suivants :

I.7.2.1 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 :

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

La transformée 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 :

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

Cet algorithme 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 Transformée de Fourier à court 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 locale sur une portion de signal de faible durée, limitée par une fenêtre temporelle (par exemple fenêtre de Hamming), pendant laquelle le signal est quasi stationnaire, généralement de 10 à 32 millisecondes.

I.7.2.3 Analyse LPC (Linear Predicting Coding) [28] [29]

L'analyse par prédiction linéaire se fonde sur la corrélation entre les échantillons successifs du signal. L'échantillon s(n) à l'instant n, peut être prédit approximativement comme une combinaison linéaire des p échantillons précédents :

A partir de ses valeurs aux instants $n-1, n-2$ ... *s(n)* peut être prédit par :

$$\hat{s}(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-p) = \sum_{i=1}^{p} a_i s(n-i) \qquad \text{(I.6)}$$

Les coefficients de prédiction a_i sont supposés constants sur une fenêtre n d'analyse du signal. En introduisant une excitation normalisée $v(n)$ et un gain d'excitation G on obtient :

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + Gv(n) \qquad \text{(I.7)}$$

$Gv(n)$ est identifiée à l'erreur de prédiction introduite par le modèle, ou résidu d'ordre p:

$$\mathcal{E}(n) = s(n) - \hat{s}(n) \qquad \text{(I.8)}$$

Soit la transformée en z de la fonction (I.7):

$$S(z) = \left(\sum_{i=1}^{p} a_i z^{-i}\right) S(z) + GV(z) \qquad \text{(I.9)}$$

Où $V(z)$ désigne la transformée en z de $v(n)$.

On défini un filtre linéaire de prédiction dont la fonction de transfert est :

$$H(z) = \frac{S(z)}{GV(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}} = \frac{1}{A(z)} \qquad \text{(I.10)}$$

Ce filtre autorégressif (AR) tout-pôles, représenté sur la figure I.10 peut être assimilé au modèle acoustique linéaire de production de la parole [30], [31], [32], repose sur le cadre théorique du modèle source-filtre, où les différents sons voisés et non voisés sont modélisés par différentes sources :

- Les sons voisés sont des sons quasi-périodiques. Ils sont caractérisés par la présence de la fréquence fondamentale (pitch *F*0). par le biais d'excitations périodiques.
- Les sons non voisés sont assimilables à du bruit. Ils sont caractérisés par des durées très courtes et des fortes variations d'amplitude.

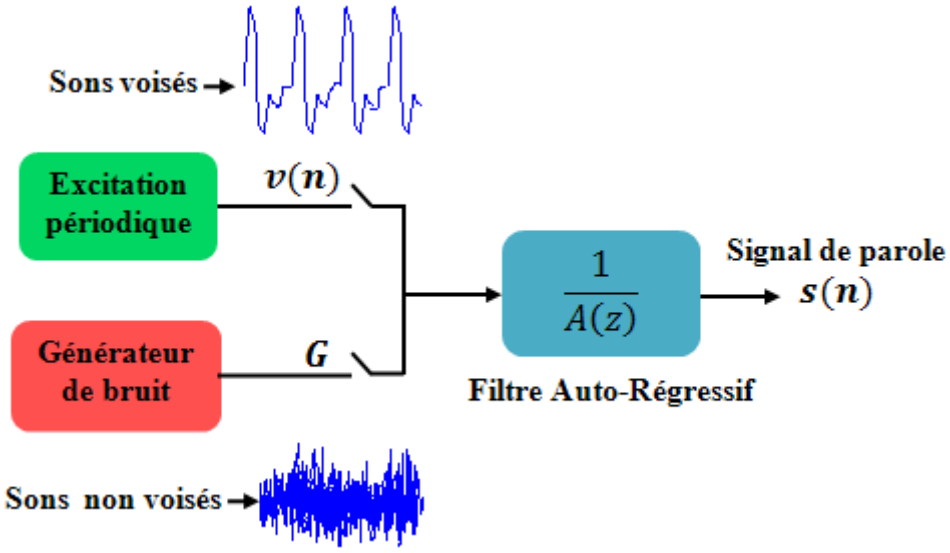


Figure I.10: Modèle autorégressif (AR) de la prédiction linéaire.

De manière à utiliser la prédiction linéaire pour reconstruire le signal $s(n)$, nous pouvons définir l'erreur de prédiction $\mathcal{E}(n)$ qui est la différence entre le signal réel $s(n)$ et le signal approximé $\hat{s}(n)$:

$$\mathcal{E}(n) = s(n) - \hat{s}(n) = Gv(n) \tag{I.11}$$

Nous pouvons en déduire la fonction de transfert d'erreur :

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{i=1}^{p} a_i z^{-i} \tag{I.12}$$

On s'intéresse à l'erreur quadratique moyenne commise sur un ensemble de m trames du signal autour de $s(n)$:

$$E_{moy} = \sum_m \mathcal{E}^2(m) = \sum_m \left[s(m) - \sum_{i=1}^{p} a_i s_i(m-i)\right]^2 \tag{I.13}$$

Pour minimiser E_{moy}, on annule ses dérivées partielles par rapport aux coefficients a_i, ce qui fournit un système de p équations. La résolution de ce système d'équations peut être effectuée à l'aide de deux méthodes : La méthode de covariance ou la méthode d'autocorrélation [33].

Les coefficients a_i étant calculés, nous pouvons procéder à la détermination de plusieurs paramètres qui en découlent directement, en l'occurrence les coefficients cepstraux de prédiction linéaire (LPCC).

I.7.2.4 Analyse LPCC (Linear Predictive Cepstral Coefficient) [34]

Les paramètres acoustiques LPCC (c_m) sont calculés à partir des coefficients (a_p) du codage LPC par l'algorithme récursif suivant:

$$\begin{cases} c_0 = \ln G \\ c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \rightarrow 1 \leq m \leq p \\ c_m = \sum_{k=m-p}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k} \rightarrow m > p \end{cases} \quad \text{(I.15)}$$

I.7.2.5 Analyse MFCC (Mel Frequency Cepstral Coefficient)

La paramétrisation MFCC est la plus utilisée en reconnaissance automatique de la parole [35] [36] [37], et notamment en reconnaissance du locuteur. 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 de la même manière et également à toutes les fréquences.

❖ **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 1 kHz, alors qu'il est logarithmique pour le reste des fréquences (supérieures à 1 kHz). Nous pouvons donc utiliser la formule approximative suivante afin de faire correspondre à chaque fréquence en Hz une fréquence sur l'échelle Mel:

$$f_{mel} = 2595 \times \log_{10}\left(1 + \frac{FHz}{700}\right) \quad \text{(I.16)}$$

❖ **L'utilité de l'échelle Mel**

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.11, 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.

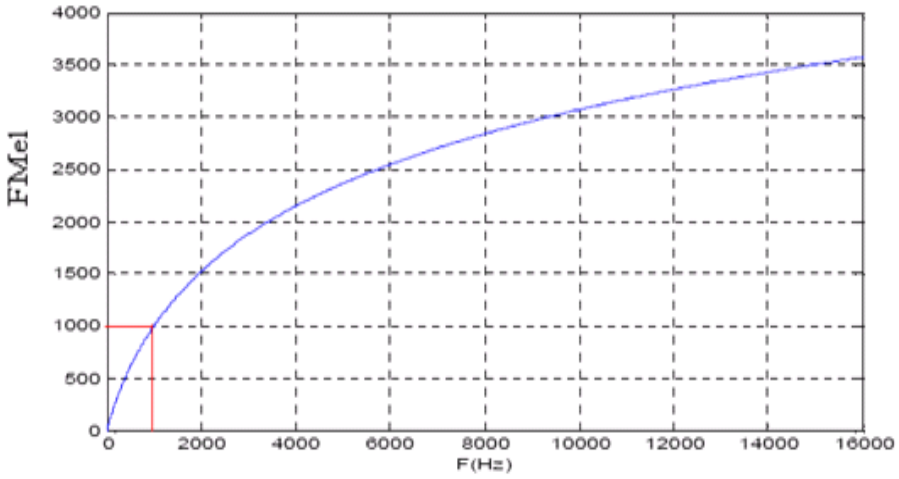


Figure I.11: Transformation du Hz en Mel.

❖ **Calcul des coefficients cepstraux (MFCC) [35] [37]**

Les différentes étapes pour calculer les coefficients MFCCs d'un signal parole sont : la segmentation 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.12).

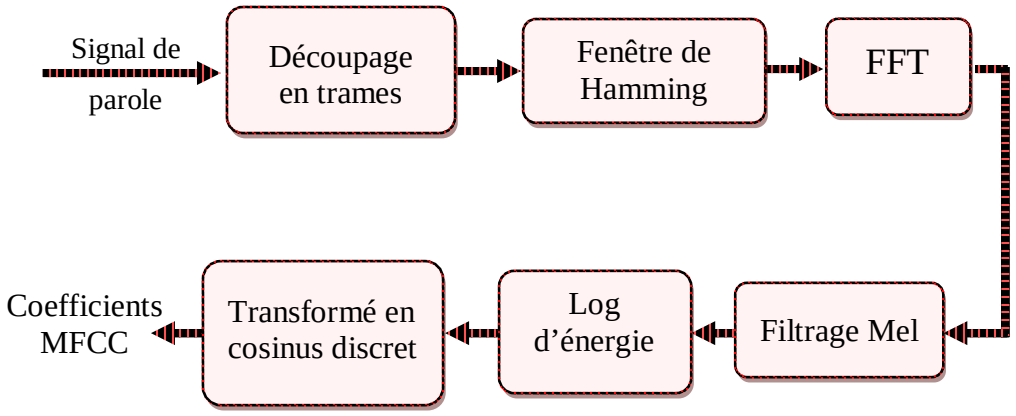


Figure I.12 : Calcul des coefficients MFCC avec une échelle Mel.

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.13. La localisation des fréquences centrales des filtres est donnée par:

$$f_{mel} = 1000 \times \frac{\log(1+f/1000)}{log\,2} \qquad \text{(I.17)}$$

Où f est la fréquence en Hz.

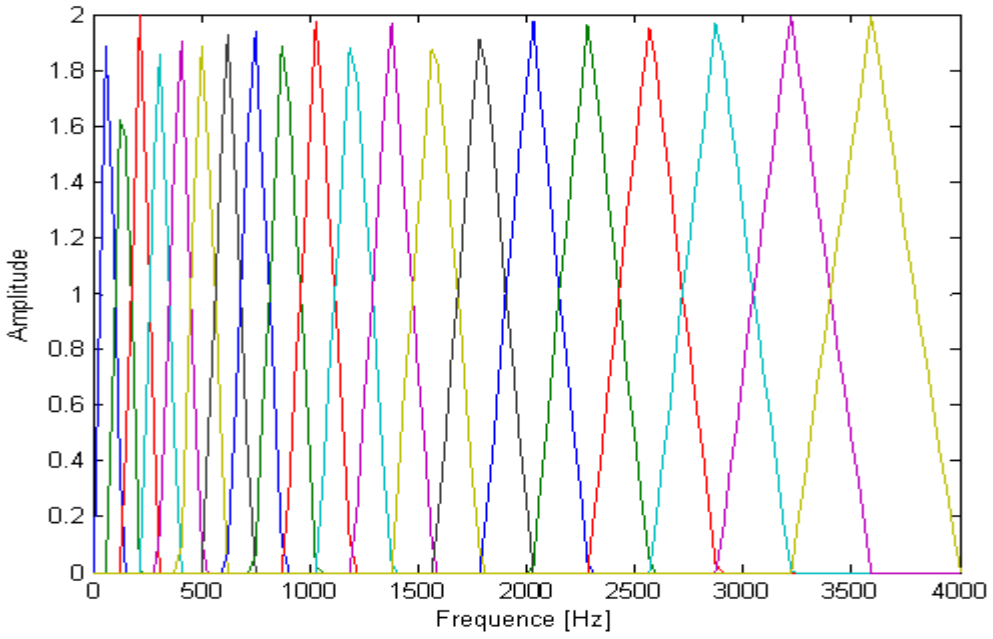


Figure I.13: Banc de filtres triangulaires: équidistance en échelle Mel.

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 [38] [39], l'expression de ces coefficients est données par :

$$c_n = \sum_{k=1}^{K} S_k \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1,2\ldots,L \qquad \text{(I.18)}$$

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.7.2.6 Analyse PLP (Perceptual Linear Predictive)

La méthode PLP, [40], [41], [42], [43], Perceptual Linear Prediction (ou Perceptually based Linear Prediction), est une méthode inspirée du principe de prédiction linéaire. Elle combine ce principe à une représentation du signal qui suit l'échelle humaine de l'audition. Le schéma général de cette méthode est donné à la figure I.14.

Cette méthode peut être résumée en deux phases de traitements successifs. Le signal de parole est tout d'abord analysé pour obtenir un spectre suivant une échelle d'audition. Ce spectre est ensuite modifié par une interpolation et une transformée de Fourier inverse, le signal obtenu étant passé dans un filtre pour réduire la dimension du spectre et augmenter la résolution fréquentielle.

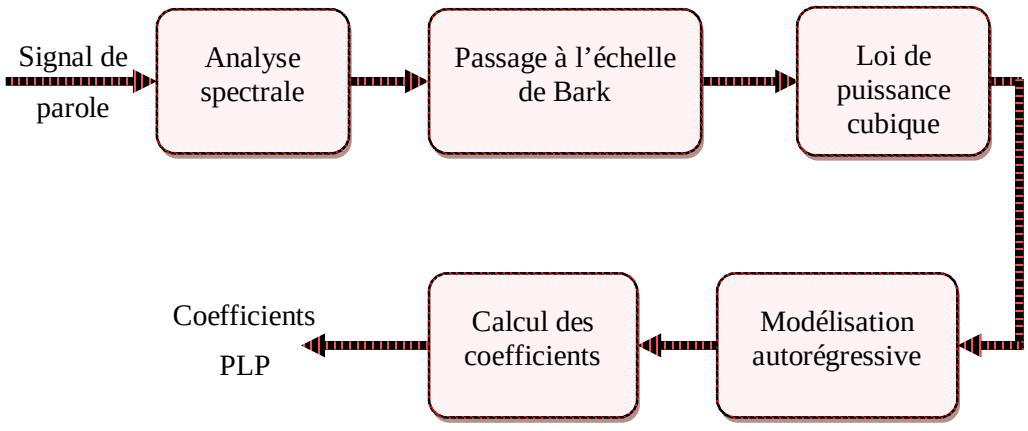


Figure I.14: Méthode de calcul des coefficients PLP.

Dans la première étape il est précisément procédé à :

- une analyse en bandes critiques selon une échelle Bark par un banc de filtres ;
- une préaccentuation des valeurs obtenues selon une courbe suivant approximativement les mêmes principes que les traitements effectués par l'oreille, avec accentuation des basses fréquences et atténuation des hautes fréquences.

Dans la deuxième étape, on procède à :

- une interpolation des sorties des filtres du banc pour obtenir un spectre sur une échelle fréquentielle auditive ;
- une transformée de Fourier inverse qui permet de ramener le spectre obtenu dans le domaine temporel ;
- une résolution d'un ensemble d'équations linéaires pour obtenir les coefficients issus d'un filtre tout pôle d'ordre 5 (ce qui permet d'obtenir au moins deux sommets caractéristiques selon).

Cette méthode a pour avantage de permettre une analyse ou un codage de la parole qui respecte le principe de la prédiction linéaire, qui suit l'échelle fréquentielle observable dans l'oreille et, enfin, elle réduit l'espace de représentation.

❖ **Rasta PLP**

La méthode PLP [40], [41], dont l'algorithme repose sur des spectres à court terme de la parole, résiste difficilement aux contraintes qui peuvent lui être imposées par la réponse fréquentielle d'un canal de communication. Pour atténuer les effets de distorsions spectrales linéaires, H. Hermansky [42] propose de modifier l'algorithme PLP en remplaçant le spectre à court terme par un spectre estimé où chaque canal fréquentiel est modifié par passage à travers un filtre. Cette modification est à la base de la méthode RASTA PLP, RASTA étant l'acronyme de RelAtive SpecTrAl. La mise en place de ce filtrage permet, lorsqu'il est effectué dans le domaine spectral logarithmique, de supprimer les composantes spectrales constantes, supprimant ainsi les effets de convolution du canal de communication.

Différentes études réalisées avec cette méthode [43] ont permis de confirmer les bonnes qualités de cette méthode relativement aux distorsions et ses moindres qualités face aux bruits qualifiés d'additifs, signe de la présence de plusieurs sources sonores dans un même environnement.

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 [44] [45].

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 I.19. 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).

$$\frac{\partial c(t)}{\partial t} \approx \Delta c(t) = \frac{\sum_{K=-N}^{N} K * c(t+K)}{\sum_{K=-N}^{N} K^2} \qquad \text{(I.19)}$$

Où $c(t)$ représente le coefficient à dériver, $\Delta c(t)$ sa dérivée première à l'instant t et où les coefficients Δ sont calculés sur une fenêtre temporelle de longueur $2N + 1$ trames.

I.8 Modélisation des locuteurs

Ce paragraphe parcourt les techniques les plus utilisées en reconnaissance du locuteur. Comme dans le cas de reconnaissance de 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 cinq grandes familles :

- L'approche vectorielle où le locuteur est représenté par un ensemble de vecteurs de paramètres (MFCC, PLP, etc.) dans l'espace acoustique. Ses principales techniques sont la reconnaissance à base de DTW et la reconnaissance par quantification vectorielle [46].
- 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**

L'algorithme DTW (Dynamic Time Warping) appliqué à la reconnaissance vocale [46] [47], 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.

❖ **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 [48] [49]. 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.

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 [48] [49].

I.8.2 L'approche statistique

L'approche statistique repose sur la modélisation de la distribution des vecteurs de paramètres correspondant `a un locuteur. Ce principe relativement simple offre néanmoins de très bonnes performances, notamment en mode indépendant du texte avec l'apparition des GMM.

- **Méthodes Statistiques du Second Ordre**

Le principe des Méthodes Statistiques du Second Ordre (MSSO) est de représenter une séquence de vecteurs acoustiques par une distribution gaussienne multi-dimensionnelle. Le modèle d'un locuteur se résume alors par le triplet $\{\mu, \sum, X\}$ où μ est un vecteur moyen, $\sum$ est une matrice de covariance, qui sont tous les deux estimés à partir de la séquence de X vecteurs acoustiques.

Les MSSO sont généralement associées à des mesures de similarité particulières en vue de la reconnaissance. Ces mesures ont pour particularité de faire intervenir le triplet $\{\mu_0, \sum_0, \gamma\}$. Ce dernier est estimé sur la séquence de vecteurs de test de manière analogue au triplet $\{\mu, \sum, X\}$. Les mesures reposent ainsi essentiellement sur une ressemblance entre les matrices $\sum$ et $\sum_0$ [50].

- **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 [51].

❖ **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 [52] [53]. 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 [53]. Nous détaillerons cette méthode dans le chapitre quatre.

I.8.3 L'approche connexionniste

Les réseaux de neurones ont été assez largement utilisés en reconnaissance du locuteur [54] [55]. 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.

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 [56]. 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 [57].

I.8.5 L'approche discriminante

La plus employée en RAL sont les Support Vector Machine (SVM) [58]. 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 à l'approche GMM. Ces deux méthodes sont aussi combinées dans un nouveau formalisme [59], par exemple le GMM-SVM Super-Vecteur qui profite des capacités génératives du GMM et discriminantes du SVM.

I.9 Prise de décision

La prise de décision en RAL basée sur le formalisme probabiliste est différente selon la tâche choisie. Le plus souvent, cette stratégie se formalise dans un cadre bayésien pour les tâches de vérification et d'identification.

I.9.1 Décision en identification

En identification de locuteur, 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.

$$Taux\ d'identification\ correct\ \% = \frac{Tests\ \ correctement\ \ \ identifi\ és}{Tests\ \ totales} \qquad (I.20)$$

Notons que les performances d'identification du locuteur diminuent avec l'augmentation du nombre des locuteurs de la base de données [60]. Il faut aussi noter que, dans ce cas, les ressources nécessaires et les temps de traitement augmentent.

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 :

$$p(FA) = \frac{Nomdre\ \ d'imposteurs\ \ accept\ és}{Nomdre\ \ d'acc\,ès\ imposteurs} \tag{I.21}$$

$$p(FR) = \frac{Nomdre\ \ de\ clients\ \ rejet\ és}{Nomdre\ \ d'acc\,ès\ clients} \tag{I.22}$$

Contrairement à l'identification, les performances de vérification ne sont pas affectées par la taille de la population de référence, puisque la tâche à accomplir se ramène toujours à un choix binaire entre deux hypothèses.

I.10 Mesures de performances

La thèse étant centrée sur la tâche de vérification du locuteur (VAL), nous nous intéressons uniquement aux techniques d'évaluation pour cette tâche. L'évaluation de la qualité d'un système de VAL dépend de plusieurs facteurs. Ce sont les performances en termes de taux d'erreurs qui vont en déterminer la qualité.

I.10.1 Faux Rejet (FR) et Fausse Acceptation (FA)

Les performances d'un système de VAL s'évaluent en fonction de deux taux d'erreurs. La probabilité de fausses acceptations FA (False alarms), correspondant à l'acceptation à tort d'imposture et la probabilité de faux rejets FR (Miss probability), correspondant au rejet à tort du client correspondant à l'identité proclamée. C'est à partir des taux d'acceptation PFA et de rejet PFR que l'évaluation du système est réalisée. 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 comme application de sécurité, ou minimiser le taux de faux rejets pour augmenter l'ergonomie d'utilisation (figure I.15). Ainsi il n'est pas possible de minimiser conjointement ces deux taux. Le taux d'erreurs égales ou EER (Equal Error Rate) est le point de fonctionnement où $FA = FR$, cette mesure est très utilisée pour comparer les performances des systèmes de RAL [61].

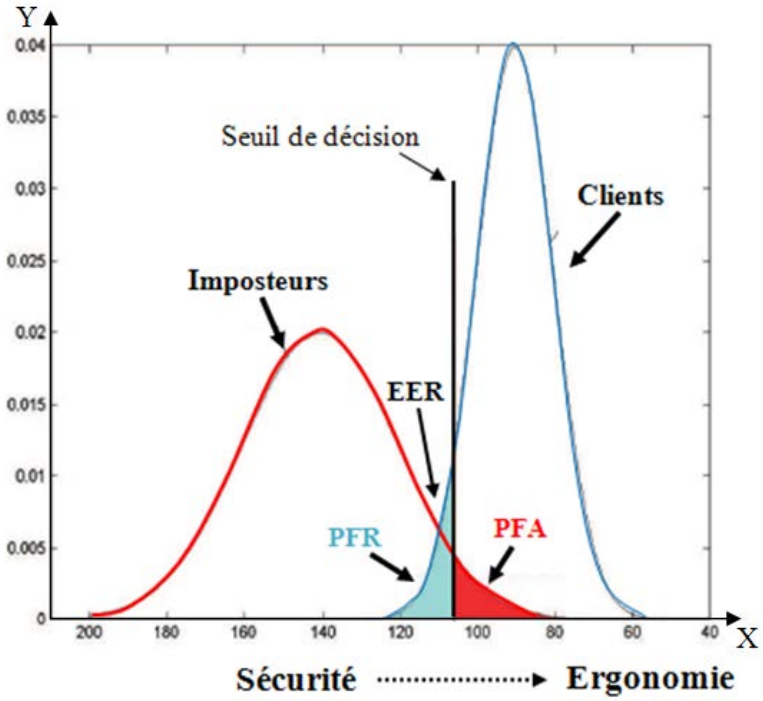


Figure I.15: Types d'erreurs dans un système VAL.

I.10.2 Les courbes DET (Detection Error Tradeoff)

La courbe DET (Detection Error Tradeoff curve) illustrera les résultats expérimentaux présentés dans ce document, elle a été présentée par [62] comme une variante des courbes ROC (Receiving Operating Characteristic) [63]. Cette courbe DET permet d'évaluer, pour chaque seuil de vérification les valeurs du couple FA, FR. La figure I.16 illustre un exemple de courbe DET, dont les axes sont les taux d'erreurs PFA et PFR. 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 [64].

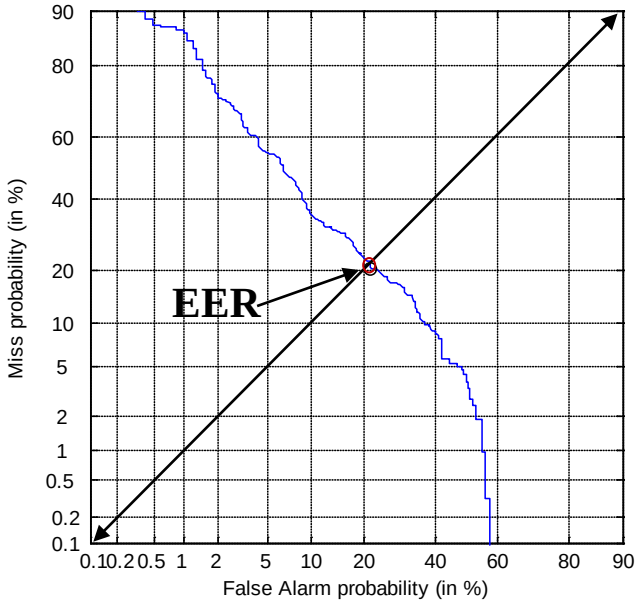


Figure I.16: Courbe DET ainsi que l'EER.

I.11 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 et l'extraction de vecteurs caractéristiques ;
- la modélisation du locuteur ;
- la décision.

Dans ce chapitre, nous avons décrit 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 que sont la modélisation du locuteur et la décision quand a son acceptation ou rejet (Vérification automatique de locuteur), ou la décision dans la détermination de son identité (Identification automatique de locuteur).

CHAPITRE II

Les réseaux NGN et VoIP

II.1 Introduction

Les réseaux NGN (Next Generation Network) représentent la nouvelle génération de réseaux censée réaliser la convergence totale des services (voix, vidéo et données) en une seule architecture réseaux, le Tout-IP ou ALL-IP (figure II.1), qui englobe le développement de nouveaux services basés sur le protocole IP (Internet Protocol).

C'est dans ce contexte que ce deuxième chapitre est consacré à la présentation des réseaux NGN. Dans la première partie, nous décrivons l'architecture des réseaux NGN, puis nous procéderons à l'étude des principales caractéristiques, ainsi qu'aux différents protocoles utilisés dans ces réseaux. La deuxième partie de ce chapitre est consacrée à un service directement lié à l'évolution vers les réseaux NGN; à savoir le service de la voix sur IP (VoIP).

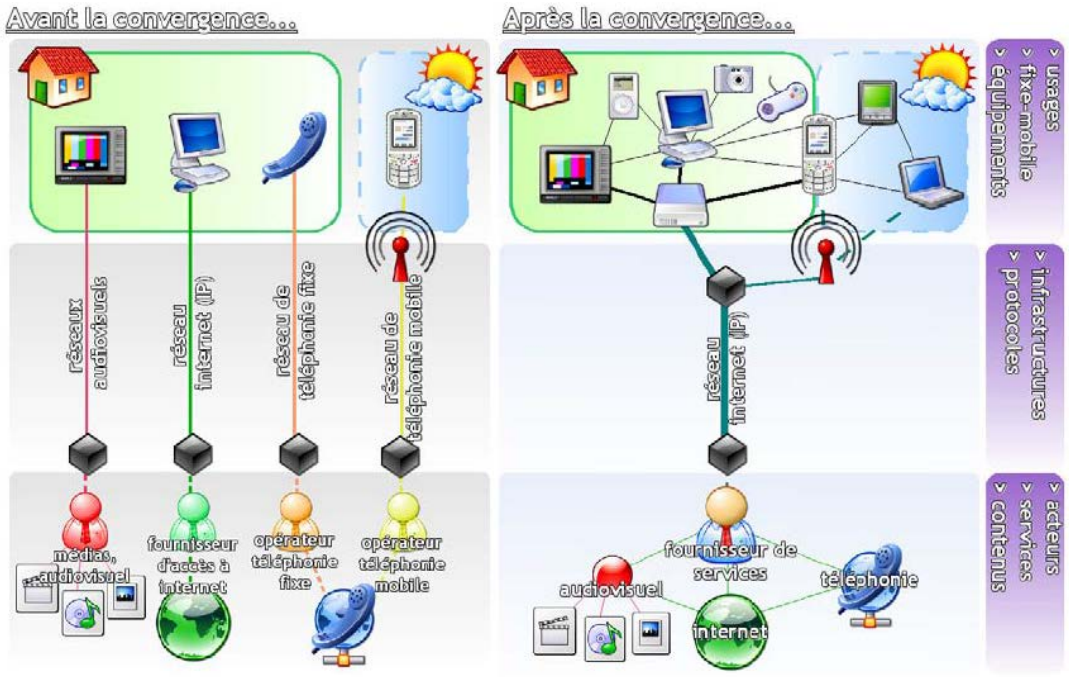


Figure II.1: La convergence Tout-IP [64].

II.2 Les réseaux de nouvelle génération NGN

II.2.1 Définition

Deux recommandations à propos des réseaux NGN sont publies par L'ITU-T (International Telecommunication Union - Telecommunication Standardization Sector). La première recommandation, Y.2001 [65], décrit les principales caractéristiques d'un réseau NGN, tandis que la deuxième, Y.2011 [66], offre une architecture fonctionnelle. Le groupe technique TISPAN (Telecom and Internet converged Services and Protocols for Advanced Neworks) de l'ETSI (European Telecommunications Standards Institute) a aussi défini une architecture fonctionnelle pour les réseaux NGN largement inspirée de celle proposée par l'ITU-T. L'ensemble de ces documents propose un modèle de conception pour les NGN. La toute première recommandation publiée par l'ITU-T, définit un réseau NGN comme un "réseau en mode paquet, en mesure d'assurer des services de télécommunication et d'utiliser de multiples technologies de transport à large bande à qualité de service imposée et dans lequel les fonctions liées aux services sont indépendantes des technologies sous-jacentes liées au transport [65]. L'ETSI a présenté les réseaux de nouvelle génération comme des concepts communs basés sur une évolution progressive vers le « Tout IP ». Ils reposent sur une architecture en couches indépendantes (transport, contrôle, services) communiquant via des interfaces ouvertes et normalisées. Les services doivent être évolutifs et accessibles indépendamment du réseau d'accès utilisé [67]. Les définitions des organismes de normalisation, l'ETSI et l'ITU-T, s'accordent globalement pour définir les NGN comme un réseau de transport en mode paquet permettant la convergence des réseaux Voix/données et Fixe/Mobile, ces réseaux permettront de fournir des services multimédia accessibles depuis différents réseaux d'accès.

II.2.2 Intérêts des réseaux NGN

Les réseaux traditionnels de téléphonie fixe des opérateurs historiques sont basés sur la commutation de circuits entre les lignes d'abonnés, et sur une organisation hiérarchique des commutateurs selon différentes zones d'appels. La problématique de passage à une architecture NGN s'inscrit avant tout dans une logique de diminution des coûts, avec le passage à une infrastructure unique basée sur IP pour le transport de tout type de flux, voix ou données, et pour toute technologie d'accès (DSL, FTTH, RTC, WiFi, etc.). L'impact majeur d'un passage à une architecture NGN pour les réseaux de téléphonie commutée est que le

commutateur traditionnel est scindé en deux éléments logiques distincts : Le media Gateway pour assurer le transport et le Softswitch pour assurer le contrôle d'appel. Une telle évolution permet théoriquement des gains en termes de performance et d'optimisation des coûts, mais elle peut aussi faciliter le déploiement de nouveaux services.

Toutefois cette évolution amène à repenser l'architecture réseau, dont le but principal est l'amélioration de l'évolutivité du réseau :

- La montée des trafics de données : Les trafics données dépassent les trafics de téléphonie dans plusieurs pays. Le transport de données devient structurant :
  1. Déploiement de réseaux de transport paquets (ATM, IP).
  2. Mutualisation du transport (voix et données) sur un réseau paquets.
- Amélioration de l'évolutivité du réseau.
- Possibilité pour les utilisateurs d'avoir un accès non restreint aux opérateurs de leurs choix avec une multiplicité de services de télécom.
- Assurer l'inter fonctionnement avec des réseaux basés sur des technologies anciennes (commutation de circuit).

II.2.3 Architecture des Réseaux NGN

L'UTI-T présente une architecture d'un réseau NGN scindée en deux couches [68], une couche de transport et une couche service.

La couche transport a pour rôle l'acheminement des données entre les différents équipements, quelque soit la technologie d'accès utilisée. Ses fonctions sont :

- Adapter le protocole de commutation de paquets à la couche de liaison physique,
- Authentifier le terminal sur le réseau,
- Gérer le transport des données (qualité de service, accès aux ressources, etc.),
- S'interconnecter physiquement avec les réseaux traditionnels (passerelles).

La couche service est relative à l'utilisation du réseau et offre :

- L'authentification de l'utilisateur sur le réseau et les services associés,
- La fourniture de services et le contrôle d'accès à ces derniers.

Pour faire travailler de concert ces deux couches, il existe une couche transversale : la couche de gestion. Elle effectue une jonction entre les couches service et transport.

La figure ci-dessous présente l'architecture d'un réseau NGN inspirée à partir de la recommandation UTI-T [68].

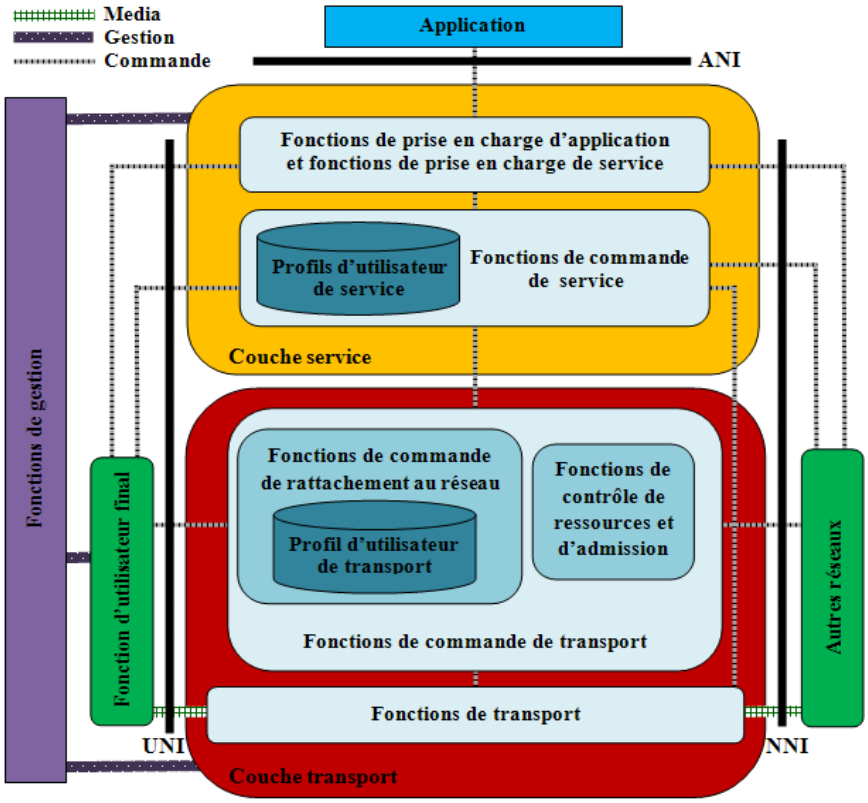


Figure II.2: Architecture d'un réseau NGN [68].

L'UTI-T définit trois interfaces dans un réseau NGN [68]:

- UNI (User-Network Interface) - Démarcation entre l'utilisateur et le réseau.
- ANI (Application-Network Interface) : Séparation logique entre le réseau et les applications ;
- NNI (Network-to-Network Interface) : Interconnexion entre un NGN et d'autres réseaux (NGN ou traditionnels).

Les entités de chaque zone, délimitées par les interfaces, interagissent les unes avec les autres à l'aide de flux. Ces flux sont les liens entre les différents éléments qui composent le réseau et son écosystème. Nous distinguons trois catégories de flux dans un NGN :

- Les flux de contrôle (Control) : qui transportent les données liées à l'établissement et au contrôle des sessions ;

- Les flux de données (Media) : c'est-à-dire les données échangées entre les applications et les terminaux utilisateurs ;
- Les flux de gestion du réseau (Management).

Cependant un réseau NGN n'est pas uniquement un ensemble de spécifications techniques qui décrivent une architecture technique. C'est avant tout un changement vers un nouveau paradigme de communication orienté vers les services. Il offre des services multimédia en s'appuyant sur un réseau support mutualisé. Pour cela, plusieurs éléments sont essentiels et globalement partagés par tous [69]:

- Cœur de réseau unique et mutualise pour tous les types de réseaux d'accès et de services ;
- Architecture de cœur de réseau en 2 couches : Transport et Services;
- Evolution du transfert des données en le mode paquet ;
- Fonctions liées aux services sont indépendantes des technologies sous-jacentes liées au transport. Les fonctions sont bien définies et dialoguent entre elles par des interfaces normalisées ;
- Support de multiples technologies d'accès, de terminaux multiples et de la convergence des réseaux voix/données;
- Interfonctionnement avec des réseaux basés sur des technologies anciennes par l'intermédiaire d'interfaces ouvertes ;
- Notion de mobilité généralisée, l'accès aux services et cohérence des services quelque soit le lieu ou la technologie d'accès fixe ou mobile.

II.2.4 Caractéristiques des réseaux NGN

Les principales caractéristiques des réseaux NGN sont l'utilisation d'un unique réseau de transport en mode paquet, ainsi que la séparation des couches de transport des flux et de contrôle des communications, qui étaient implémentées dans un même équipement pour un commutateur traditionnel. Ces grands principes se déclinent techniquement comme suit, concernant les équipements actifs du cœur de réseau NGN :

- Remplacement des commutateurs traditionnels par deux types d'équipements distincts :
    1. Des serveurs de contrôle d'appel dits Softswitch ou Media Gateway Controller ;

2. Des équipements de médiation et de routage dits Media Gateway, qui s'appuient sur le réseau de transport mutualisé NGN ;

- Apparition des nouveaux protocoles de contrôle d'appel et de signalisation;

II.2.5 Les entités fonctionnelles du cœur de réseau NGN

II.2.5.1 Le Media Gateway (MG)

Le Media Gateway est située au niveau du transport des flux média entre le réseau RTC (Réseau Téléphonique Commuté) et le réseau en mode paquet, ou entre le cœur de réseau NGN et les réseaux. Il a pour rôle:

- Le codage et la mise en paquets du flux média reçu du RTC et vice-versa (conversion du trafic TDM (Time Division Multiplexing) en trafic IP ;
- La transmission, selon les instructions du Media Gateway Controller, des flux média reçus de part et d'autre.

II.2.5.2 Le serveur d'appel ou Media Gateway Controller (MGC)

Dans un réseau NGN, le MGC possède de « l'intelligence » et c'est lui qui gère :

- L'échange des messages de signalisation transmise de part et d'autre avec les passerelles de signalisation, et l'interprétation de cette signalisation ;
- Le traitement des appels et dialogue avec les terminaux H.323, SIP, communication avec les serveurs d'application pour la fourniture des services ;
- Le choix du MG de sortie selon l'adresse du destinataire, le type d'appel, la charge du réseau, etc.
- La réservation des ressources dans le MG et le contrôle des connexions internes au MG (commande des Media Gateways).

II.2.5.3 Le Signalling Gateway (SG)

La fonction Signalling Gateway a pour rôle de convertir la signalisation échangée entre le réseau NGN et le réseau externe interconnecté selon un format compréhensible par les équipements chargés de la traiter, mais sans l'interpréter (ce rôle étant dévolu au Media Gateway Controller). Elle assure notamment l'adaptation de la signalisation par rapport au protocole de transport utilisé. Cette fonction est souvent implémentée physiquement dans le même équipement que la Media Gateway, d'où le fait que ce dernier terme est parfois employé abusivement pour recouvrir les deux fonctions MG et SG.

II.2.6 Les protocoles NGN

La convergence des réseaux voix/données ainsi que le fait d'utiliser un réseau en mode paquet pour transporter des flux multimédia, ayant des contraintes de temps réel, a nécessité l'adaptation de la couche Contrôle. En effet, ces réseaux en mode paquet étaient généralement utilisés comme réseau de transport mais n'offraient pas de services permettant la gestion des appels et des communications multimédia. Cette évolution a conduit à l'apparition de nouveaux protocoles, principalement concernant la gestion des flux multimédia, au sein de la couche Contrôle. On peut classer les protocoles de contrôle en différents groupes :

- Les protocoles de contrôle d'appel permettant l'établissement, généralement à l'initiative d'un utilisateur, d'une communication entre deux terminaux ou entre un terminal et un serveur ; les deux principaux protocoles sont H.323 et SIP;
- Les protocoles de commande de Media Gateway qui sont issus de la séparation entre les couches Transport et Contrôle, permettent au Softswitch ou Media Gateway Controller de gérer les passerelles de transport ou Media Gateway. MGCP (Media Gateway Control Protocol) de l'IETF et H.248/Megaco, développé conjointement par l'UIT et l'IETF, sont actuellement les protocoles prédominants ;
- Les protocoles de signalisation entre les serveurs de contrôle (ou Media Gateway Controller) permettant la gestion du plan contrôle :

II.2.7 Les services offerts par les NGN

Les NGN offrent les capacités, en termes d'infrastructure, de protocole et de gestion, de créer et de déployer des nouveaux services multimédia sur des réseaux en mode paquet. La grande diversité des services est due aux multiples possibilités offertes par les réseaux NGN en termes de :

- Support multimédia (données, texte, audio, visuel);
- Mode de communication, Unicast (communication point à point), Multicast (communication point-multipoint), Broadcast (diffusion) ;
- Mobilité (services disponibles partout et tout le temps) ;
- Portabilité sur les différents terminaux.

Parmi ces services offerts on peut citer :

- La messagerie instantanée ;

- La messagerie unifiée ;
- La diffusion de contenus multimédia ;
- La voix sur IP (VoIP).

En Algérie le nouveau réseau multiservices d'Algérie Télécom (RMS) de nouvelle génération (NGN) présente de nouvelles opportunités de communication et aussi de nouveaux services très performant à large bande, Le RMS d'Algérie Télécom est équipé d'une plate forme SIP, permettant d'offrir et de contrôler différents services VoIP.

II.3 Les Réseaux Voix sur IP (VoIP)

La VoIP (Voice over Internet Protocol), ou Voix sur IP est une technique permettant de transporter la voix sur un réseau IP. Le principe de cette technique consiste à encapsuler un signal audio numérisé dans le protocole IP pour le transporter sur un réseau. L'approche VoIP s'applique donc au transport de la voix sur Internet, que le réseau soit Internet, Intranet ou Extranet. Les premières applications de la transmission de la voix par le protocole IP (avec IBM en 1996) étaient caractérisées par une qualité de voix très mauvaise: retards importants souvent supérieurs à une seconde, échos, paroles saccadées. Vu l'évolution profonde du secteur de télécommunication et l'introduction du concept NGN, la voix sur IP est considérée comme un service directement lié à ce nouveau paradigme.

II.3.1 Principe de la transmission VoIP

Pour une communication en VoIP, le signal vocal doit être compressé et codé à l'aide d'un codec dédié à la vois sur IP, ensuite, l'information à transmettre est découpée en paquets par une procédure de paquétisation avant l'envoi sur le réseau IP. Les paquets d'informations, qui circulent sur Internet, empruntent des chemins différents et arrivent fréquemment dans le désordre. Les paquets sont alors stockés dans des mémoires tampons, ou buffer, pour être ré-séquencés et permettre la décompression de l'information et sa transformation en signal sonore.

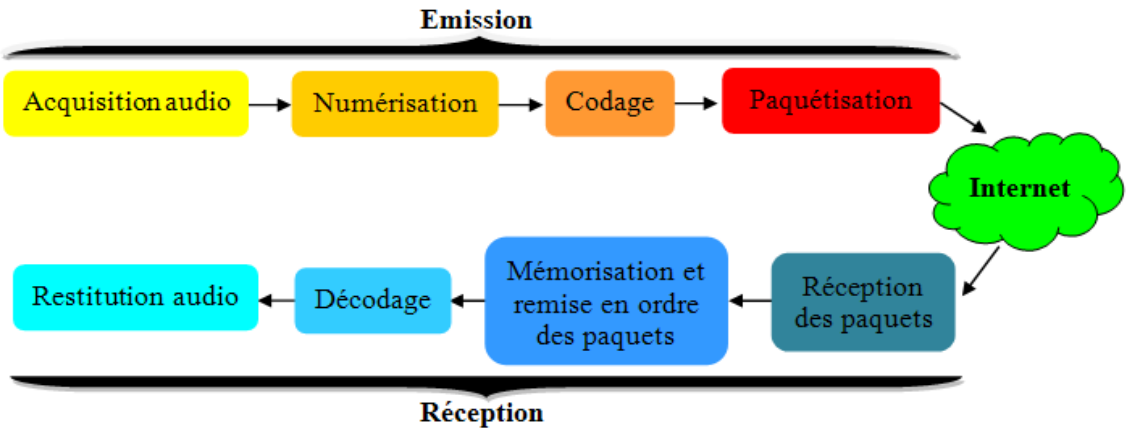


Figure II.3: Principe de la transmission de la voix par paquets.

La transmission de la vois sur un réseau IP est basée sur deux parties ; à savoir l'émission et la réception [70]:

- À l'émission, on trouve quatre étapes essentielles qui sont:
  1. La voix à transmettre est échantillonnée et numérisée par un convertisseur analogique numérique ;
  2. Le signal numérique obtenu est comprimé et encodé grâce à des algorithmes de compression spécifiques ;
  3. Le signal est découpé en paquets ;
  4. Les paquets sont transmis à travers le réseau Internet ;
- À la réception, les paquets sont réassemblés, le signal de données ainsi obtenu est décomprimé puis convertis en signal analogique sonore.

Le principal intérêt de la transmission de la VoIP est l'exploitation de la bande passante disponible par tous les utilisateurs du réseau [70]. Le transport de la voix sur IP peut prendre des formes distinctes selon l'appareil utilise à émission et la réception.

- La communication entre deux ordinateurs (PC to PC) entraîne une procédure lourde et peu conviviale. Elle suppose de nombreux accords sur l'heure de l'appel et les applications utilisées (même logiciel de téléphonie par exemple). De plus, il faut avoir pris auparavant connaissance de l'adresse IP du poste du correspondant.

- La communication entre un ordinateur et un poste téléphonique (PC to phone), elle désigne une utilisation plus large. Cette utilisation a pour but de se rapprocher de la téléphonie classique, tout en tirant parti de l'Internet.
- La Communication entre postes téléphoniques (phone to phone), ce type de communication reste encore relativement coûteux car il implique l'installation d'un nombre suffisant de passerelles sur les réseaux téléphoniques locaux.

II.3.2 Codecs dédie à la VoIP

Les codecs vocaux 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 super 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 ou la qualité d'expérience. Le choix du codec est un compromis entre la qualité de vocale et la capacité de l'infrastructure IP à délivrer une bande passante et des paramètres de QoS qui vont impacter cette qualité. En général, plus le débit binaire de la parole est grand, plus la qualité de la parole est bonne et plus l'application est gourmande en bande passante et en stockage.

II.3.2.1 ITU G.711

G.711 [71] est un codec qui a été mis en place par l'ITU en 1972 pour la téléphonie numérique. Le codec a deux variantes : A-Law est utilisé en Europe et lors des communications internationales, μ-Law est utilisé dans les États-Unis d'Amérique et le Japon. G.711 utilise une quantification à 8 bits et une compression logarithmique. Le débit résultant, pour une seule direction, est de 64 kbit/s, donc un appel consomme 128 kbit/s. Ce codec peut être librement utilisé (open-source) dans des applications Voix sur IP. Les meilleures performances de ce codec sont obtenues dans les réseaux locaux où nous avons beaucoup de bande passante disponible. De plus, ce codec est caractérisé par une très bonne qualité audio perçue (un MOS de 4,2 sur 5) et par une simple implémentation, et donc il n'a pas besoin d'un processeur puissant.

II.3.2.2 ITU G.729 (voir annexe A)

Le standard G.729 [72] décrit un algorithme pour le codage de signaux vocaux à 8 kbit/s au moyen de la prédiction linéaire à excitation par séquences codées à structure algébrique conjuguée (CS-ACELP) (Conjugate-Structure Algebraic-Code-Excited Linear-Prediction). Les annexes A, B et D à J du standard G.729 étendent les fonctionnalités du codec tel que le codage à un taux de 6,4 kbit/s et 11,8 kbit/s, une multi-cadence de fonctionnement, DTX et fournit une version à complexité réduite de l'algorithme. Le codec G.729 est généralement utilisé dans les applications VoIP.

II.3.2.3 ITU G.723.1

Le codec G.723.1 [73] est aussi généralement utilisé dans les applications VoIP. Le codeur est fondé sur les principes du codage prédictif linéaire (Linear Predictive Coding LPC) par analyse et synthèse, en vue de minimiser un signal d'erreur pondéré par une courbe de perception. Le G.723.1 possède deux débits binaires associés : 5,3 kbit/s et 6,3 kbit/s. Pour le débit supérieur, on fait appel à l'excitation par quantification d'impulsions multiples selon le critère du maximum de vraisemblance (MP-MLQ, Multi-Pulse Maximum Likelihood Quantization). Pour le débit inférieur, on fait appel à l'excitation par séquences codées à structure algébrique (ACELP, Algebraic-Code-Excitation).

II.3.2.4 GSM-FR

Le codec GSM 06.10 Full-Rate [74] décrit le transcodage dans la télécommunication cellulaire. Le schéma de codage est basé sur Regular Pulse Excitation – Long Term Prediction (RPE-LTP) paradigme de codage de la parole.

II.3.2.5 GSM-HR

GSM 06.20 Half Rate (HR) [75] nécessite moins de la moitié de la bande passante du GSM-FR au prix d'une moins bonne qualité audio. Le codec utilise l’algorithme Vector-Sum Excited Linear Prediction (VSELP).

II.3.2.6 AMR

Le codec audio Adaptive Multi-Rate (AMR) [76] est largement utilisé dans les réseaux cellulaires GSM et UMTS. Le codec encode le signal à huit différents débits, de l'ordre de 4,75 kbit/s à 12,2 kbit/s. Le débit le plus élevé de 12,2 kb/s est compatible avec le standard

GSM Enhanced Full Rate (GSM-EFR). Le système de codage est basé sur l'algorithme Algebraic Code Excited Linear Prediction (ACELP).

II.3.2.7 iLBC

Internet Low Bitrate Codec (iLBC) [77] est un codec conçu pour la communication voix sur IP. L'algorithme utilise Block-Independent Linear Predictive Coding (BI-LPC) et prend en charge des débits binaires de 13,3 et 15,2 kbit/s. Généralement les codecs à bas débit exploitent les dépendances entre les trames. Le traitement indépendant des trames appliquées par iLBC offre une meilleure robustesse du codec, similaire à celle du codec G.711 avec le masquage de pertes de paquets (Packet Loss Concealment PLC).

II.3.2.8 Speex

Speex [78] est un codec libre (Open source) ciblée pour la VoIP. Le codec utilise CELP comme technique de codage. Speex supporte un large intervalle de débits : de 3,95 à 24,6 kbit/s pour les signaux à bande étroite (narrow band). L'encodage est contrôlé par le paramètre de qualité qui varie de 0 à 10. Le mode de plus faible qualité 0 (correspondant à un débit de 2,15 kbit/s) est principalement utilisé pour le bruit de confort (confort noise).

II.3.2.9 Silk

Le codec Silk [79] est utilisé par l'application Skype. Le débit binaire pour les signaux à bande étroite (narrowband) peut être réglé entre 6 et 20 kbit/s. Le codec fournit également la transmission discontinue DTX (discontinuous transmission), un générateur de bruit de confort CNG (Comfort Noise Generator) et les mécanismes de masquage de pertes de paquets (Packet Loss Concealment PLC).

II.3.3 Protocoles dédié à la VoIP

Actuellement, il existe plusieurs approches de normalisation, pour fixer une qualité de service de téléphonie sur IP et permettre l'interopérabilité des applications et des équipements. La recommandation H.323 spécifiée par l'ITU-T en 1996, le Session Initiation Protocol (SIP) de l'IETF (Internet Engineering Task Force) ou encore le protocole GLP (Gateway Location Protocol). Les deux protocoles les plus utilisés et présents sur le marché sont le H.323 et le SIP.

II.3.3.1 Recommandation H.323

La recommandation H323 est le premier ensemble de normes réglementant la ToIP [80]. Elle est basé sur une architecture complète de communication pour transporter la voix sur différents type de réseaux et précise le rôle des divers éléments d'un système H.323 qui sont [70]:

- Gatkeeper : le composant le plus important d'un réseau H.323. Il gère les communications des terminaux rattachés à sa zone. Il fournit les services aux éléments qui sont enregistrés auprès de lui tel que la conversion des adresses, le contrôle d'admission, la gestion de la bande passante et la capacité de routage ;
- La MCM (multipoint control unit) : équipement optionnel de gestion des conférences. Elle contrôle l'entrée et la sortie des participants, minimise le trafic en diffusant les flux entre chaque émetteur et ses récepteurs. Elle se charge aussi de la conversion de codec vidéo pour offrir un affichage correct à un participant doté d'un écran de qualité médiocre ;
- la passerelle (Media Gateway) Elle assure l'interface avec des terminaux (H.323 ou autre), s'occupe de la conversion de signalisation de la conversion des codecs audio et vidéo et de l'adaptation aux supports de transmission. Elle est constituée de deux sous-entités :
    1. La partie traitement ou Media Gateway est la passerelle multimédia proprement dit ; l'entité de contrôle ;
    2. Le Media Gateway Controller, centralise les signalisations de contrôle et commande des différentes Media Gateways nécessaires à la traversée des réseaux rencontrés dans l'acheminement des communications.
- Terminal : S'agit d'un équipement utilisateur tel qu'un PC ou un téléphone IP qui supporte au moins un codec audio et éventuellement d'autres codecs audio et vidéo.

La figure II.4 représente les divers éléments d'une architecture H.323 que nous allons détailler par la suite.

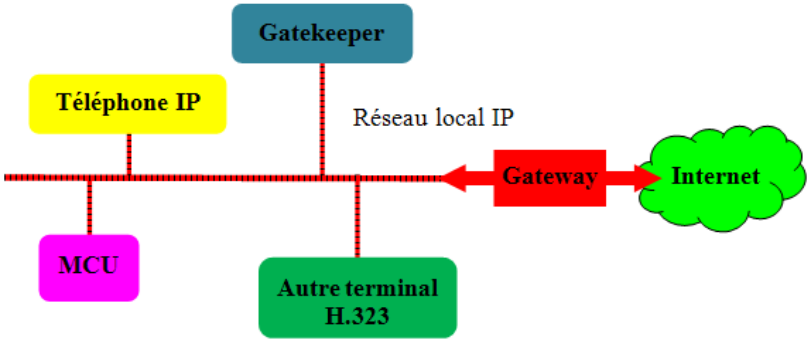


Figure II.4: Architecture d'un système conforme à H.323 [70].

II.3.3.1.1 Architecture H.323

Le protocole H.323 se dessine en 3 grandes parties (figure II.5) ; la signalisation, la négociation de codec, et le transport de l'information. En effet, pour établir une communication audio ou vidéo sur IP, le signal doit être encodé en utilisant des codecs normalisés définis dans la norme H323. Le protocole H323 normalise aussi la signalisation à utiliser pour l'établissement d'une communication. La voix ou la vidéo est transmise en utilisant le protocole UDP, associé aux protocoles RTP et RTCP pour le transfert des données en temps réel.

- Parmi les codecs possibles figurent G.711, G.723 .1 et G.729 pour les signaux audio, et H.261 et H.263 pour les signaux vidéo.
- La signalisation pour l'établissement des appels est mise en œuvre à l'aide de quatre protocoles :
    1. H.225 RAS (Registration, Admission and Status) : La signalisation RAS est utilisée pour la communication entre les terminaux et le Gatekeeper. RAS rend possible l'enregistrement, contrôle l'admission des appels et informe de l'état du système.
    2. H.225 Call signaling (Q.931) : Cette signalisation d'appel permet d'établir une connexion entre deux extrémités de communication. Les messages utilisés sont ceux du protocole de signalisation de la norme Q.931 modifiés par la recommandation H.225. La signalisation d'appel décrit la façon dont les signaux audio/video/données doivent être associés, codés et mis en paquets pour être acheminés sur un réseau local.

3. H.245 : signalisation de contrôle qui permet l'échange d'informations gérant les opérations de communication. Lorsque l'appelé décroche, le protocole H.245 permet l'établissement de canaux RTP/RTCP permettant le transfert de données multimédia et le contrôle de ce transfert.
4. T.120 : L'établissement et la gestion d'un transfert interactif de données entre un ou plusieurs terminaux.

- Les protocoles temps réel sur IP utilisés sont RTP et RTCP. RTP fournit un transport de bout en bout sur un réseau pour les applications transmettant des données en temps réel, telles que la voix ou la vidéo, en unicast et en multicast. RTP ne se préoccupe pas de la réservation de ressources et ne garantit pas la qualité de service des transferts de données en temps réel. Le transport des données bénéficie aussi du protocole de contrôle RTCP qui fournit un contrôle minimal et des fonctions d'identification particulièrement utiles dans le cas de réseaux multicast. RTP et RTCP sont conçus pour être indépendants des réseaux sous-jacents. Le protocole RAS utilise un transport UDP alors que les protocoles Call Signaling et H.245 s'appuient sur un transport TCP.

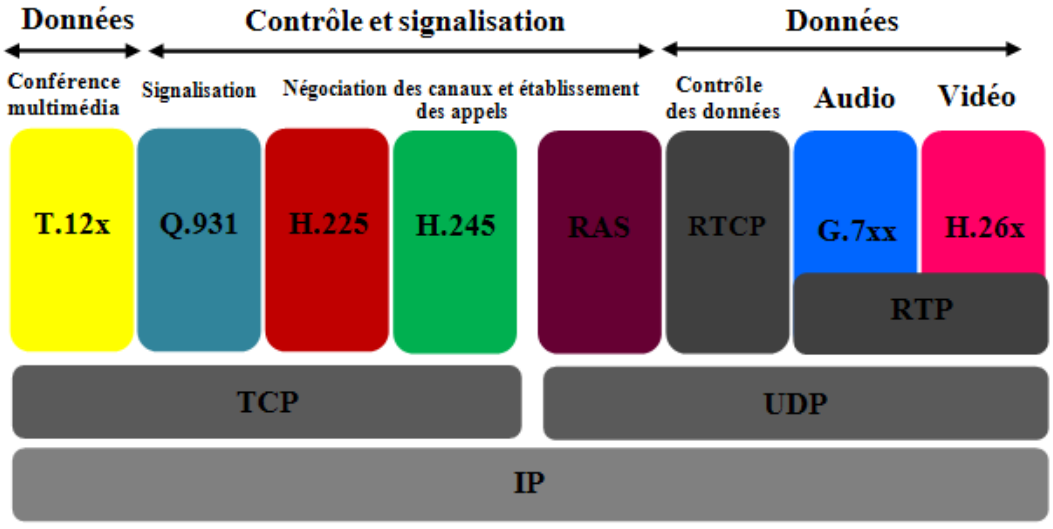


Figure II.5: Architecture du protocole H.323 [70].

II.3.3.2 SIP

Le protocole SIP (Session Initiation Protocol) est un protocole standardisé par l'IETF son rôle est d'ouvrir, modifier et libérer les sessions multimédia entre un ou plusieurs utilisateurs. L'ouverture de ces sessions permet de réaliser de l'audio ou vidéoconférence, de l'enseignement à distance, de la voix (téléphonie) et de la diffusion multimédia sur IP. Pour fonctionner, SIP a donc besoin d'autres standards ou protocoles, il est souvent décrit comme un protocole « chapeau » puisqu'il s'appuie sur d'autres briques protocolaires comme UDP [RFC768] ou TCP [RFC793] pour la couche Transport [81]. Notons que SIP possède l'avantage de ne pas être attaché à un médium particulier et est sensé être indépendant du protocole de transport. De plus, il peut être étendu et s'adapter aux évolutions futures. La pile protocolaire SIP pour la signalisation et le média est représentée par la figure II.6

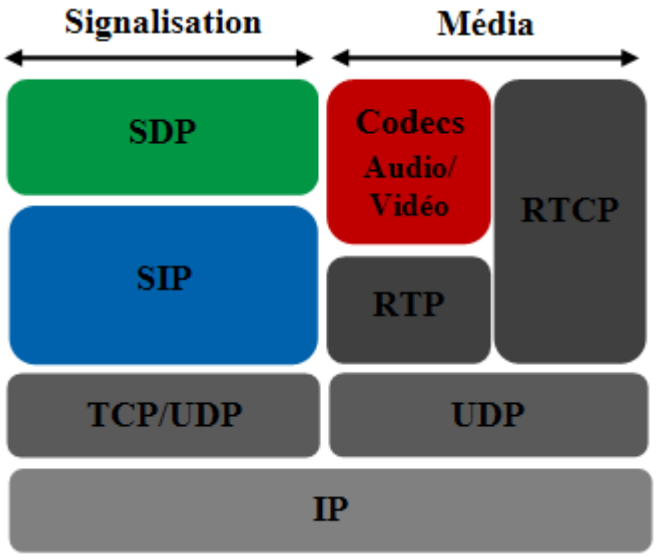


Figure II.6: Pile protocolaire de SIP [81].

II.3.3.2.1 Architecture SIP

Nous trouvons dans l'architecture fonctionnelle deux types de composants à savoir l'agent client (UA, User Agent) qui représente l'utilisateur et gère les requêtes, et les serveurs SIP qui reçoivent les requêtes et soutiennent l'établissement des échanges [70].

- L'agent client SIP : il constitue l'élément de base de l'architecture et se subdivise en deux parties : i) une partie client notée UAC (User Agent Client)

qui envoie les requêtes vers les serveurs SIP et traite les réponses, ii) une partie serveur notée UAS (user agent server) qui reçoit les requêtes et les traite. Le rôle d'un agent client comprend la gestion des sessions, la négociation des paramètres et l'envoi du média sous forme de messages RTP ou RTCP ainsi que son traitement par le biais de codecs.

- Les serveurs SIP : il en existe de 4 types :
  1. **Registrar Server** : il s'occupe exclusivement de l'enregistrement des terminaux SIP. Il reçoit les messages de type REGISTER. Il doit identifier les utilisateurs,voire les authentifier. Il doit être relié à un Proxy Server ou à un Redirect Server qui sera en charge de l'appel ;
  2. **Proxy Server** : il sert de relais aux messages SIP. Il joue le rôle de serveur d'un côté et de client de l'autre. Il interprète, transforme ou traduit un message avant de transférer ;
  3. **Redirect Server** : il gère la signalisation d'appel comme le Poxy Server, mais il ne relaie pas les messages. Il redirige directement l'UA vers la destination requise en lui indiquant l'adresse IP et le port à contacter ;
  4. **Location Server** : il est utilisé par les deux types de serveur précédents pour obtenir des informations sur les différentes localisations possibles d'un utilisateur.

II.3.4 Contraintes de la VoIP

II.3.4.1 Délai de transit (Delay)

Le délai de transit ou latence est le temps que va mettre en moyenne un paquet IP contenant un échantillon de voix pour traverser l'infrastructure entre deux interlocuteurs. Il influence fortement la QoS dans la VoIP et il comporte quatre composantes :

- Le délai de codage et décodage : C'est la durée de numérisation et de compression de la voix à l'émission puis de conversion en signal vocal à la réception. Ce temps dépend du type de codec choisi et varie de quelques millisecondes avec le codec G.711 (débit 64 kbps) à plus de 50 ms en G.723 (débit 6,3 ou 5,3 kbps).

- Le délai de propagation : C'est la durée de transmission en ligne des données numérisées. Cette durée est normalement très faible par rapport aux autres composantes du délai de transit, de l'ordre de quelques millisecondes.
- Le délai de transport : C'est la durée passée à traverser les routeurs, les commutateurs et les autres composants du réseau et de l'infrastructure de téléphonie IP.
- Le délai des buffers de gigue : C'est le retard introduit à la réception en vue de lisser la variation de temps de transit, et donc de réduire la gigue de phase. Les éléments d'infrastructure, notamment les routeurs, peuvent également mettre en œuvre des buffers de gigue.

II.3.4.2 Gigue (Jitter)

La variation de délai ou gigue est la conséquence du fait que tous les paquets contenant des échantillons de voix ne vont pas traverser le réseau à la même vitesse. Cela crée une déformation de la voix ou un hachage. La gigue est indépendante du délai de transit. Le délai peut être court et la gigue importante ou inversement. La gigue est une conséquence de congestions passagères sur le réseau, ce dernier ne pouvant plus transporter les données de manière constante dans le temps.

II.3.4.3 Perte de paquets

La perte de paquets est une source majeure de la distorsion de la parole dans la Voix sur IP. Les pertes de paquet peuvent être dues à plusieurs raisons, la congestion est la cause la plus fréquente de perte de paquet [82], [83]. Pour résumer, les causes majeures de pertes de paquets sont ; le désordre à l'arrivée des paquets (jitter) ; la congestion dans les nœuds du réseau (routeurs) ; les erreurs de transmission et le retard à l'arrivée du paquet.

II.3.4.4 Qualité de service (QoS) dans la VoIP

Il n'est pas facile de transformer un réseau d'échange de données en une architecture de transmission pour les applications critiques telle que la téléphonie. La qualité de service reste donc la question centrale de la voix sur IP.

Les principales contraintes de la transmission IP sont :

- Le délai ou latence : Dans le standard ITU G.114, des seuils de retard ont été définis pour garantir une qualité acceptable de la conversation En général, un

délai inférieur à 150 ms permet de tenir une conversation satisfaisante. Les études montrent également qu'un délai de 150 à 400 ms est tolérable pour des courtes durées d'une communication. Lorsque le retard dépasse 400 ms, une conversation téléphonique normale devient très difficile à tenir [84].

- La gigue : c'est la variation de délai, ce dernier pourrait être constant, ce qui préserve la synchronisation du signal entre l'émetteur et le récepteur, ou variable ce qui détruit la base de temps du signal et oblige le destinataire à maintenir une mémoire tampon de resynchronisation. Une gigue pouvant aller jusqu'à 50 ms est tolérée en VoIP.
- Les pertes de paquets : elles sont chroniques et font partie de la transmission IP. Elles sont nombreuses au moment de la congestion. Un taux de perte de paquets de l'ordre de 2% est acceptable.
- La qualité sonore : le phénomène d'écho devient gênant lorsque le temps d'aller retour du signal dépasse 40 ou 50 ms.
- A ces contraintes, on peut rajouter la contrainte de bande passante

II.4 Conclusion

Dans ce chapitre, nous avons présenté les réseaux NGN ainsi que la notion de la VoIP. La première partie de ce chapitre a pour but de définir les différents facteurs qui ont conduit à mettre en place un réseau de nouvelle génération (NGN), ainsi que son architecture et les différents protocoles et équipements qui le composent. La deuxième partie se concentre, sur la transmission de la voix sur IP, il traite les différents codecs et les protocoles de signalisation H.323 et SIP qui permet l'établissement, le maintien et la terminaison de sessions.

CHAPITRE III

La reconnaissance automatique du locuteur distribuée

- **II.1 Introduction**
- **II.2 Système distribué**
- **II.3 La reconnaissance automatique de locuteur distribuée sur IP (DSR)**
- **II.4 Architectures client-serveur**
- **II.5 Sockets**
- **II.6 Protocoles réseau et transport**
- **II.7 Conclusion**

III.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.

III.2 Système distribué

Il existe moult définitions d'un system distribué. Selon Tanenbaum [85]: « 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 [86]: « 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.

III.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 [87], publié par ETSI en Février 2000, suivie par DSR ES 202 050 [88] 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) [89], [90] pour gérer le processus de la reconnaissance automatique de locuteur comme illustré dans la Figure III.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.

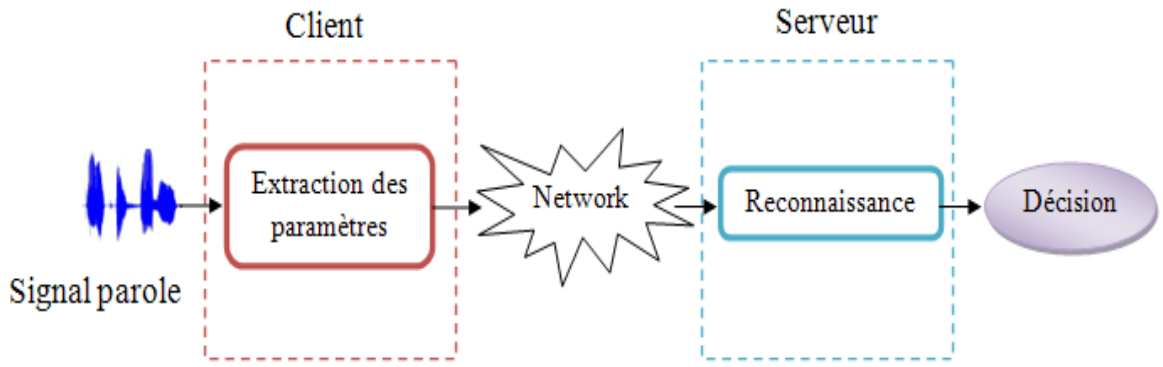


Figure III.1: Schéma de principe d'un système DSR basé sur la transmission des vecteurs des paramètres du côté client (front-end) au coté serveur (back-end) pour faire la reconnaissance.

- 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 III.2.

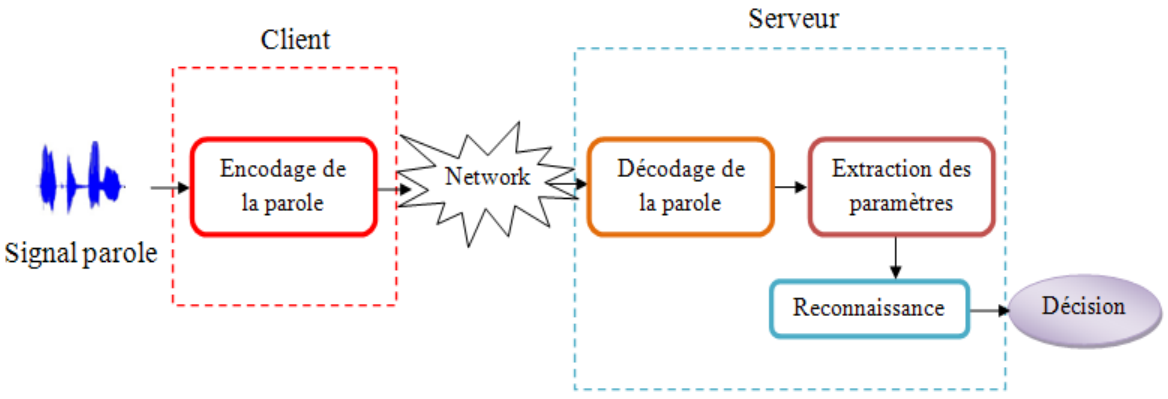


Figure III.2: Schéma de principe d'un système DSR basé sur la transmission de la parole codée (bit-stream) du coté client au coté serveur où il faut décoder le bit-stream et resynthétiser la parole. A partir de celle ci, l'extraction des paramètres est effectuée en vue de la reconnaissance.

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 [91]. Néanmoins, il a été constaté que les performances de reconnaissance peuvent être significativement dégradées lorsque la parole est reçue par ces réseaux [92]. 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.

III.4 Architecture client-serveur

Le modèle le plus simple d'application distribué est le modèle client-serveur (voir figure III.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 [93]. 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.

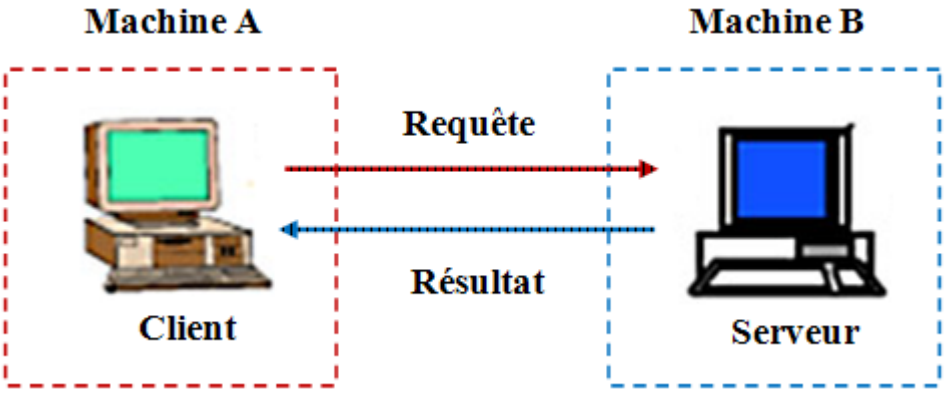


Figure III.3 : Interaction dans le modèle client-serveur.

III.5 Sockets

Les sockets sont une interface de programmation entre les applications et les couches réseau (figures III.4). Il s'agit d'une interface souple, d'assez bas niveau (couches 4 et 3 selon la norme modèle OSI). 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.

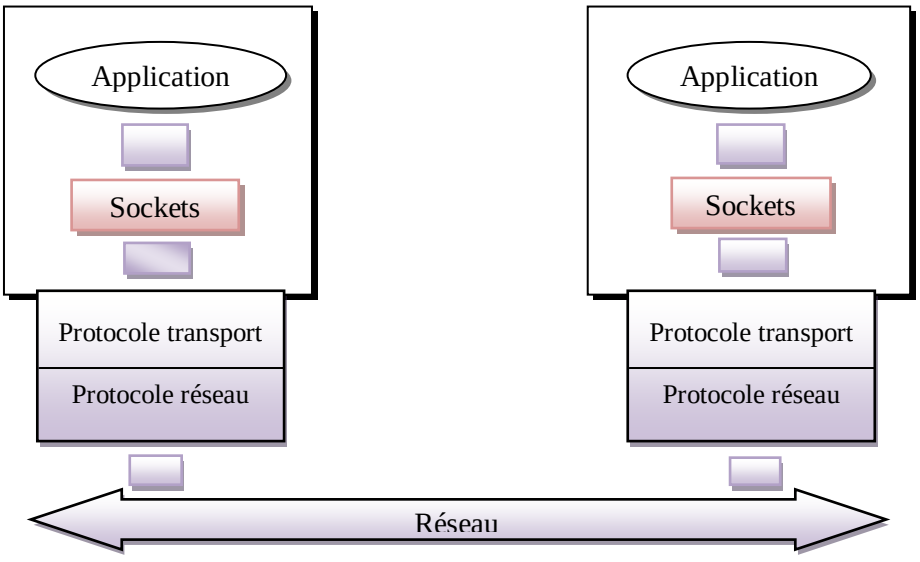


Figure III.4: Modèle de la communication via socket.

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 III.6). 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.

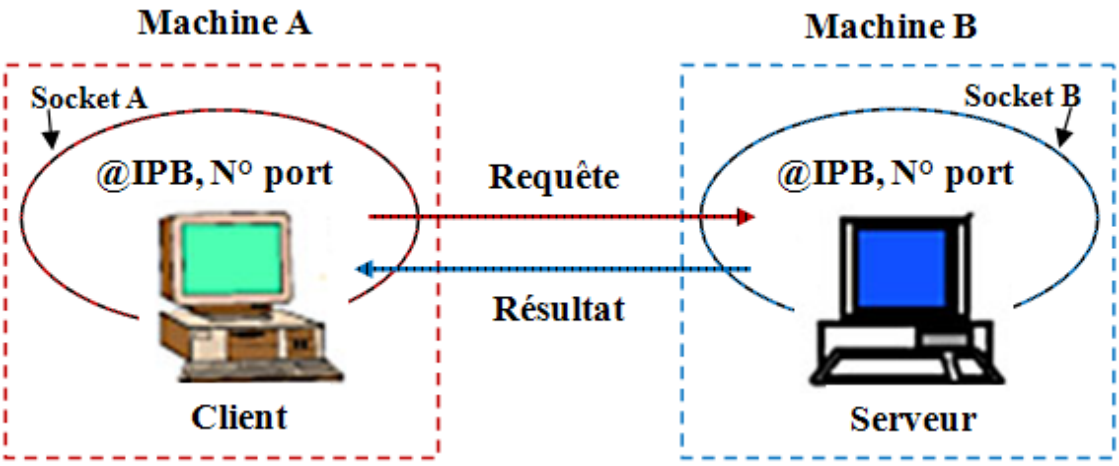


Figure III.5: Schéma d'un socket.

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 [94]. 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).

III.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.

III.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 [95]. Il correspond à la couche 3 (réseau) de la hiérarchie des couches ISO [96] [97]. 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 III.6 représente la structure de l'entête IP basé sur 20 octets.

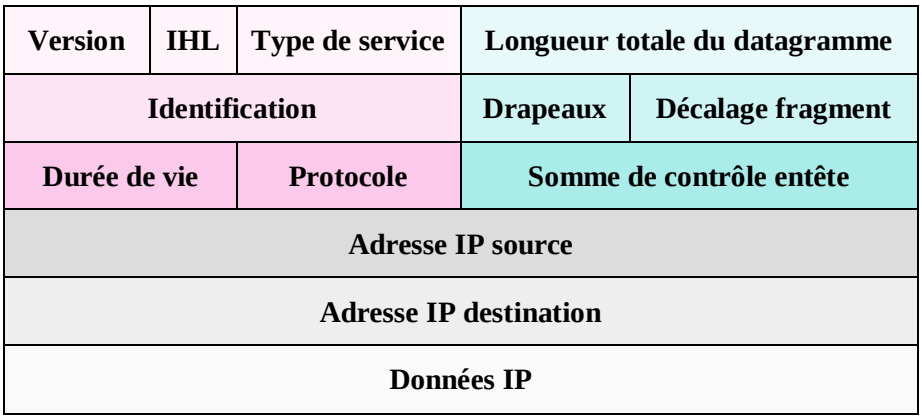


Figure III.6: Structure de l'entête IP basé sur 20 octets [97].

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 III.7, TCP/IP suit un modèle en couches légèrement différent du modèle OSI à 7 couches.

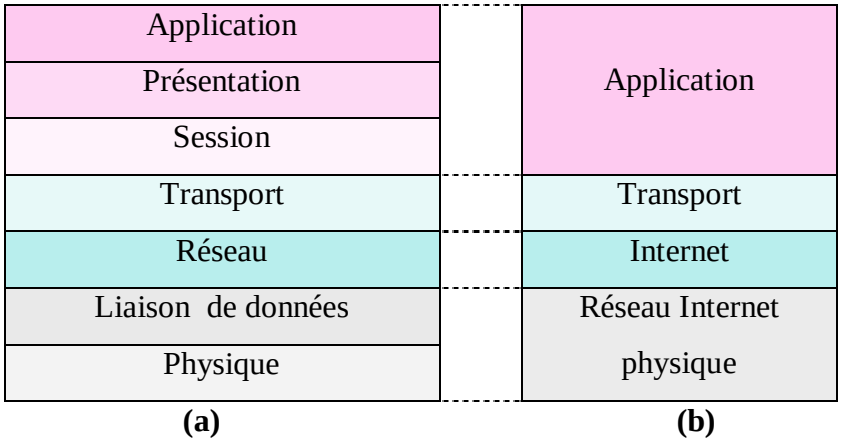


Figure III.7: (a) : Modèle de référence OSI, (b) : Modèle TCP/IP (Internet) [98].

III.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 III.8) et le transfère à la couche transport (TCP) où l'entête est analysé.

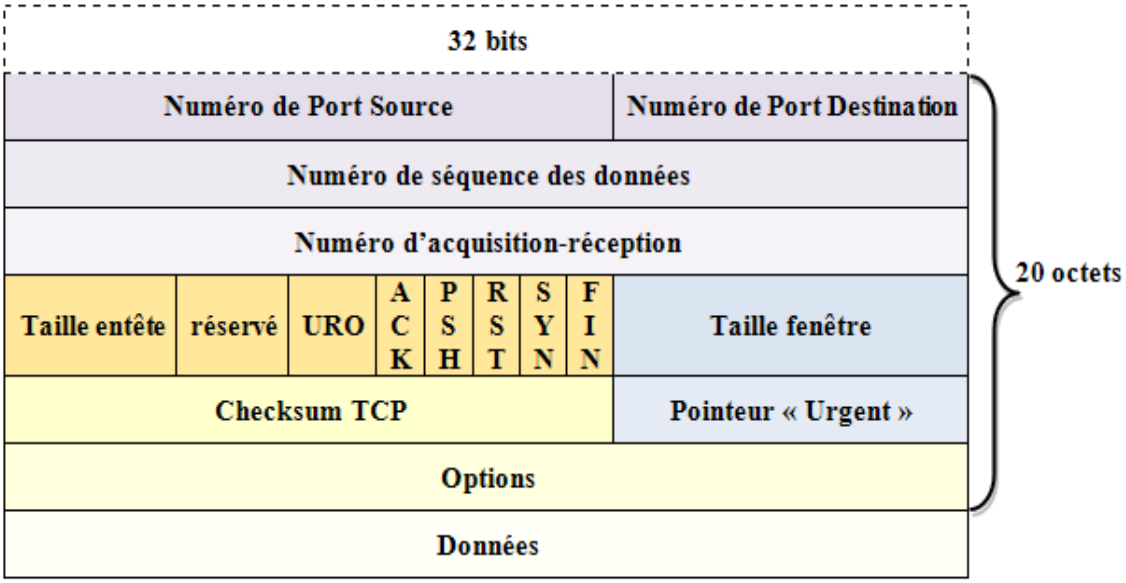


Figure III.8: Structure d'un segment TCP [97].

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 [95].

III.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 III.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.

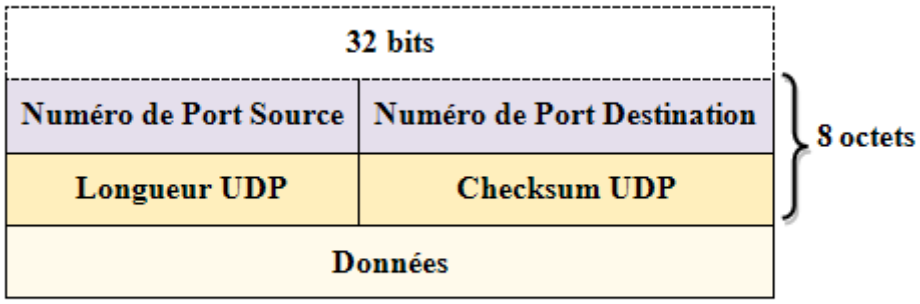


Figure III.9: Structure d'une trame UDP [97].

III.7 Conclusion

Dans ce chapitre, nous avons décrit l'architecture d'un système de reconnaissance distribuée (DSR). 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.

Cependant, dans tous les cas, le protocole UDP n'intègre pas de contrôle de congestion. C'est à l'application d'implémenter son propre mécanisme afin de respecter les autres flux présents sur le réseau.

CHAPITRE IV

Approche de reconnaissance proposée basée sur la fusion des modèles et des scores à base de GMM-UBM et GMM-SVM

Chapitre IV Approche de reconnaissance proposée basée sur la fusion des modèles et des scores à base de GMM-UBM et GMM-SVM

IV.1 Introduction

Ce chapitre est consacré aux techniques de modélisation en reconnaissances du locuteur et à la nouvelle approche proposée dans notre travail. Il présente tout d'abord l'approche de reconnaissance du locuteur basée sur le mélange de gaussiennes (GMM : Gaussian Mixture Model) [99] et l'utilisation d'un modèle générique appelé modèle du monde ou UBM (Universal Background Model), introduit par Reynolds [99], et Carey et Parris dans [100]. Cette approche est généralement nommée GMM-UBM [101]. L'objectif d'un tel paradigme est d'aboutir à une modélisation générative, l'estimation de la distribution qui a pu générer les vecteurs cepstraux du signal d'apprentissage.

La deuxième approche présente la méthode discriminante la plus employée en RAL, qui est basée sur l'utilisation des Machines à Vecteurs de Support (SVM: Support Vector Machine) [102]. A l'origine, les SVM ont été conçus comme une fonction discriminante permettant de séparer au mieux des régions complexes dans des problèmes de classification à 2 classes [103]. Ils démontrent aujourd'hui des performances similaires à l'approche GMM-UBM. Ces deux méthodes sont aussi combinées dans un nouveau formalisme, le GMM-SVM Super-Vecteur [104], [105] qui profite des capacités génératives du GMM et discriminantes du SVM.

L'approche de reconnaissance proposée dans notre travail, basée sur la fusion des modèles et des scores à base de GMM-UBM et GMM-SVM, est présentée en fin de chapitre.

IV.2 Première approche: Le paradigme GMM-UBM

Les approches génératives utilisées en reconnaissance du locuteur reposent essentiellement sur le paradigme GMM-UBM (figure IV.1). 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) [106], [107]. 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.

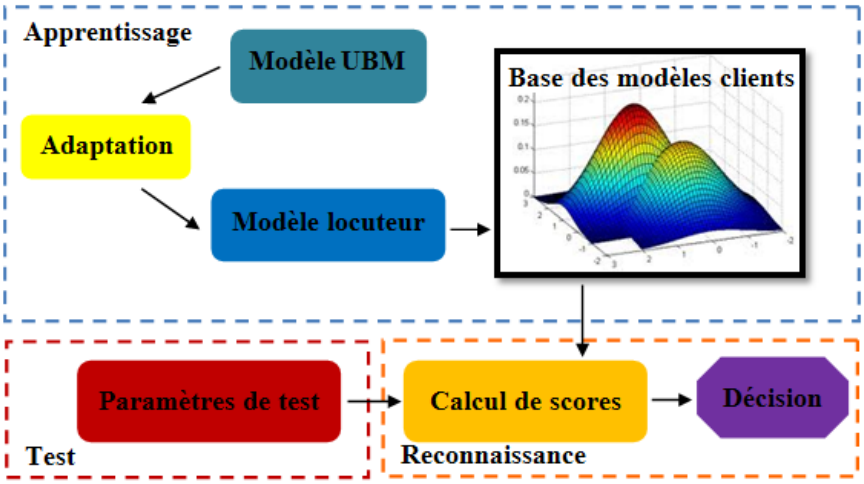


Figure IV.1: Structure du paradigme GMM-UBM en RAL.

IV.2.1 Modèle de mélange de gaussiennes

L'étape la plus importante dans l'implémentation d'un système d'identification, basé sur le rapport de vraisemblance, est le choix de la fonction de vraisemblance $p(X|\lambda)$. Le choix de cette fonction dépend du type de l'application et des paramètres utilisés. Pour l'identification du locuteur en mode dépendant du texte, où le système a beaucoup d'information a priori concernant le message prononcé, une information temporelle additionnelle peut être incorporée en employant le Modèle de Markov Caché (HMM). Dans les applications en mode indépendant du texte, où le système n'a aucune connaissance sur le message prononcé, (qui est le cadre de notre travail), la fonction la plus utilisée est le modèle de mélange de gaussiennes (GMM).

IV.2.1.1 Description du modèle

Le mélange de gaussiennes est une somme pondérée de plusieurs distributions gaussiennes, chacune a un vecteur de moyenne μ et une matrice de covariance Σ . Ce modèle est représenté par la relation suivante:

$$p(x|\lambda) = \sum_{i=1}^{M} p_i b_i(x) \qquad \text{(IV.1)}$$

Où x est un vecteur de dimension D, M est le nombre des composantes gaussiennes, p_i est le poids du mélange, qui vérifie la condition $\sum_{i=1}^{M} p_i = 1$ et $b_i(x)$ est une distribution gaussienne multidimensionnelle :

$$b_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{(1/2)}} exp\left\{-\frac{1}{2}\left(x-\mu_i\right)'\Sigma_i^{-1}(x-\mu_i)\right\} \quad \text{(IV.2)}$$

Avec μ_i est le vecteur de moyenne, Σ_i la matrice de covariance.

Le mélange de gaussiennes est caractérisé par le vecteur de moyenne μ_i , la matrice de covariance Σ_i et le poids de mélange p_i de toutes les densités gaussiennes. Ces paramètres sont regroupés dans un seul modèle λ , tel que :

$$\lambda = \{p_i, \mu_i, \Sigma_i\} \quad i = 1, \dots, M \quad \text{(IV.3)}$$

Pour la modélisation du locuteur, chaque individu est représenté par son vecteur de paramètres dans le modèle λ. La forme du modèle de mélange de gaussiennes (GMM) dépend du choix de la matrice de covariance : une matrice de covariance pour chaque composante gaussienne, une matrice de covariance pour chaque modèle ou une seule matrice de covariance pour tous les modèles.

IV.2.1.2 Interprétation du modèle

L'utilisation des GMM pour la modélisation du locuteur est motivée pour deux raisons. La composante gaussienne fut, en premier lieu, utilisée pour représenter les caractéristiques spectrales des formes phonétiques issues de la voix d'une personne. L'espace acoustique correspondant à la voix d'un locuteur peut être caractérisée par un ensemble de classes acoustiques représentant des événements phonétiques, voyelles et les consonnes. Ces classes acoustiques reflètent une certaine configuration des cordes vocales dépendante du locuteur qui sont utiles pour caractériser son identité. La forme spectrale de la $i^{ième}$ classe acoustique peut être représentée par le vecteur moyen μ_i de la $i^{ième}$ composante et les variations de la forme spectrale moyenne peuvent être représentées par la matrice de covariance Σ_i.

La deuxième motivation pour l'usage de mélange de gaussiennes pour la modélisation du locuteur est une observation empirique. Une combinaison linéaire des distributions gaussiennes est capable de représenter une grande classe des paramètres dans une simple distribution. Le modèle mono gaussien classique du locuteur représente la distribution des

vecteurs acoustiques d'un locuteur par une position (moyenne) et une forme elliptique (matrice de covariance) et le modèle de QV (quantification vectorielle) représente la distribution de ces vecteurs par un ensemble discret de poids de pondération caractéristiques. Dans un certain sens, le GMM peut être considéré comme une hybridation entre ces deux modèles en utilisant un ensemble discret de fonctions gaussiennes, chacune avec leur propre vecteur de moyenne et matrice de covariance, ce qui permet une meilleure modélisation pour le locuteur.

IV.2.1.3 Estimation des paramètres

L'objectif de l'apprentissage du modèle du locuteur est d'estimer, à partir des données extraites des segments de paroles, les paramètres du GMM qui donnent la meilleure distribution possible des vecteurs acoustiques. Plusieurs techniques sont disponibles pour estimer les paramètres d'un GMM. L'estimation par maximum de vraisemblance (MV) reste la plus populaire. Le vecteur λ_i de chaque locuteur est estimé de façon à maximiser la vraisemblance du GMM. En effet pour une séquence de T vecteurs d'apprentissage $X = \{x_1, \dots, x_T\}$ supposons que ces observations sont indépendantes, le maximum de vraisemblance du GMM est donné par :

$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \qquad \text{(IV.4)}$$

Malheureusement, la maximisation analytique de cette fonction n'est pas facile. L'algorithme EM (Expectation Maximization) [108] permet de résoudre ce problème. Cet algorithme maximise, de façon itérative, la fonction de vraisemblance. Cette maximisation fait intervenir la fonction auxiliaire $Q(\lambda, \lambda^n)$ qui est définie comme étant l'espérance mathématique du logarithme de vraisemblance [109].

$$Q(\lambda, \lambda^n) = \sum_{i=1}^{M} \sum_{t=1}^{T} p(i|x_t, \lambda^n)\, log\ p(x_t, i|\lambda) \qquad \text{(IV.5)}$$

Avec λ , désignant les paramètres à estimer, $(p_i, \mu_i\ et\ \Sigma_i)$ et λ^n ,l'ensemble des paramètres estimés à l'itération n. Ce qui donne après calcul:

$$Q(\lambda, \lambda^n) = \sum_{i=1}^{M} \sum_{t=1}^{T} p(i|x_t, \lambda^n) \left[log p_i - \frac{D}{2} \log(2\pi) - \frac{1}{2} log|\Sigma_i| \right] -$$

$$\sum_{i=1}^{M} \sum_{t=1}^{T} p(i|x_t, \lambda^n) \left[-\frac{1}{2} (x_t - \mu_i)' \Sigma_i^{-1} (x_t - \mu_i) \right] \qquad \text{(IV.6)}$$

Où $p(i|x_t, \lambda^n)$ est une probabilité a posteriori donnée par :

$$p(i|x_t, \lambda^n) = \frac{p_i^n b_i^n(x_t)}{\sum_{k=1}^{M} p_k^n b_k^n(x_t)} \quad \text{(IV.7)}$$

Supposant que $b_i(x_t)$ est une densité gaussienne à matrices de covariance diagonales, l'expression de la fonction auxiliaire devient :

$$Q(\lambda, \lambda^n) = \sum_{i=1}^{M} \sum_{t=1}^{T} p(i|x_t, \lambda^n)\, log p_i$$
$$-\frac{1}{2} \sum_{i=1}^{M} \sum_{t=1}^{T} p(i|x_t, \lambda^n) \left[Cste + log\sigma_i^2 - \frac{(x_t - \mu_i)^2}{\sigma_i^2} \right] \quad \text{(IV.8)}$$

Les paramètres sont estimes en annulant la dérivée de la fonction auxiliaire Q par rapport à chacun de ceux-ci. Le cas des poids des composantes de mélange p_i est assez simple puisqu'il s'agit de paramètres scalaires. Ceci dit, il faut tenir compte de la contrainte qui existe sur ces paramètres $\sum_{i=1}^{M} p_i = 1$. La maximisation sous contrainte se résout simplement en introduisant un multiplicateur de Lagrange associé à cette contrainte.

$$\frac{\partial}{\partial p_i}\left(Q(\lambda, \lambda^n) + \mathcal{L}(\sum_{i=1}^{M} p_i - 1) \right) = \frac{\sum_{t=1}^{T} P(i|x_t, \lambda^n)}{p_i} +$$
$$\mathcal{L} \sum_{i=1}^{M} \sum_{t=1}^{T} P(i|x_t, \lambda^n) + \sum_{i=1}^{M} p_i = 0 \quad \text{(IV.9)}$$

$$p_i = \frac{1}{T} \sum_{t=1}^{T} p(i|x_t, \lambda^n) \quad \text{(IV.10)}$$

Le vecteur de moyenne est donné par :

$$\frac{\partial}{\partial \mu_i}\left(Q(\lambda, \lambda^n) \right) = \frac{\sum_{t=1}^{T} P(i|x_t, \lambda^n)(x_t - \mu_i)}{\sigma_i^2} = 0 \quad \text{(IV.11)}$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda^n) x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda^n)} \quad \text{(IV.12)}$$

Et finalement, la matrice de covariance est donnée par :

$$\frac{\partial}{\partial \sigma_i^2}\left(Q(\lambda, \lambda^n) \right) = \frac{1}{2} \sum_{t=1}^{T} P(i|x_t, \lambda^n) \left(\frac{1}{\sigma_i^2} - \frac{(x_t - \mu_i)^2}{(\sigma_i^2)^2} \right) = 0 \quad \text{(IV.13)}$$

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda^n) {x_t}^2}{\sum_{t=1}^{T} p(i|x_t, \lambda^n)} - {\bar{\mu}_i}^2 \quad \text{(IV.14)}$$

IV.2.1.4 Difficultés algorithmiques

IV.2.1.4.1 Initialisation

La procédure d'apprentissage du GMM doit être initialisée avec un certain modèle initial $\lambda^{(0)}$, l'algorithme EM garantit de trouver un maximum local de la vraisemblance indépendamment du point de départ. Cependant, la vraisemblance d'un GMM a plusieurs maxima locaux, et différentes initialisations peuvent mener à des maxima locaux différents [110]. Pour étudier l'effet de l'initialisation du modèle sur les performances d'identification du locuteur, Reynolds et al [111] formèrent plusieurs modèles pour l'identification du locuteur utilisant différentes méthodes d'initialisation.

La première méthode d'initialisation emploie un HMM pour segmenter automatiquement le signal de parole d'apprentissage. Les données d'apprentissage ont été segmentées dans des classes phonétiques qui correspondent aux composants initiaux de mélange. Les vecteurs de moyennes et les variances globales sont utilisés comme un modèle initial pour l'algorithme EM. La deuxième méthode d'initialisation consiste à choisir aléatoirement des vecteurs à partir de l'ensemble des vecteurs acoustiques d'apprentissage pour les vecteurs de moyennes et une matrice identité pour la matrice de covariance.

IV.2.1.4.2 Limitation de variance

Lorsqu'on utilise un GMM avec une matrice de covariance pour chaque modèle, on observe que les éléments de la matrice de covariance peuvent devenir très petits. Cela apparaît particulièrement pour un GMM avec un nombre de composantes gaussiennes supérieur à 32. Ces petites variances produisent une singularité dans le modèle de la fonction de vraisemblance et peuvent modifier le modèle de calcul du score utilisé dans le classificateur de maximum de vraisemblance, ce qui dégrade les performances d'identification du locuteur. Ces singularités peuvent apparaître quand la quantité d'apprentissage est assez faible, ou quand les données d'apprentissage sont corrompues par le bruit [112].

Pour éviter ces fausses singularités, on applique la contrainte suivante ; dans toutes les matrices de covariances, on remplace les valeurs inferieures à une certaine valeur minimale par cette valeur minimale. Pour une composante gaussienne arbitraire, on applique après chaque itération EM la contrainte suivante :

$$\bar{\sigma}_i^2 = \begin{cases} \sigma_i^2 & si \quad \sigma_i^2 > \sigma_{min}^2 \\ \sigma_{min}^2 & si \quad \sigma_i^2 < \sigma_{min}^2 \end{cases} \tag{IV.15}$$

Cette contrainte sur l'algorithme EM fourni des estimations plus robustes du paramètre que la version sans contrainte [110], [113]. Pour optimiser les performances, le choix de cette valeur minimale doit être déterminé empiriquement pour chaque type de données, des paramètres acoustiques et de l'ordre du modèle. Dans ce travail la valeur minimale a été fixée à 0.01.

IV.2.1.4.3 Ordre du modèle

La détermination du nombre de composantes M dans un mélange requis pour modéliser un locuteur est un problème important et très difficile. Jusqu'à présent, il est impossible théoriquement d'estimer le nombre de composantes de mélange a priori. Le choix d'un nombre très réduit de composantes gaussiennes peut produire un modèle du locuteur qui ne modélise pas exactement la distribution de ses paramètres acoustiques. Le choix d'un nombre très grand de composantes gaussiennes peut aussi augmenter le temps d'exécution et produire une complexité accrue de calcul et de classification.

IV.2.2 Modèle non locuteur UBM

Le modèle générique, encore appelé modèle du monde ou UBM est la modélisation unique, pour tous les locuteurs, de l'hypothèse inverse dans le test bayésien. Il représente ainsi le modèle du non-locuteur pour tous les locuteurs, c'est en ce sens qu'il est qualifié d'« universel ». Dans le cadre de cette thèse l'UBM intervient aux différents niveaux dans le système de RAL. Il représente l'hypothèse inverse dans le test de vérification, sert d'initialisation pour l'estimation des modèles de locuteur, et il permet de structurer l'espace acoustique pour tous les modèles. Dans notre travail, la construction du modèle UBM est basée sur la collection de toutes les données d'apprentissage pour former un seul modèle UBM estimé par l'algorithme EM du toolkit ALIZE [114].

IV.2.3 Estimation du modèle locuteur

Pour palier le manque de données d'apprentissage pour réaliser une estimation robuste des GMM de locuteurs, l'adaptation MAP du modèle du monde est utilisée [100]. Cette adaptation permet d'estimer de manière robuste des modèles spécifiques au locuteur en ajoutant de l'information a priori sur la distribution des paramètres.

IV.2.3.1 Adaptation Maximum a Posteriori

L'adaptation Maximum a posteriori (MAP) [106] est utilisée en Vérification Automatique du Locuteur (VAL), pour adapter le modèle du monde aux données d'apprentissage des locuteurs. Seules les moyennes du GMM sont adaptées en VAL. L'estimation par le maximum a posteriori consiste à définir des distributions a priori $p(\lambda)$ pour les paramètres du modèle et à maximiser leurs probabilités a posteriori $p(\lambda|X)$ sur un signal d'apprentissage X. Le critère d'adaptation pour l'estimation des nouveaux paramètres s'écrit comme suit :

$$\hat{\lambda} = arg \underbrace{max}_{\lambda} p(\lambda|X) = arg \underbrace{max}_{\lambda} p(X|\lambda)p(\lambda) \quad \text{(IV.16)}$$

Une des formes les plus simples d'adaptation MAP, qui est la plus communément appliquée en reconnaissance du locuteur, consiste à adapter uniquement les vecteurs de moyenne des gaussiennes. Les moyennes du modèle du monde sont les a priori pour celles du locuteur [107]. Dans ce cas, La maximisation du paramètre de moyenne $\tilde{\mu}$ pour une Gaussienne i du GMM, s'exprime sous la forme :

$$\tilde{\mu}_i = \alpha_i {\mu_i}^c + (1 + \alpha_i)\mu_i^{\omega} \quad \text{(IV.17)}$$

Avec ${\mu_i}^c$ la $i_{ème}$ moyenne du GMM client et μ_i^{ω} la $i_{ème}$ moyenne du GMM du monde. α est un coefficient de pondération qui permet d'affecter plus ou moins de poids aux paramètres a priori par rapport aux paramètres estimés sur les données d'apprentissage. Il est défini par :

$$\alpha_i = \frac{\eta_i}{\eta_i + \tau} \quad \text{(IV.18)}$$

$$\eta_i = \sum_{t=1}^{T} Pr(i|x_t, \lambda) \quad \text{(IV.19)}$$

$$Pr(i|x_t, \lambda) = \frac{\omega_i p_i x_t}{\sum_{j=1}^{M} \omega_j p_j x_t} \quad \text{(IV.20)}$$

η_i est le nombre de trames associées à la $i_{ème}$ Gaussienne définie par l'équation IV.19 où $Pr(i|x_t, \lambda)$ est la probabilité que la $i_{ème}$ Gaussienne du GMM λ, ait généré le vecteur x_t.

τ est le relevance factor. Il contrôle le degré d'adaptation de chaque Gaussienne en terme de trames attribuées. Un relevance factor de 14 est couramment utilisé en VAL. 14 signifie qu'une confiance égale est accordée au modèle de référence et aux données d'apprentissage, lorsque 14 trames d'apprentissage sont associées à une composante du

GMM. En dessous de 14 trames, le facteur α devient très faible et l'influence des paramètres a priori est très forte. L'adaptation MAP permet ainsi de faire varier l'influence des données a priori, en fonction de la représentativité des données d'apprentissage pour chaque Gaussienne du modèle. Les paramètres du modèle du monde sont utilisés comme a priori. Cela permet d'initialiser les paramètres du modèle de locuteur de façon robuste.

IV.2.3.2 Test de vérification

Le test de vérification permet d'obtenir la mesure de similarité entre un modèle de locuteur et un signal de test. Cette mesure est appelée score de vérification, correspondant à la vraisemblance d'une séquence de données de test $X = [x_1 \cdots x_t]^T$ sur un modèle de locuteur *S*. Le score de vérification s'écrit sous la forme d'un rapport de vraisemblance (likelihood ratio) :

$$LR(X|S) = \frac{p(X|S)}{p(X|UBM)} \quad \text{(IV.21)}$$

En pratique, on utilise le logarithmique des vraisemblances, ce qui donne le logarithme du rapport de vraisemblance Log Likelihood Ratio (LLR), pour éviter les problèmes de précision numérique dûs aux multiplications de faibles valeurs. Le score de vérification utilisé en VAL est alors défini comme :

$$Score(X|S) = LLR(X|S) = \frac{1}{T}(\log(p(X|S)) - \log(p(X|UBM))) \quad \text{(IV.22)}$$

Où $p(X|S)$ et $p(X|UBM)$ sont les vraisemblances du vecteur X (le plus souvent cepstral) respectivement sur le modèle du locuteur S et sur le modèle du monde UBM.

IV.3 Deuxième approche: Machine à Vecteurs de Support (SVM)

La Machines à vecteurs de support (Support Vector Machine : SVM) est une technique discriminante relativement récente dans la théorie de l'apprentissage statistique. Elle sépare deux classes ayant comme labels +1et -1 par un hyperplan de séparation [115]. C'est une technique très utilisée en régression, classification et fusion.

Les SVM fournissent une approche très intéressante de l'approximation statistique. Souvent, le nombre d'exemples pour l'apprentissage est insuffisant pour que les estimateurs fournissent un modèle avec une bonne précision. D'un autre côté, l'acquisition d'un grand

nombre d'exemples s'avère être souvent très coûteuse et peut même mener à des problèmes de sur-apprentissage dans le cas où la capacité du modèle est très complexe. Pour ces deux raisons, il faut arriver à un compromis entre la taille des échantillons et la précision recherchée. Dans ces cas spécifiques, comme la reconnaissance de locuteur, il serait intéressant de trouver une mesure de la fiabilité de l'apprentissage, et du taux d'erreur qui sera commis durant la phase de test. Ces nouvelles techniques unifient deux théories : la minimisation du risque empirique et la capacité d'apprentissage d'une famille de fonctions.

Le principe des SVM consiste à projeter les données de l'espace d'entrée (appartenant à deux classes différentes) non-linéairement séparables dans un espace de plus grande dimension appelé espace de caractéristiques, de façon à ce que les données deviennent linéairement séparables. Dans cet espace, on construit un hyperplan optimal séparant les classes tel que :

- Les vecteurs appartenant aux différentes classes se trouvent de différents côtés de l'hyperplan.
- La plus petite distance entre les vecteurs et l'hyperplan (la marge), soit maximale.

IV.3.1 Construction de l'hyperplan optimal

Pour bien décrire la technique de construction de l'hyperplan optimal séparant des données appartenant à deux classes différentes dans deux cas différents : le cas des données linéairement séparables et le cas des données non-linéairement séparables, nous utiliserons les notations suivantes :

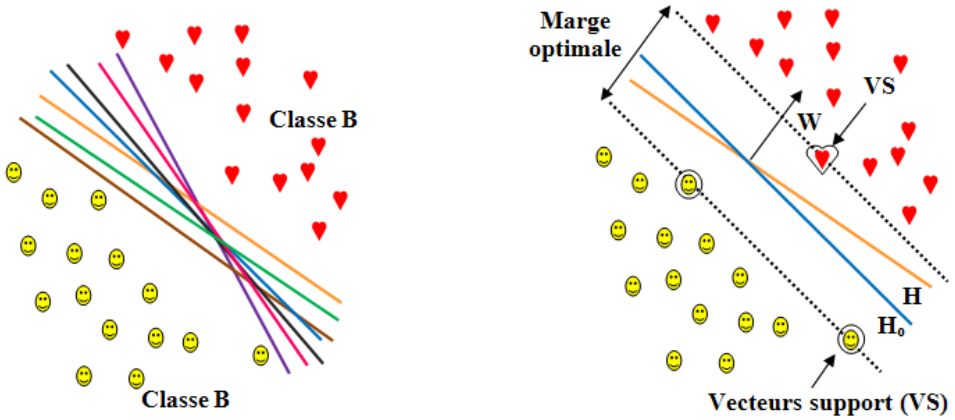
Considérons l points $\{(x_1, y_1),\ (x_2, y_2), \dots, (x_l, y_l)\}$ $x_i \in \Re^N$ avec $i = 1, \dots, l$ et $y_i \in \{\mp 1\}$. Classons ces points en utilisant une famille de fonctions linéaires définie par $\omega x + b = 0$ avec $\omega \in \Re^N$ et $b \in \Re$ de telle sorte que la fonction de décision concernant l'appartenance d'un point à l'une des deux classes soit donnée par :

$$f(x) = sgn(\omega x + b) \tag{IV.23}$$

IV.3.1.1 Cas de données linéairement séparables

Il existe une multitude d'hyperplans séparateurs qui séparent les données linéairement séparables de ces deux classes (figure IV.2 (a)). Mais seulement un seul de ces hyperplans

maximise la marge entre les données et la frontière de séparation. La marge étant la distance entre l'hyperplan et les données les plus proches appelées vecteurs supports. Le problème dans la modélisation par SVM est de trouver l'hyperplan optimal. Dans ce paragraphe, nous présentons la méthode générale de construction de l'Hyperplan Optimal (H_O) qui sépare des données appartenant à deux classes différentes linéairement séparables. La figure IV.2 (b) (où H_n : un hyperplan quelconque et VS sont les vecteurs supports**)** donne une représentation visuelle de l'H_O et de la marge optimale dans le cas des données linéairement séparables.



(a) Hyperplans séparateurs non optimaux **(b)** Hyperplan Optimal

Figure IV.2 : Données linéairement séparables.

Nous présentons brièvement le problème de la classification linéaire à 2 classes, reliées étroitement avec le formalisme des SVM. Un SVM peut être exprimé comme un classifieur bi-classe, les stratégies d'extensions multi-classes étant souvent exprimées comme des extensions du modèle binaire.

Considérons un jeu de données d'apprentissage $(x_t y_t)_{t=1\ldots T}$

$$H1:\ \omega x + b = 1 \qquad \text{(IV.24)}$$

$$H2:\ \omega x + b = -1 \qquad \text{(IV.25)}$$

Telle que les deux conditions suivantes soient respectées :

- Condition 1: Il n'y a aucun point qui se situe entre $H1$ et $H2$. Cette contrainte se traduit par les inégalités : $\omega x_i + b \geq +1$ pour $y_i = +1$ et $\omega x_i + b \leq -1$ pour $y_i = -1$. Ces deux inégalités peuvent être combinées en une seule: $y_i(\omega x_i + b) \geq +1$.
- Condition 2 : La distance ou la marge entre $H1$ et $H2$ est maximale.

Dans ce cas, la distance entre $H1$ et $H2$ est donnée par: $M = \frac{2}{|\omega|}$. Maximiser M revient à minimiser $|\omega|$ ou à minimiser $|\omega|^2$ avec $|\omega|^2 = \omega^T \omega$ (carré de la norme euclidienne du vecteur ω).

Le problème de séparation par hyperplan optimal peut être formulé comme suit :

$$\begin{cases} \underbrace{min}_{\omega,b} \frac{1}{2}\omega^T \omega \\ sous\ contraintes \\ y_i(\omega x_i + b) \geq +1 \quad i = 1, \dots, l \end{cases} \tag{IV.26}$$

Ce problème d'optimisation quadratique peut être résolu en introduisant des multiplicateurs de Lagrange $\propto_i > 0$. Le Lagrangien associé au problème précédent d'optimisation est :

$$L(\omega, b, \propto) = \frac{1}{2} \|\omega\|^2 - \sum_{i=1}^{l} \propto_i (y_i(\omega x_i + b) - 1) \tag{IV.27}$$

Le Lagrangien doit être minimisé par rapport à ω et b et maximisé par rapport à $\propto$.

$$\frac{\partial L}{\partial \omega} = 0 \tag{IV.28}$$

$$\frac{\partial L}{\partial b} = 0 \tag{IV.29}$$

avec $\propto_i \geq 0$. A partir des relations (IV.28) et (IV.29) nous pouvons déduire :

$$\omega = \sum_{i=1}^{l} \propto_i y_i x_i \tag{IV.30}$$

Et

$$\sum_{i=1}^{l} \propto_i y_i = 0 \tag{IV.31}$$

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

$$\begin{cases} L_{D=} \sum_{i=1}^{l} \propto_i - \frac{1}{2} \sum_{i,j=1}^{l} \propto_i \propto_j y_i y_j x_i x_j \\ sous\ contraintes \\ \sum_{i=1}^{l} \propto_i y_i = 0 \quad et \quad \propto_i \geq 0 \quad et\ i = 1, \dots, l \end{cases} \tag{IV.32}$$

La fonction de décision est alors :

$$f(x) = sgn(\sum_{i=1}^{l} \propto_i y_i (x_i x) + b) \tag{IV.33}$$

Cette fonction de décision est donc seulement influencée par les points correspondants à des $\propto_i$ non nuls. Ces points sont appelés les vecteurs de support. Ils correspondent, dans un cas linéairement séparable, aux points les plus proches de la limite de décision, c'est à dire aux points se trouvant exactement à une distance égale à la marge. Il s'agit là d'une propriété très intéressante des SVM : seuls les vecteurs de support sont nécessaires pour décrire cette limite de décision, et le nombre de vecteurs de support pour le modèle optimal est généralement petit devant le nombre de données d'entraînement.

IV.3.1.2 Cas des données non-linéairement séparables

La plupart des applications pratiques correspondent en réalité à des classes non linéairement séparables, où il n'existe aucune frontière de séparation linéaire capable de séparer les données des deux classes. Le problème est dit non linéairement séparable. Afin de traiter également des données bruitées ou non linéairement séparables (Figure IV.3), les SVM ont été généralisées grâce à deux outils : la marge souple (soft margin) et les fonctions noyau (kernel functions). Le principe de la marge souple est d'autoriser des erreurs de classification.

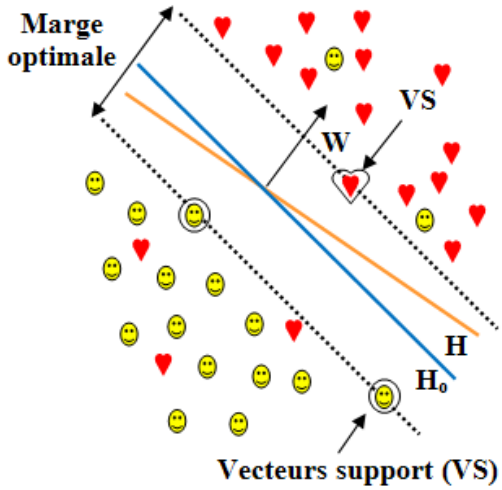


Figure IV.3 : Hyperplans séparateurs dans le cas de données non-linéairement séparables.

Le nouveau problème de séparation optimale est reformulé comme suit : L'hyperplan optimal séparant les deux classes est celui qui sépare les données avec le minimum d'erreurs, et satisfait donc les deux conditions suivantes :

- Condition 1: La distance entre les vecteurs bien classés et l'hyperplan doit être maximale.
- Condition 2: La distance entre les vecteurs mal classées et l'hyperplan doit être minimale.

Pour formaliser cela, on introduit des variables de pénalité non-négatives, ϵ_i pour $i = 1, \dots, l$ appelés variables d'écart. Le principe de la marge souple se traduit par la transformation des contraintes (V.29) qui deviennent :

$$y_i(\omega x_i + b) \geq +1 - \epsilon_i \; Pour \; i = 1, \dots, l \tag{IV.34}$$

Avec l'introduction d'un terme de pénalité, la fonction objective devient :

$$\underbrace{min}_{\omega, b, \epsilon} \frac{1}{2}\omega^T \omega + C \sum_{i=1}^{l} \epsilon_i \quad C \geq 0 \tag{IV.35}$$

Le paramètre C est défini par l'utilisateur. Il peut être interprété comme une tolérance au bruit du classificateur. C'est aussi la pénalité associée à toute violation des contraintes du cas linéairement séparable. Pour de grandes valeurs de C, seules de très faibles valeurs de ϵ sont autorisées et, par conséquent, le nombre de points mal classés sera très faible (données faiblement bruitées). Cependant, si C est petit, f peut devenir assez grand et on autorise alors bien plus d'erreurs de classification (données fortement bruitées). La nouvelle formulation du problème d'optimisation est alors :

$$\begin{cases} \underbrace{min}_{\omega, b, \epsilon} \frac{1}{2}\omega^T \omega + C \sum_{i=1}^{l} \epsilon_i \, , C \geq 0 \\ sous \; contraintes \\ y_i(\omega x_i + b) \geq +1 - \epsilon_i \\ \epsilon_i \geq 0 \quad Pour \; i = 1, \dots, l \end{cases} \tag{IV.36}$$

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

$$L(\omega, b, , \epsilon_i \propto, \mu) = \frac{1}{2}\omega^T \omega + C \sum_{i=1}^{l} \epsilon_i - \sum_{i=1}^{l} \propto_i [\, y_i(\omega^T x_i - b) + \epsilon_i - 1] - \sum_{i=1}^{l} \mu_i \, \epsilon_i$$

$$= \frac{1}{2}\omega^T \omega + \sum_{i=1}^{l} \big(C - \propto_i - \mu_i\big)\epsilon_i - \big(\sum_{i=1}^{l} y_i x_i \propto_i\big)\omega - \big(\sum_{i=1}^{l} y_i \propto_i\big) b + \sum_{i=1}^{l} \propto_i \tag{IV.37}$$

Le Lagrangien doit être minimisé par rapport à $\omega, b, , \epsilon_i$ et maximisé par rapport à $\propto$ et μ.

$$\frac{\partial L}{\partial \omega} = 0 \tag{IV.38}$$

$$\frac{\partial L}{\partial b} = 0 \quad \text{(IV.39)}$$

$$\frac{\partial L}{\partial \epsilon_i} = 0 \quad \text{(IV.40)}$$

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

$$\omega = \sum_{i=1}^{l} \propto_i y_i x_i \quad \text{(IV.41)}$$

$$\sum_{i=1}^{l} \propto_i y_i = 0 \quad \text{(IV.42)}$$

Et

$$\propto_i = C - \mu_i \quad \text{(IV.43)}$$

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

$$\begin{cases} L_D = \sum_{i=1}^{l} \propto_i - \frac{1}{2} \sum_{i,j=1}^{l} \propto_i \propto_j y_i y_j x_i x_j \\ \sum_{i=1}^{l} \propto_i y_i = 0 \quad et \quad \propto_i \geq 0 \\ \propto_i \geq 0 \quad et \quad i = 1, \dots, l \end{cases} \quad \text{(IV.44)}$$

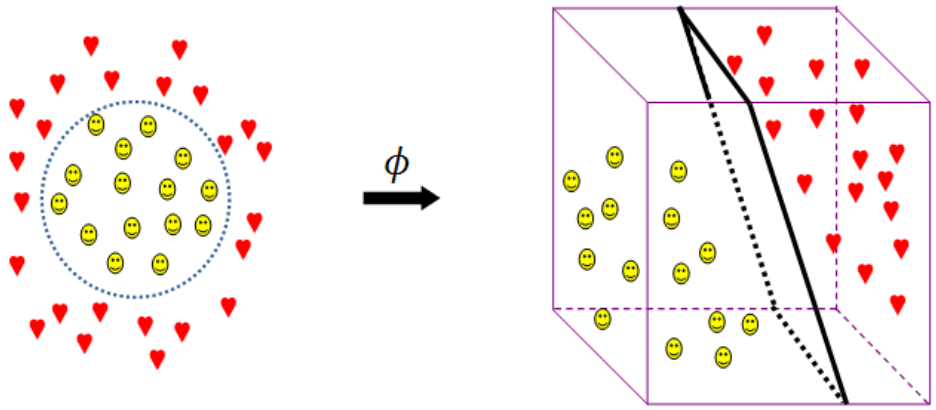
La seule différence avec le cas linéairement séparable est donc l'introduction d'une borne supérieure pour les paramètres $\propto_i$.

Choisir des frontières de décision linéaires semble être un facteur limitant. Les cas des données non linéairement séparables peuvent être considérablement enrichis en projetant les données non linéairement séparables dans un espace caractéristique $\mathcal{F}$ (feature space) de plus grande dimension permettant d'augmenter la séparabilité des données (voir figure IV.4). On peut alors appliquer le même algorithme dans ce nouvel espace, ce qui se traduit par une frontière de décision non linéaire dans l'espace initial. Considérons l'application non linéaire définie par :

$$\phi: X \to \mathcal{F} \quad \text{(IV.45)}$$

$$x \to \phi(x) \quad \text{(IV.46)}$$

Il suffit alors d'appliquer l'algorithme d'apprentissage dans $\mathcal{F}$ et non dans X en considérant l'ensemble $(\phi(x_i), y_i) \epsilon \mathcal{F} * y$ avec $i = 1 \cdots N$ et $y \epsilon \{1, -1\}$.



(a) Espace d'entrée X **(b)** Espace caractéristique $\mathcal{F}$

Figure IV.4: Principe des techniques SVM.

IV.3.2 Fonction Noyau et théorème de Mercer

Le principe des SVM est de transposer le problème dans un espace de dimension supérieure, pour traiter le cas de données non linéairement séparables, où il est probable de trouver un hyperplan comme une frontière de séparation linéaire. L'algorithme SVM requiert que les données soient représentées dans un espace doté d'un produit scalaire. Pour garantir cela, on introduit un projecteur ϕ de l'espace d'origine X dans un espace $\mathcal{F}$ (feature space) qui sera lui, doté d'un produit scalaire. Il suffit alors de remplacer chaque produit scalaire $\langle x_i, y_j \rangle$ par $\langle \phi(x_i), \phi(y_j) \rangle$. Cependant, pour certains espaces $\mathcal{F}$, il arrive que la projection ϕ ne soit pas calculable directement. Pour pallier à ce problème, on a recours à une famille de fonctions permettant de réaliser implicitement les produits scalaires dans $\mathcal{F}$. On appelle ces fonctions, fonctions noyaux (Kernel) et on les définit comme :

$$K(x, y) = \langle \phi(x), \phi(y) \rangle \qquad \text{(IV.47)}$$

Donc, chaque fonction noyau correspond à une projection ϕ et par conséquent, à un espace augmenté $\mathcal{F}$. En pratique, seule la fonction noyau intervient dans les calculs, alors il n'est pas nécessaire de connaitre la transformation ϕ. La fonction noyau ne fournit aucune information sur cette transformation et donc sur l'espace des caractéristiques. On peut donc envisager des transformations complexes, et même des espaces de redescription de dimension infinie. Le problème est de savoir quelles sont les fonctions noyau K qui admettent une telle

représentation, c'est-à-dire pour lesquelles on peut trouver un espace $\mathcal{F}$ et une projection ϕ. La solution est réalisée grâce à l'utilisation d'une fonction noyau (kernel function) respectant les conditions de Mercer [115]. Le théorème de Mercer est défini comme une fonction $K: X * X \rightarrow \mathbb{R}$. K est un noyau valide si elle est symétrique et définie positive. Sous cette condition, la fonction K est un noyau si est seulement si la matrice de Gram, appelée aussi matrice de similarité, définie par :

$$G\ (i,j) = K(x_i, x_j), \quad i,j = 1, \dots, N \tag{IV.48}$$

est définie positive.

Ainsi, la fonction $K: X * X \rightarrow \mathbb{R}$ générant une matrice définie positive possède les trois propriétés fondamentales du produit scalaire :

- Positive: $K(x_i, x_j) > 0$
- Symétrie: $K(x_i, x_j) = K(x_j, x_i)$
- Inégalité de Cauchy-Shwartz: $\left|K(x_i, x_j)\right| \leq \|x_i\| \cdot \left\|x_j\right\|$

En pratique, n'importe quelle fonction peut être utilisée, pour peu qu'elle représente un produit scalaire dans un certain espace, dans le but de former un noyau. Plus généralement un noyau valide doit satisfaire la condition de Mercer. Une liste de noyaux les plus courants peut être trouvée dans la littérature, les principaux étant : linéaire, polynomial et gaussien (tableau IV.1).

Noyau linéaire	$K(x_i, x_j) = {x_i}^T x_j$
Noyau gaussien	$K(x_i, x_j) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$
Noyau polynomial	$K(x_i, x_j) = ({x_i}^T x_j + 1)^d$

Tableau IV.1 : Quelques noyaux usuels.

IV.3.3 RAL avec SVM via les noyaux de séquences

En reconnaissance automatique du locuteur, le modèle SVM est appris en utilisant les données d'un locuteur comme client ayant comme label +1 et des données appartenant à

d'autres locuteurs comme imposteurs ayant le label −1. Les données utilisées peuvent être soit des séquences de vecteurs paramétriques, ou soit issues d'une modélisation des locuteurs par la combinaison des SVM avec d'autres techniques de modélisation. Des SVM élaborés sur des séquences de vecteurs paramétriques ont été utilisées en RAL :

- Noyau de toutes sortes avec différentes paramètres acoustiques : paramètres spectraux [102], [116], [117], [118], [119], [120], paramètres prosodiques [121] et paramètres de haut-niveau [122].
- Noyaux polynomiaux non normalisés et normalisés et un noyau radial gaussien (Radial Basis Function) RBF ont été exploités dans [102].
- Noyaux incorporant le principe d'alignement dynamique ont été proposés dans [123], [124] pour des applications dépendantes du texte.
- Noyau GLDS (Generalized Linear Discriminant Sequence) proposé dans [117], [120], ce noyau est basé sur une expansion polynomiale de vecteurs.
- Noyau linéaire, utilisé dans [121].
- Noyaux FSMS (Feature Space Mahalanobis Sequence) proposés comme une nouvelle famille de noyaux dans [125].

IV.3.4 Modèle hybride GMM-SVM appliqué à la RAL

Ces dernières années ont vu l'apparition d'une méthode simple et peu couteuse permettant de combiner les approches génératives et discriminantes, basées la combinaison des SVM et le paradigme GMM-UBM. Les premiers travaux de [126], [127], [128], [129] et [130] ont utilisé les scores de vraisemblances GMM comme de nouveaux paramètres exploités par les modèles SVM. Dans [131], des SVM ont été appris sur les différences entre les vecteurs de moyennes du modèle GMM du client et de ceux de l'UBM. D'autres travaux ont projeté les SVM dans un nouvel espace en utilisant des fonctions noyau fondées sur les scores de Fisher, un noyau de Fisher dans [132] et des noyaux bases sur la divergence de Kullback-Leibler (KL) dans [133],[134] et [135]. Aussi, une généralisation du noyau de Fisher a été proposée dans [136]. Puis, une toute autre approche basée sur les paramètres d'un modèle GMM comme vecteur d'observation du SVM a été proposé dans [137], [138]. Le supervecteur GMM est fondé sur les M vecteurs de moyennes de la modélisation GMM représenté par l'équation suivante:

$$p(x|\lambda) = \sum_{i=1}^{N} \gamma_i N(x; \mu_i, \Sigma_i) \quad \text{(IV.49)}$$

Où γ_i sont les poids de mélange, $N(.)$ est un gaussienne, μ_i et Σ_i sont la moyenne et la covariance des gaussiennes, respectivement. Le supervecteur de GMM est constitué des moyennes de tous les modèles comme illustré par figure IV.5. Le supervecteur GMM comprend donc les moyennes des modèles après l'adaptation.

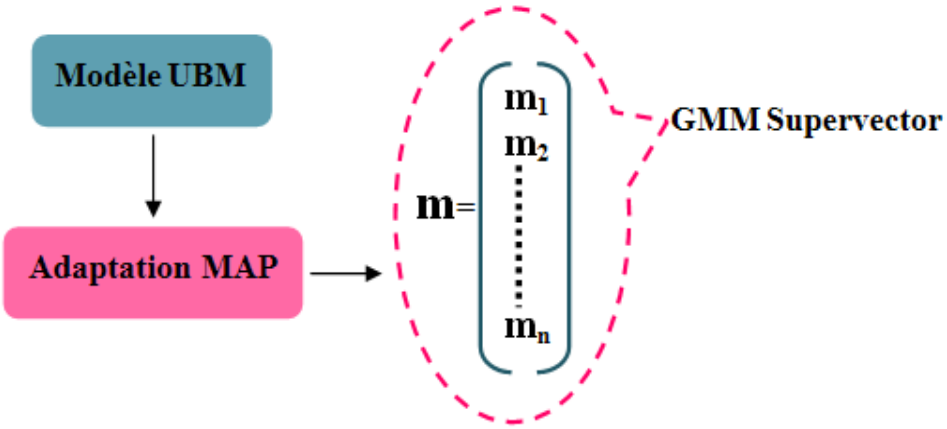


Figure IV.5: Le supervecteur GMM-SVM des moyennes GMM.

Les supervecteurs GMM s'accordent bien avec les SVM. Certains auteurs [137], [138] proposent donc d'utiliser un noyau linéaire dans l'espace des supervecteurs : les poids et les variances des gaussiennes servent à normaliser les vecteurs de moyennes avant l'apprentissage. Ce système GMM Supervector Linear Kernel (GSL) (appelé aussi dans la littérature Gaussian Supervector SVM (GSV-SVM) et GMM-SVM) est parmi les techniques les plus performantes en reconnaissance automatique du locuteur (figure IV.6).

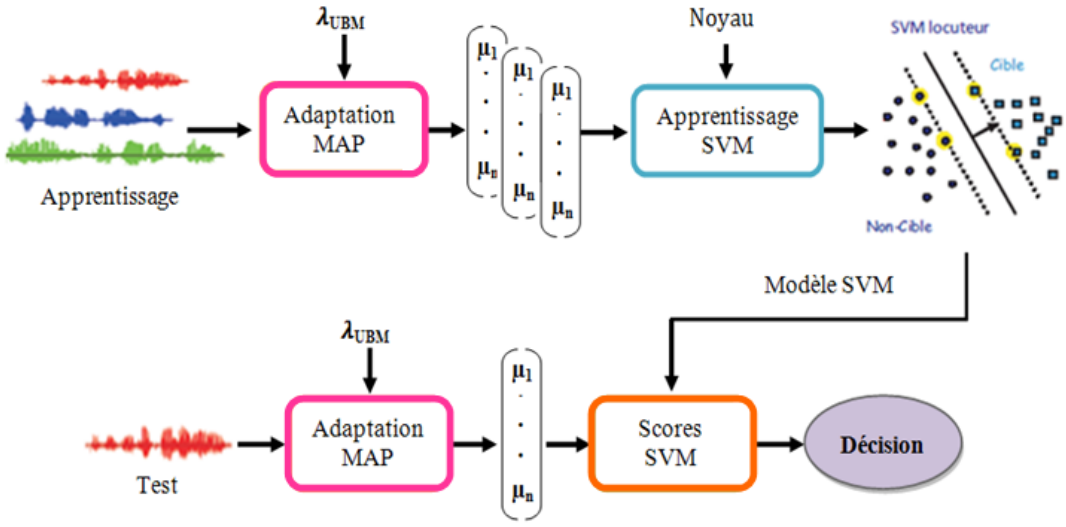


Figure IV.6 : Système RAL via le modèle hybride GMM-SVM.

Les moyennes des GMM de chaque locuteur d'apprentissage sont les vecteurs d'entrée exploités par le noyau d'apprentissage SVM, pour construire les modèles SVM qui seront enregistrés dans une base de modèles. Ensuite, il faut faire la comparaison de ces modèles avec noyau SVM de tests représentés par les moyennes de GMM des signaux tests. La combinaison des méthodes discriminantes et génératives est particulièrement intéressante. La figure IV.7 représente une expérience sur la base de données ARADIGIT [139], en utilisant le supervecteur GMM-SVM et le paradigme GMM-UBM.

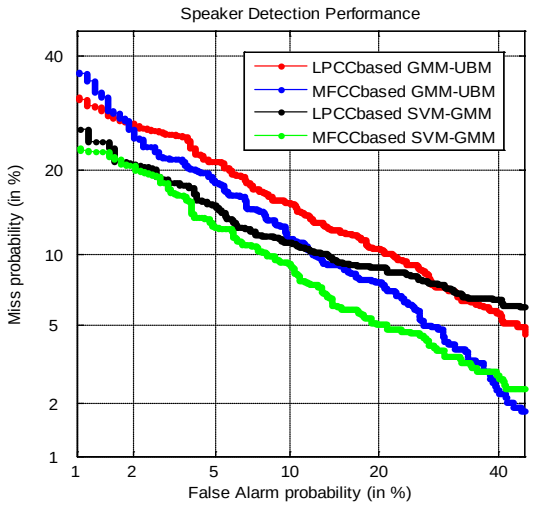


Figure IV.7: Evaluation comparative d'un système RAL à base GMM-UBM. et GMM-SVM.

Il est intéressant de remarquer que la combinaison permet d'améliorer les performances du système. Les résultats de la figure IV.7 sont en faveur du système GMM-SVM puisqu'un gain d'environ 8% est apporté.

IV.4 Normalisation des scores

La plupart des systèmes de vérification du locuteur de l'état de l'art intègrent une étape de normalisation avant la prise de décision, pour renforcer la robustesse d'un système de RAL. Cette étape permet de prendre en compte la variabilité des scores obtenus lors des différents tests. La variabilité provient principalement des différences entre les locuteurs, contenus phonétiques ou durées d'enregistrement d'un test à l'autre. La variabilité intra-locuteur doit également être prise en compte afin de fixer le seuil de décision du système automatique. La variabilité du canal de transmission [140] est souvent nommée variabilité intersession, car c'est la différence de contexte entre plusieurs enregistrements qui la caractérise. Les paramètres, extraits du signal, sont influencés différemment par le canal de transmission et projetés dans des espaces différents. Ceci implique notamment une grande variabilité des scores de vérification. La normalisation de scores a pour but de proposer un score optimal: i) pour chaque locuteur, c'est la Z-norm [141], ii) pour chaque test, c'est la T-norm [142], iii) pour chaque type de combiné, la H-norm.

Les techniques de normalisation sont essentiellement basées sur l'analyse des distributions de scores clients et imposteurs du système de RAL. En général, les approches état de l'art reposent sur une normalisation des distributions de scores imposteurs. Nous inciterons cependant le lecteur à se référer aux travaux cités en [142], [143], [144], [145], [146] et [147]. La plupart des techniques de normalisation consistent à retrancher la moyenne de la distribution des scores imposteurs aux scores de vérification, puis à les diviser par la variance.

$$Score_{Norm} = \frac{Score - \mu_{imp}}{\sigma_{imp}} \qquad \text{(IV.50)}$$

μ_{imp} et σ_{imp} représentent respectivement la moyenne et la variance des scores imposteurs. Il est à noter que le rapport de vraisemblance peut être considéré comme une première étape de normalisation des scores [148].

IV.4.1 Normalisation Z-norm

L'approche Z-norm est une normalisation dépendante du locuteur en utilisant des statistiques pour le locuteur cible. Des énoncés imposteurs sont utilisés comme accès de test. Les scores obtenus par ces accès permettent d'estimer la moyenne μ_{imp} et la variance σ_{imp}, à partir des séquences imposteurs comparées au modèle de locuteur cible (figure IV.8). La Z-norm ne nécessite pas de connaissance sur les énoncés de test. Les paramètres de normalisation peuvent être estimés à partir de l'apprentissage d'un modèle de locuteur. Cette normalisation a pour but de définir un seuil de décision, dépendant du locuteur et de la qualité de son modèle.

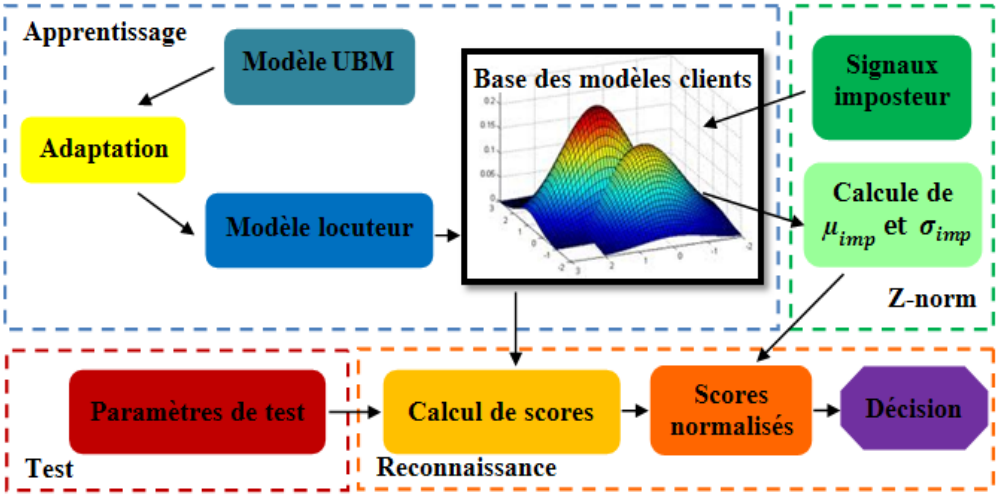


Figure IV.8 : Système GMM-UBM basé sur la normalisation Z-norm.

IV.4.2 Normalisation T-norm

La normalisation T-norm normalise les scores en calculant la distribution des scores du segment test par rapport à des modèles d'imposteurs. Les paramètres μ_{imp} et σ_{imp} sont estimés par les scores de l'énoncé de test sur cet ensemble de modèles de locuteurs imposteurs.

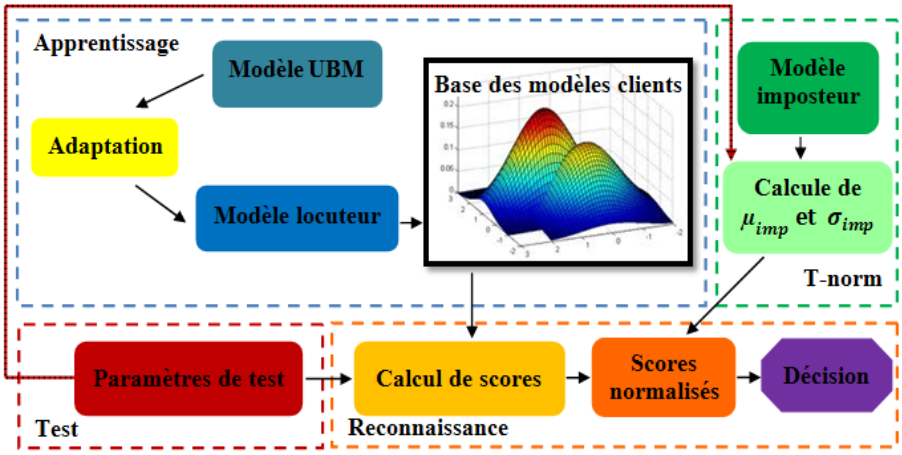


Figure IV.9: Système GMM-UBM basé sur la normalisation T-norm.

IV.4.3 Normalisation H-norm

La normalisation H-norm à été proposée par D. Reynolds [107], à cause des variations du combiné. Cette normalisation consiste à centrer et réduire les effets du combiné utilisé. Supposons que $S(X)$ est le score obtenu suite au test de X sur le modèle, la normalisation de ce score est obtenue par l'équation suivante :

$$S_{Hnorm} = \frac{S(X)-\mu_{\alpha}}{\sigma_{\alpha}} \quad \text{(IV.51)}$$

Où μ_{α} et σ_{α} sont respectivement la moyenne et la variance des accès imposteurs correspondant au type de combiné du X (carbone ou électret).

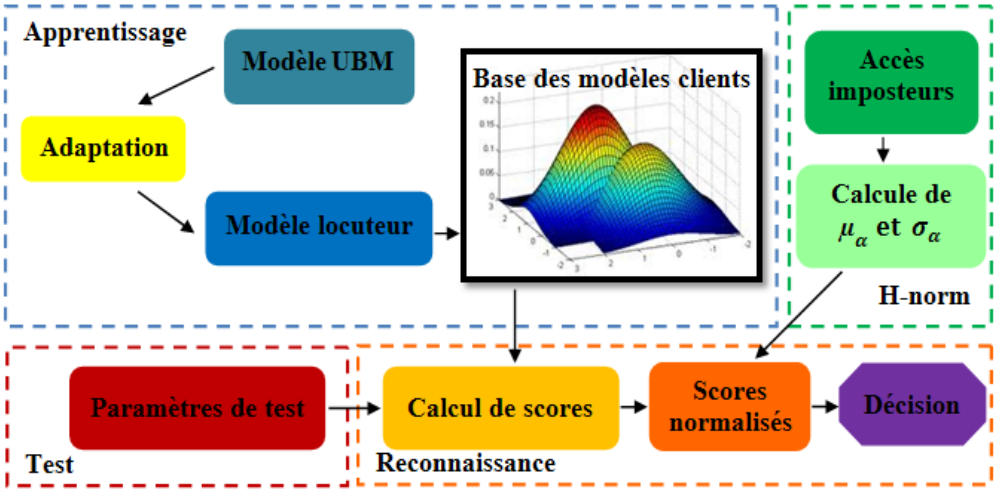


Figure IV.10 : Système GMM-UBM basé sur la normalisation H-norm.

IV.5 Approche de reconnaissance proposée basée sur la fusion des modèles et des scores à base de GMM-UBM et GMM-SVM

IV.5.1 Fusion des scores

Le but des méthodes de fusion est de combiner, associer ou fusionner plusieurs classifieurs qui en général traitent les mêmes données pour exploiter leur complémentarité et améliorer les performances de classification [149], [150]. Il existe de nombreuses méthodes de fusion de scores ; à savoir les méthodes de combinaison simple et les méthodes de modélisation dans l'espace à N dimensions, par exemple dans la figure IV.11, N = 3. La différence entre ces deux types de méthodes réside dans la façon de considérer les scores des systèmes. Les méthodes de combinaison traitent les scores séparément avant de les combiner pour obtenir un score final, alors que les méthodes de modélisation cherchent à séparer les deux scores client et imposteur dans l'espace à N dimensions.

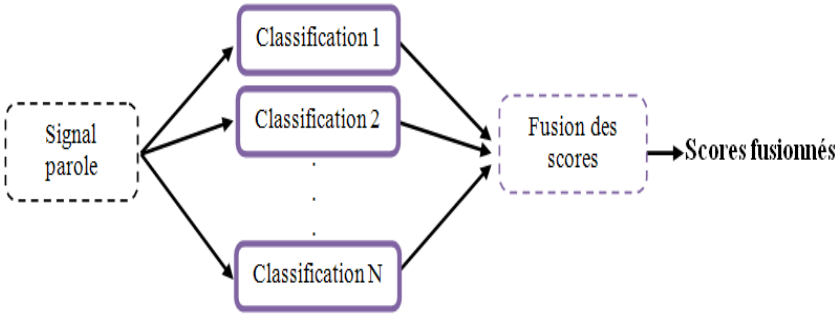


Figure IV.11 : Schéma de fusion des scores.

IV.5.2 Combinaison de scores

Les méthodes de combinaison traitent séparément les M scores disponibles (s_i pour $i = 1$ à M) issus de M systèmes avant de les combiner pour obtenir un score final s. Les fusions les plus utilisées sont la moyenne, le produit, le minimum, le maximum ou la médiane [151].

- Fusion des scores par la moyenne: $s = \frac{1}{N}\sum_{i=1}^{M} s_i$;
- Fusion des scores par le produit: $s = \frac{1}{N}\prod_{i=1}^{M} s_i$;
- Fusion des scores par le minimum: $s = \min(s_i)$;

- Fusion des scores par le maximum: $s = \max(s_i)$;
- Fusion des scores par la médiane: $s = \mathrm{med}(s_i)$;
- Fusion des scores par la somme pondérée: $s = \sum_{i=1}^{M} w_i \, s_i$.

Toutes ces méthodes de combinaison nécessitent une étape préalable de normalisation et ne peuvent être utilisées que si les scores issus des classifieurs sont homogènes.

IV.5.3 Modélisation multidimensionnelle

Les méthodes de modélisation multidimensionnelle considèrent les N scores (s_i) issus des M systèmes comme un seul vecteur de dimension M. On distingue deux types de méthodes de fusion multidimensionnelles : la modélisation des distributions et les méthodes basées sur des classifieurs. La fusion par modélisation des distributions s'appuie sur la théorie de la décision bayésienne et l'estimation des deux densités client et imposteur de distributions multidimensionnelles [152]. Dans notre problème à deux classes client et imposteur, la probabilité a posteriori ne concerne plus chaque score séparément mais le vecteur des M scores issus des M systèmes. Il s'agit donc d'une probabilité a posteriori jointe qui considère les dépendances entre les différents scores.

En reconnaissance du locuteur, nous pouvons citer parmi les travaux qui traitent la fusion; [153], [154], [155], [156],[157], [158], [159], [160], [161], [162], [163], [164], [165] et [166]. La forme de fusion la plus élémentaire consiste à combiner les deux techniques de fusion ; à savoir la modélisation multidimensionnelle et la combinaison de scores, en utilisant la somme pondérée pour calculer les scores d'appariement s_i de M différents systèmes, les poids w_i mesurent la contribution de chacun des classifieurs. Parmi les techniques de recherche des poids on cite la régression logistique (logistic regression) proposée par Brummer et du Preez dans [167]. D'autres travaux se sont basés sur une modélisation des scores des différents systèmes, en utilisant par exemple une SVM ou un réseau de neurones.

IV.5.4 Technique de Fusion proposée basée sur la Normalisation des Scores et la Régression Robuste (Normalised Score based Robust Regression Fusion)

Les techniques de fusions des scores sont fondées rigoureusement et simplement. L'utilisateur peut alors porter des modifications selon l'objectif recherché. Notre contribution se place plus particulièrement dans le cadre de fusion des différents scores résultants des

modèles GMM-UBM et GMM-SVM du système DSR sur IP. Dans le but de répondre aux contraintes lies à l'utilisation des codecs en VoIP, et aux dégradations imposées sur les performances de la RAL, nous proposons dans cette thèse une nouvelle technique de fusion des scores appelée NSRRF (Normalised Score based Robust Regression Fusion) [168]. Cette technique de fusion basée sur la normalisation Min-max des scores, s'appuie sur la régression robuste.

IV.5.4.1 Normalisation Min-max

La normalisation Min-max est une méthode de normalisation de score par remise à l'échelle [169], défini telle que,

$$NS = \frac{s_i - \min(s_i)}{\max(s_i) - \min(s_i)} \quad \text{(IV.52)}$$

La méthode Min-max met chaque score normalisé NS dans l'intervalle $[0;1]$ sous forme de score de similarité, c'est-à-dire, avec les clients proches de la borne supérieure (1) et les imposteurs proches de la borne inférieure (0). Dans le but de transformer chaque score dans un intervalle commun, chaque score issu de chaque modèle est traité séparément par des changements d'échelle pour le translater dans un intervalle précisé et similaire pour chaque modèle. Notre étude basé sur le GMM-UBM et GMM-SVM en utilisant les coefficients MFCC et LPCC extraits du G729 bit-stream, distingue entre quatre types de scores issus de la normalisation Min-max qui sont :

$$NS_{MFCC\_GMM-UBM} = \frac{S_{MFCC\_GMM-UBM} - \min(S_{MFCC\_GMM-UBM})}{\max(S_{MFCC\_GMM-UBM}) - \min(S_{MFCC\_GMM-UBM})} \quad \text{(IV.53)}$$

$$NS_{LPCC\_GMM-UBM} = \frac{S_{LPCC\_GMM-UBM} - \min(S_{LPCC\_GMM-UBM})}{\max(S_{LPCC\_GMM-UBM}) - \min(S_{LPCC\_GMM-UBM})} \quad \text{(IV.54)}$$

$$NS_{MFCC\_GMM-SVM} = \frac{S_{MFCC\_GMM-SVM} - \min(S_{MFCC\_GMM-SVM})}{\max(S_{MFCC\_GMM-SVM}) - \min(S_{MFCC\_GMM-SVM})} \quad \text{(IV.55)}$$

$$NS_{LPCC-GMM-SVM} = \frac{S_{LPCC\_GMM-SVM} - \min(S_{LPCC\_GMM-SVM})}{\max(S_{LPCC\_GMM-SVM}) - \min(S_{LPCC\_GMM-SVM})} \quad \text{(IV.56)}$$

Où $\min(S_{MFCC\_GMM-UBM}, S_{LPCC\_GMM-UBM}, S_{MFCC\_GMM-SVM}, NS_{LPCC\_GMM-SVM})$ et $\max(S_{MFCC\_GMM-UBM}, S_{LPCC\_GMM-UBM}, S_{MFCC\_GMM-SVM}, NS_{LPCC\_GMM-SVM})$ sont respectivement le minimum et le maximum des scores obtenus par les modèles GMM-UBM et GMM-SVM basés sur les paramètres MFCC et LPCC. Les distributions clients et imposteurs de ces scores normalisés sont représentés par les figures IV.12.

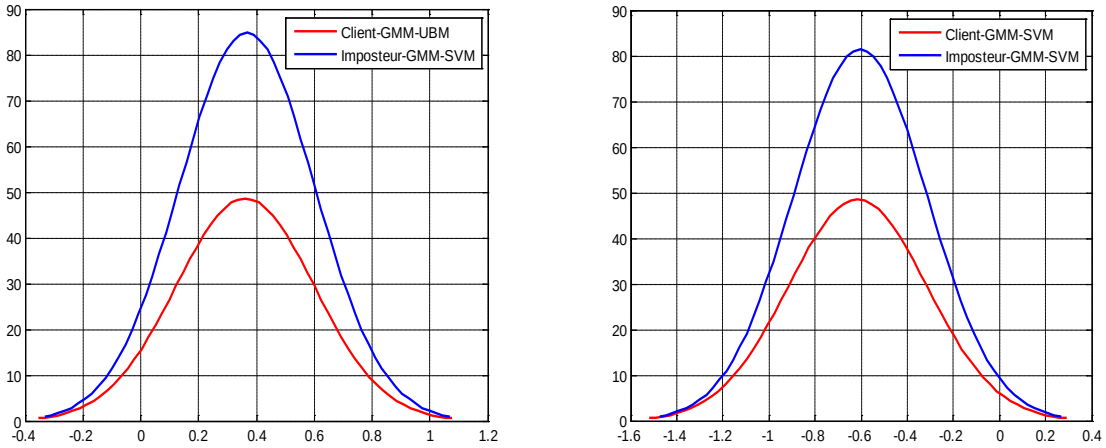


Figure IV.12 : Distributions des scores obtenus par les modèles GMM-UBM et GMM-SVM en utilisant les coefficients LPCC basé sur LSP du G.729 bit-stream.

On remarque que les deux modèles donnent des distributions client et imposteur pas trop différentes dans leur forme et leur plage de variation. Les distributions ont des classes de variances et de formes environ équivalentes, mais les deux classes client et Imposteur sont trop chevauchés entre eux, ce qui rendra nécessaire l'étape de normalisation des scores.

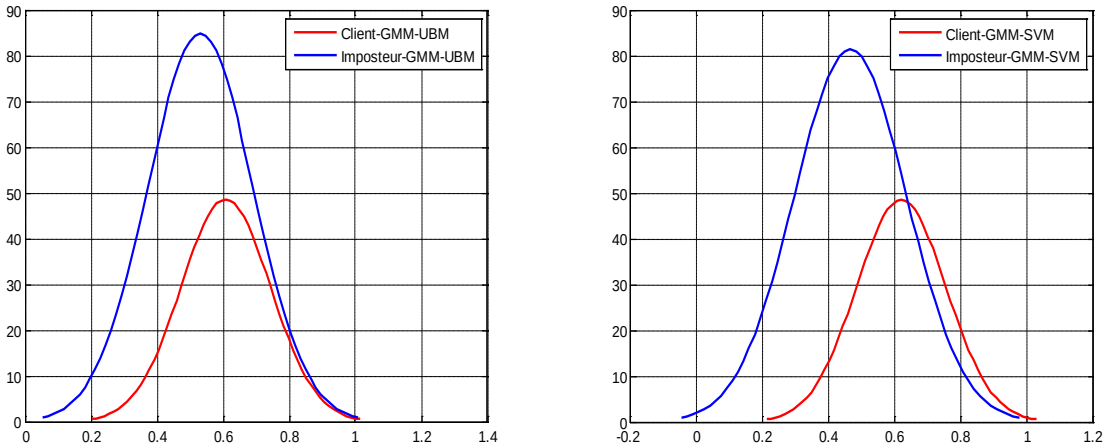


Figure IV.13: Normalisation Min-max appliquée aux scores obtenus par les modèles GMM-UBM et GMM-SVM en utilisant les LPCC basé sur LSP du G.729 bit-stream.

La normalisation Min-max a transformé les scores client et imposteur en scores de similarité où les scores Client sont plus grands que les scores Imposteur. Les scores à

fusionner varient simultanément tous dans le même sens. Ces scores normalisés Min-max sont maintenant comparables et peuvent être combinés.

IV.5.4.2 Fusion par Régression Robuste

L'estimateur robuste est l'estimateur insensible à de petites déviations vis à vis du modèle pour lequel l'estimateur a été optimisé. Ces petites déviations sont toutes les données entachées d'une petite erreur et les quelques données entachées d'une très grosse erreur comme le montre la figure IV.14.

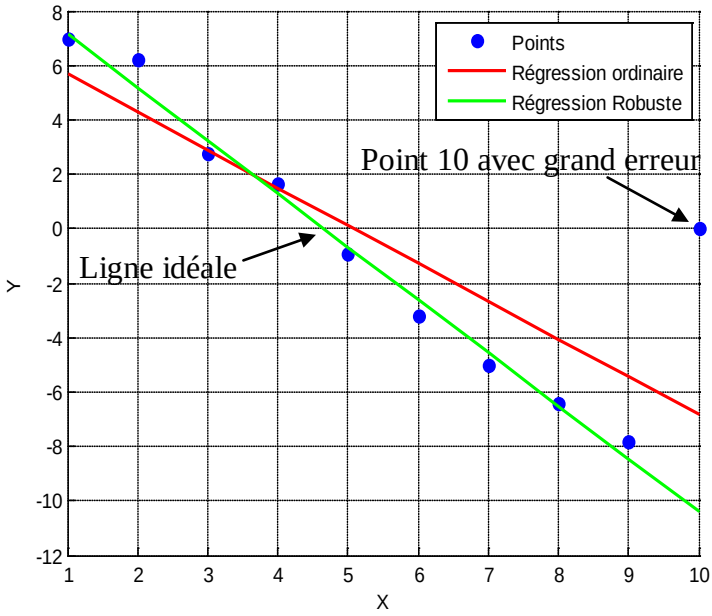


Figure IV.14: Exemple de régression robuste et les pointes d'aberrantes [170].

Ainsi, la régression robuste est une analyse de régression possédant la capacité d'être relativement insensible aux larges déviations dues à certaines observations aberrantes. Dans ce cadre, on se propose dans cette thèse d'appliquer l'estimation robuste de la fonction de régression pour fusionner les scores normalisés Min-max.

Les scores issus de la normalisation Min-max noté NS_j sont fusionnés linéairement par la somme pondérée tel que:

$$S_{Fusion} = \sum_{j=1}^{M} W_j * NS_j \quad \text{(IV.57)}$$

Où, W_j sont les valeurs de poids à estimer, des scores obtenus de M modèles à fusionner par la régression robuste. La méthode la plus courante de régression robuste est la M-estimation, présentée par Huber (Huber, 1964) [171], [172], qui considère le modèle linéaire:

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$
$$y_i = x_i^{'}\beta + \varepsilon_i \quad \text{(IV.58)}$$

Où

y_i : Est le vecteur d'observations ;

$x_i^{'}$: Est une matrice de régresseurs ;

β : Est le vecteur des paramètres inconnus (la pente de régression);

ε_i : Est le vecteur de perturbations (bruit associé au modèle).

Les données aberrantes sont celles pour lesquelles la réponse y_i aux variables dépendantes x_i n'obéit manifestement pas au même modèle que la majorité des autres données.

Il s'agit d'estimer les coefficients β à partir des n observations (y_i, x_i) :

$$\widehat{y_i} = \alpha + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik}$$
$$\hat{y}_i = x_i^{'} b \quad \text{(IV.59)}$$

Où $\hat{\beta} = b$.

Les résidus sont définis comme:

$$e_i = y_i - \hat{y}_i \quad \text{(IV.60)}$$

On remarque également que si, pour une observation *i*, le résidu associé est important, celui-ci peut être une valeur aberrante. Pour faire le calcul de l'estimateur les résidus sont élevés au carré tel que :

$$(e_i)^2 = (y_i - \hat{y}_i)^2 \quad \text{(IV.61)}$$

L'idée des M-estimateurs est de réduire l'incidence de telles observations, Au lieu d'utiliser la fonction $z \to z^2$, on va prendre une autre fonction notée ρ et chercher à minimiser la fonction des résidus (la fonction objective) tel que:

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - x_i^{'} b) \quad \text{(IV.62)}$$

La fonction objective ρ doit respecter les propriétés suivantes:

- $\rho(e) > 0$
- $\rho(0) = 0$
- $\rho(e) = \rho(-e)$
- $\rho(e_i) \geq \rho(e_{i'})\ pour |e_i| = |e_{i'}|$

Pour minimiser $\sum \rho e_i$ on doit mettre la dérivée partielle à zéro. Soit $\psi = \rho'$ est la dérivée de ρ représenté par l'équation d'estimation des coefficients suivante:

$$\sum_{i=1}^{n} \psi(y_i - x_i' b) x_i' = 0 \quad \text{(IV.63)}$$

On définit la fonction de poids par $w(e) = \frac{\psi(e)}{e}$, et $w_i = w(e_i)$ alors $w_i = \frac{\psi(y_i - x_i' b)}{(y_i - x_i' b)}$

Donc, on remplace $\psi(y_i - x_i' b)$ sur l'équation d'estimation on trouve:

$$\sum_{i=1}^{n} w_i (y_i - x_i' b) x_i' = 0 \quad \text{(IV.64)}$$

Pour résoudre cette dernière équation on minimise $\sum (w_i)^2 (e_i)^2$ en utilisant la méthode Iterated Reweighted Least Squares (IRLS) [173]. Une fois que le poids W est estimé, les scores client et les scores imposteur sont fusionnés tel que:

$$T_i = \sum_{i=1}^{M} W_i V_i \quad \text{(IV.65)}$$
$$R_k = \sum_{k=1}^{M} W_k U_k \quad \text{(IV.66)}$$

Où $W_i\ et\ W_k$ sont respectivement les poids des scores clients V et les scores imposteurs U obtenus à partir de M modèle.

Notre travail est basé sur le modele GMM-UBM et GMM-SVM, chaque modèle exploite deux types de coefficients MFCC et LPCC, on obtient donc deux types de scores normalisée Min-max et fusionnés par la régression rebuste.

$$S_{GMMUBM-Fusion} = W_{MFCC\_GMM-UBM} * NS_{MFCC\_GMM-UBM} + W_{LPCC\_GMM-UBM} * NS_{LPCC\_GMM-UBM} \quad \text{(IV.67)}$$

$$S_{GMMSVM-Fusion} = W_{MFCC\_GMM-SVM} * NS_{MFCC\_GMM-SVM} + W_{LPCC\_GMM-SVM} * NS_{LPCC\_GMM-SVM} \quad \text{(IV.68)}$$

Ensuite, une fusion finale par la régression robuste est appliqué aux scores GMM-UBM et GMM-SVM tel que

$$S_{Fusion\ -Final} = W_{GMMUBM\ -Fusion} * S_{GMMUBM\ -Fusion} + W_{GMMSVM\ -Fusion} * S_{GMMSVM\ -Fusion} \quad \text{(IV.69)}$$

IV.6 Conclusion

Nous avons présenté dans ce chapitre les deux principales approches de modélisation en RAL pour l'application ciblée par ce travail, GMM-UBM et GMM-SVM.

Dans la première partie de ce chapitre, nous avons tenu à présenter l'architecture du système GMM-UBM, à base de modèles de mélange de gaussiennes GMM et le modèle du monde UBM, dont l'objectif est d'aboutir à une modélisation générative. Nous avons abordé ensuite le rôle structurel du modèle du monde.

Dans la deuxième partie, nous avons présenté la méthode de modélisation discriminante introduite par Vladimir Vapnik les SVM. Cette méthode de classification est basée sur la recherche d'un hyperplan qui permet de séparer au mieux des classes de données dans les deux cas linéairement séparable et non linéairement séparables basée sur les fonctions noyaux (kernel). Les SVMs sont aujourd'hui l'une des méthodes les plus utilisées en RAL, grâce à leur pouvoir de généralisation, leur capacité à traiter des problèmes de grande dimension et leur mise en œuvre aisée. Malgré la grande capacité de généralisation des SVM, l'erreur de classification ne peut pas être complètement éliminée. Dans ce cas, nous exposons la combinaison des méthodes génératives et discriminantes, nommée GMM-SVM adoptée comme une réponse aux problèmes d'hétérogénéité et de complexité généralement liés aux systèmes intégrant les paradigmes génératif et discriminant.

Cependant, les modèles de reconnaissance ont des limitations principales en termes de performances sur les réseaux IP. L'utilisation des codecs entraîne des pertes de performances en reconnaissance sur IP. Dans le but d'améliorer ces performances en exploitant les sores des deux modèles étudie auparavant, nous avons proposé une approche originale de fusion: la régression robuste. Le principe de cette stratégie est l'estimation robuste des poids pour fusionner les score normalisés Min-max des systèmes GMM-UBM et GMM-SVM. Cette stratégie permet de garantir une amélioration importante des performances de la reconnaissance sur IP, mais elle permet surtout de réduire les contraintes et les dégradations liées aux applications du codec G729 et VoIP.

CHAPITRE V

Evaluation expérimentale

V.1 Introduction

Ce chapitre traite de la mise en œuvre d'un système de reconnaissance automatique de locuteur distribuée, 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 plateforme Alize, la base de données utilisée, la base de données transcodée G729, et dans un second temps, l'analyse acoustique appliquée. Nous présenterons également les performances de la reconnaissance du locuteur, en mode indépendant du texte, basée sur la méthode GMM-UBM et GMM-SVM. Ces performances sont évaluées en fonction de l'ordre des modèles, et du nombre des coefficients. Ensuite, nous examinons les performances de la reconnaissance du locuteur distribuée à travers une étude comparative utilisant les modèles GMM-UBM et GMM-SVM. La normalisation Min-max aux scores issus des deux modèles est appliquée. Enfin, nous présentons les résultats obtenus avec la méthode robuste de fusion proposée pour améliorer les performances de la reconnaissance du locuteur distribuée.

V.2 Outils de programmation utilisés

V.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.

V.2.2 C++

Le 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.

V.2.3 Perl

Perl est un langage de programmation reprenant des fonctionnalités du langage C et des langages de scripts comme le Shell (sh), ce langage étant particulièrement adapté au traitement et à la manipulation de fichiers texte.

V.2.4 ALIZE

ALIZE est une plateforme libre qui permet de faciliter le développement d'applications dans le domaine du traitement automatique du locuteur. Elle a été développée au Laboratoire d'Informatique d'Avignon (LIA) par Frédéric Wils sous la direction de Jean-François Bonastre depuis février 2003. Basés sur cette plateforme, différents outils ont été implémentés, dont LIA_SpkDet pour la reconnaissance vocale. Originellement LIA_SpkDet faisait partie du package LIA_RAL mais depuis la version 1.3 LIA_RAL s'est retrouvée divisée (voir figure V.1).

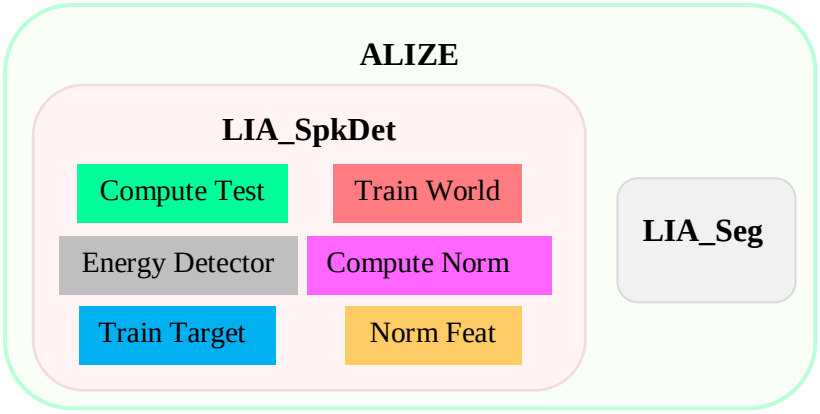


Figure V.1: Organisation de la Plateforme ALIZE.

D'un point de vue de l'architecture de base, la plate-forme est construite autour de plusieurs modules de données et de calcul :

- Le module de features qui va stocker les caractéristiques (features) issues soit d'un fichier, soit d'un calcul réalisé à partir des données audio. Il s'agit également d'un tampon de durée virtuellement illimitée.
- Le module de mixtures/distributions sert à stocker les modèles de parole (mixtures de gaussiennes) calculés à partir des features ou chargés à partir de fichiers.

- Le module de statistiques regroupe les algorithmes de base les plus courants (calcul de vraisemblance, EM, etc.) et permet de conserver et d'accumuler les résultats de calculs pour réaliser des moyennes sur un ensemble de features.

Cependant à l'heure actuelle il n'existe pas une architecture sur IP permettant d'exploiter les fonctions de ce programme. Dans le contexte de la VoIP il apparaît évident d'utiliser une architecture client-serveur.

V.3 Description des bases de données

V.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 [174]. 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)

V.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, ou bien extraire directement les LSP à partir du G.729 bit-stream et reconstruire la base de données LSP-G.729-bit-stream.

Dans ce travail nous exploitons trois 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;
3. LSP-G.729-bit-stream : la base de données extraite directement du LSP basé G.729 bit-stream, envoyé par le client au serveur en utilisant le protocole UDP.

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

V.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.

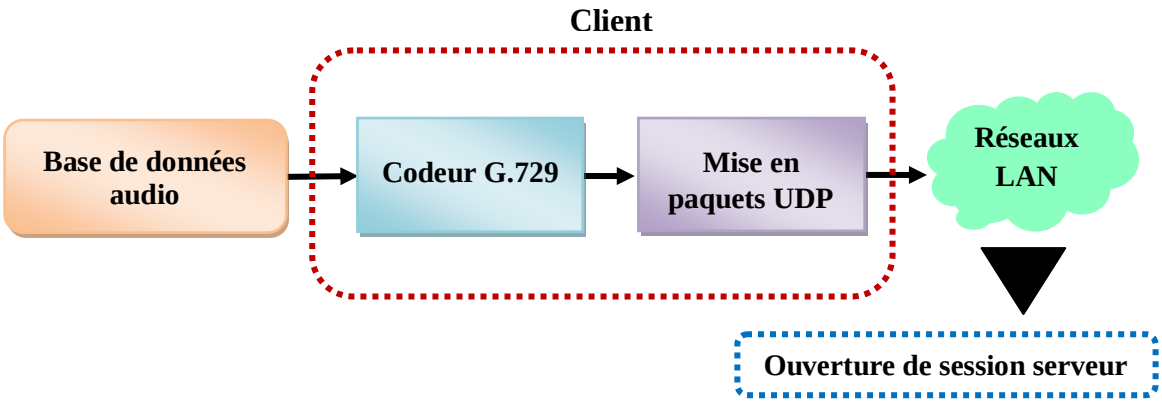


Figure V.2 : Implémentation coté client.

V.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.

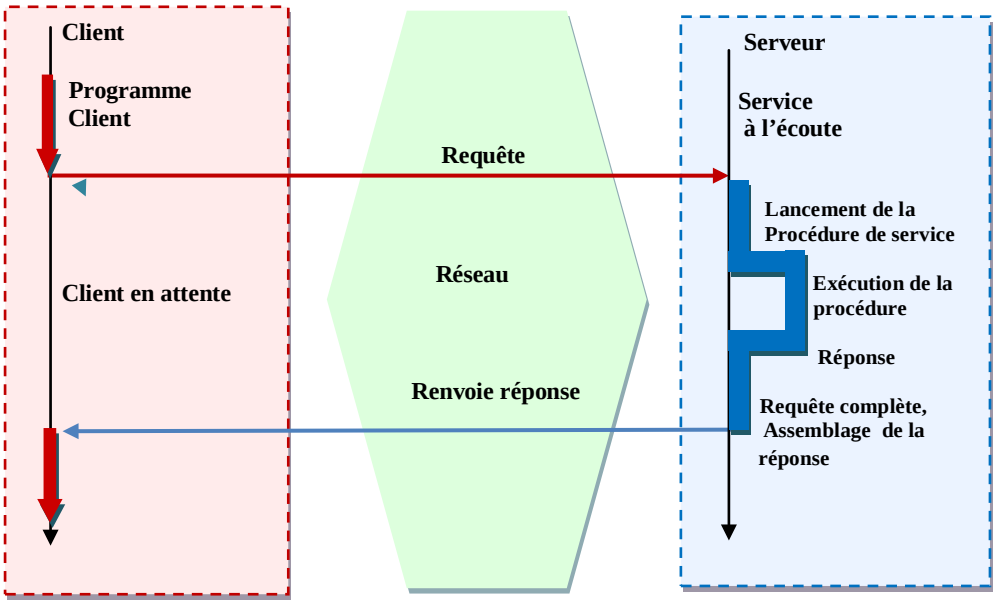


Figure V.3 : communication entre client serveur.

V.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.

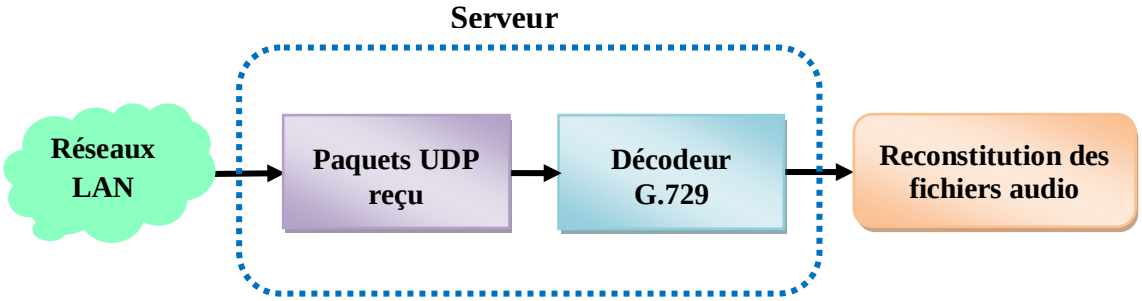


Figure V.4: Implémentation coté serveur.

V.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, deux différentes méthodes sont utilisées pour l'extraction des vecteurs caractéristiques ; MFCC et LPCC, 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 et LPCC directement à partir des LSP (lignes de raies spectrales) basés sur le G.729 bit-stream (LSP-G.729-bit-stream), comme le montre la figure ci-dessous.

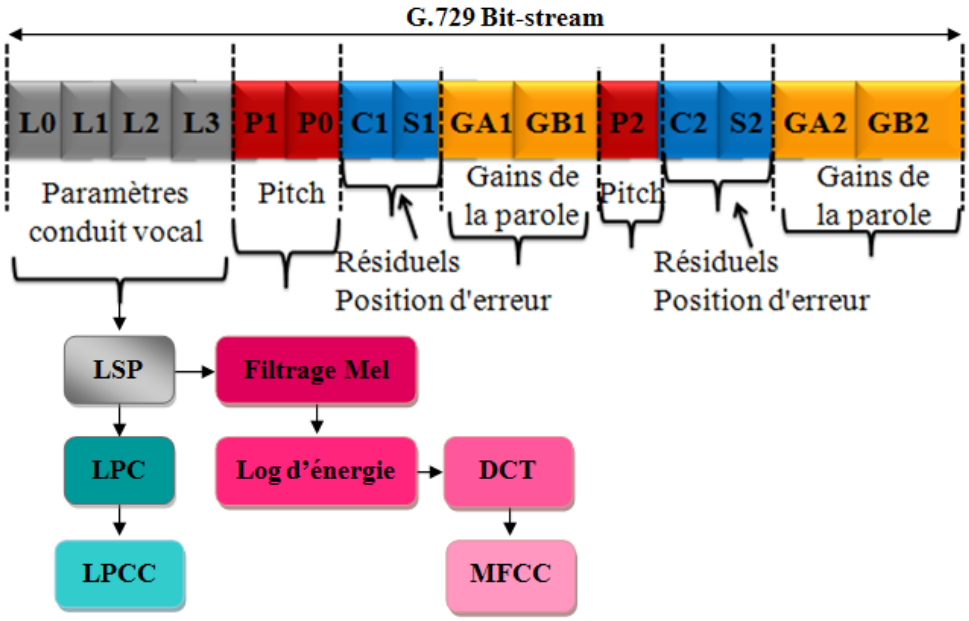


Figure V.5: Extraction des MFCC et LPCC basée sur LSP-G.729-bit-stream.

V.6 Evaluation des performances

Avant de procéder à la phase d'évaluation des performances en mode distribué nous examinons d'abord le chemin optimal qui donne les meilleures performances de reconnaissance du locuteur sans codec.

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

Dans cette expérience, nous étudions les performances des deux systèmes GMM-UBM et GMM-SVM, en utilisant 40 paramètres MFCC et LPCC (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 reconnaissance du locuteur dans le cas où on a très peu de données d'apprentissage (4 à 5 secondes de parole).

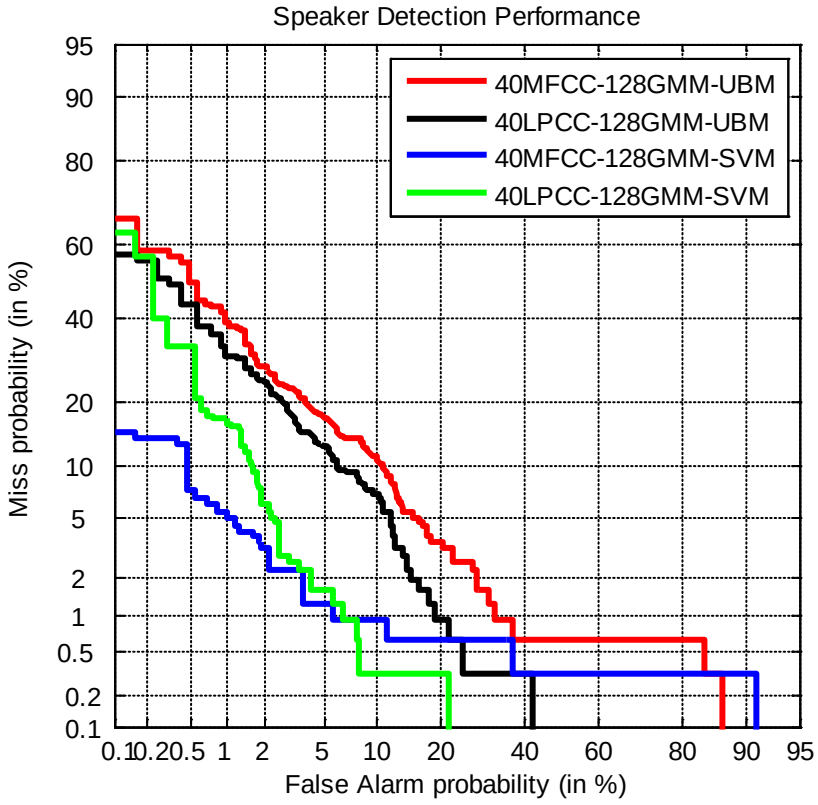


Figure V.6: Performance d'un système RAL pour 128 gaussiennes de modèle GMM-UBM et GMM-SVM.

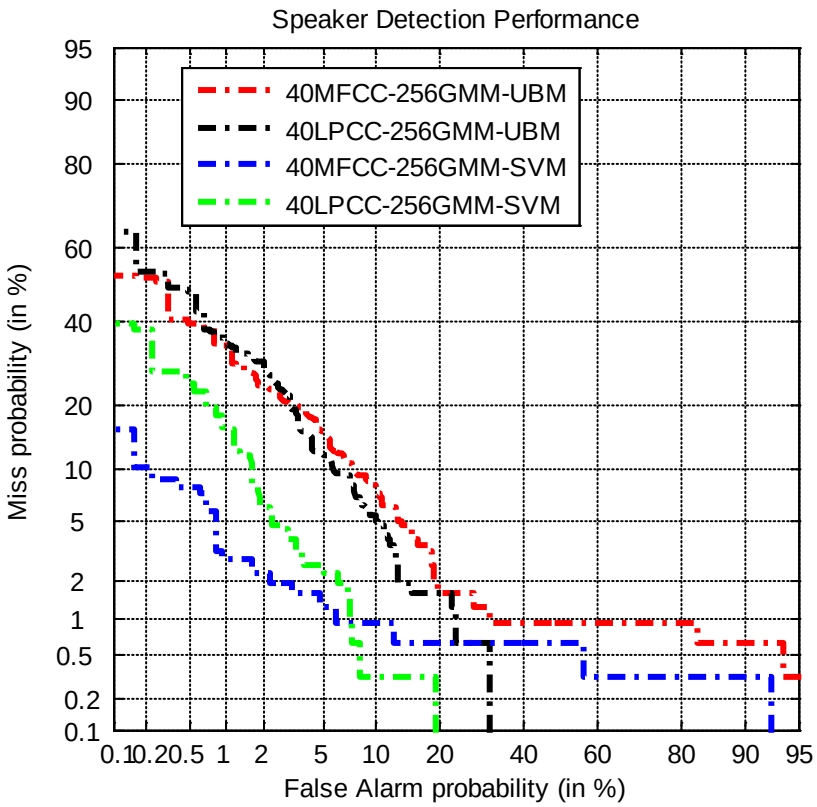


Figure V.7: Performance d'un système RAL pour 256 gaussiennes de modèle GMM-UBM et GMM-SVM.

Les figures V.6 et V.7 montrent que l'augmentation du nombre de gaussiennes de 128 à 256 n'apporte pas une amélioration des performances, pour les deux type de coefficients MFCC et LPCC basée sur les deux systèmes GMM-UBM et GMM-SVM, au contraire les performances se dégrade 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).

La figure V.6 montre que les performances de vérification sont meilleures dans le cas des coefficients LPCC, par rapport au MFCC. Pour un nombre de modèles de 128 GMM-UBM le taux de reconnaissance correct est de 91%. Il est intéressant de remarquer aussi que la combinaison SVM-GMM permet d'améliorer les performances du système, les résultats de

la figure V.6 sont en faveur du système GMM-SVM puisqu'un gain d'environ 8% basé MFCC et 7% basé LPCC est apporté, par rapport au système GMM-UBM.

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 et LPCC sur les performances d'identification en exploitent la base ARADIGIT8K. Pour cela on fixe l'ordre du model GMM-UBM à 128.

Les figures V.8 et V.9 présentent l'évaluation des performances de vérification du modèle GMM-UBM pour 40 et 60 coefficients MFCC et LPCC extraites directement de la base ARADIGIT8K. On remarque qu'avec 40 coefficients LPCC on obtient le taux de reconnaissance correct de 91% pour le modèle GMM-UBM, 90% en utilisant les MFCC, et un meilleur taux correct de 98% pour le modèle GMM-SVM basé MFCC.

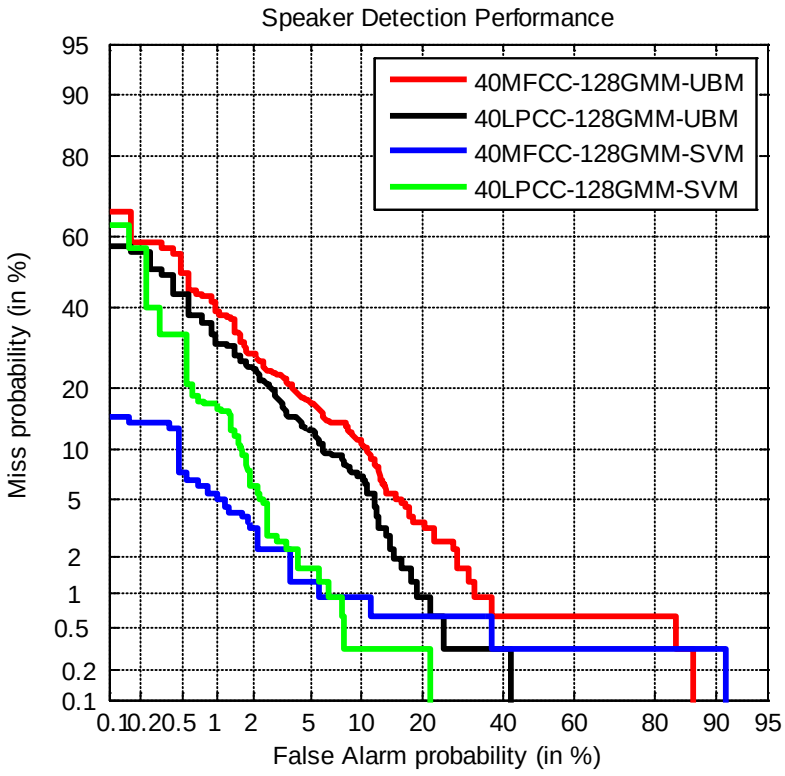


Figure V.8: Performance d'un système RAL à base de 128 gaussiennes de modèle GMM-UBM et GMM-SVM en utilisant 40 paramètres MFCC et LPCC.

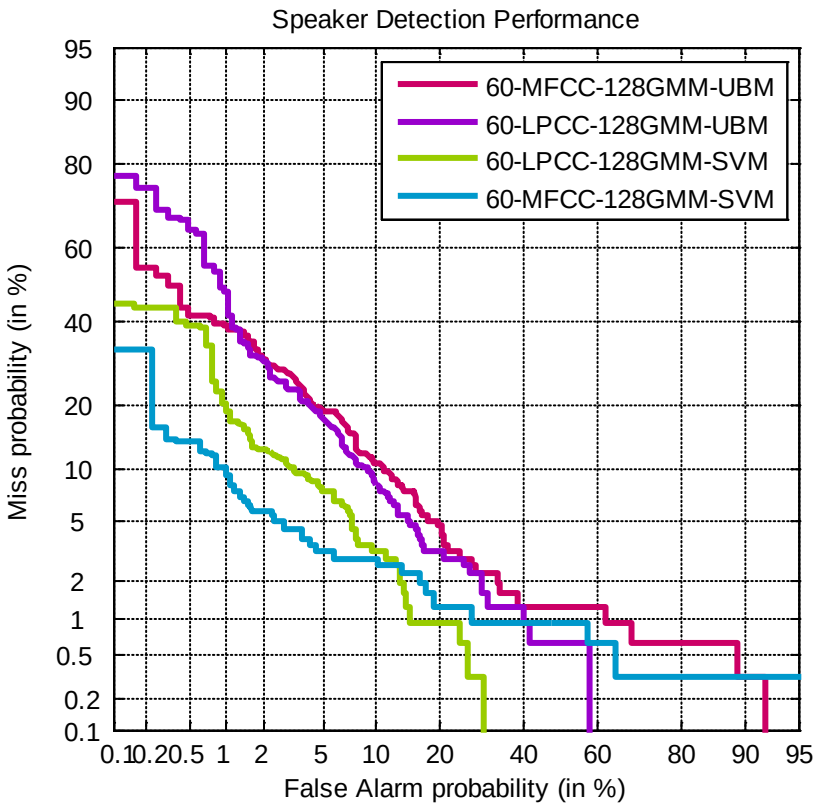


Figure V.9: Performance d'un système RAL à base de 128 gaussiennes de modèle GMM-UBM et GMM-SVM en utilisant 60 paramètres MFCC et LPCC.

Les figures V.8 et V.9 présentent l'évaluation des performances de vérification du modèle GMM-SVM pour 40 et 60 coefficients MFCC et LPCC. On remarque bien que lorsque le nombre des paramètres (MFCC et LPCC) augmente les performances tendance à rester stable ou diminuer. Il est donc inutile de prendre un nombre des paramètres supérieur à 40. La figure V.9 montre que l'utilisation de 40 coefficients MFCC avec le modèle GMM-SVM permet d'obtenir le meilleur taux de reconnaissance correct, soit 96%.

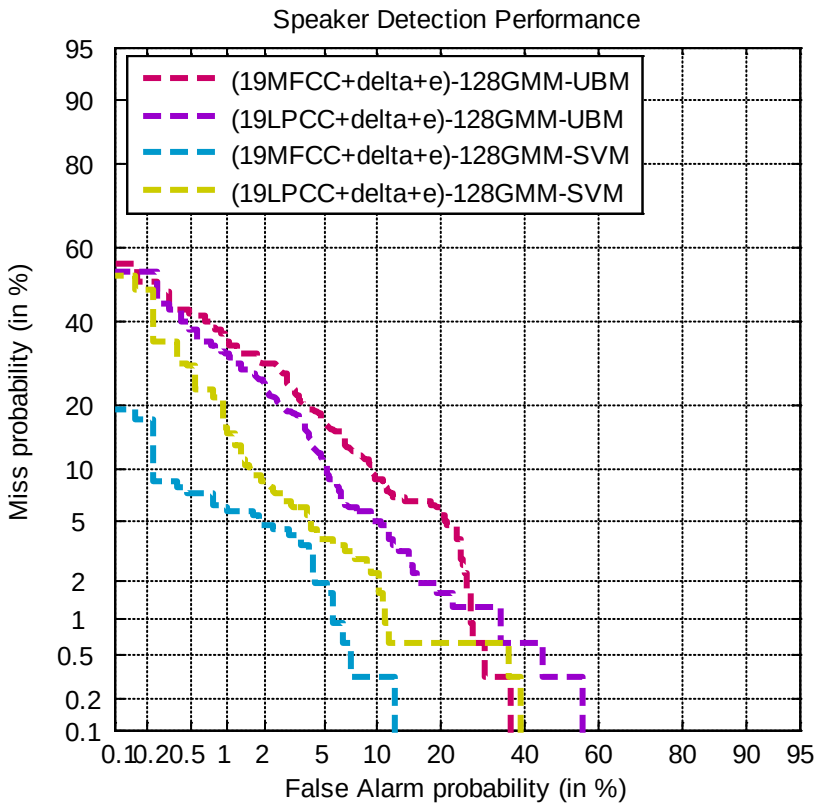


Figure V.10: Performance d'un système RAL à base de 128 gaussiennes de modèle GMM-UBM et GMM-SVM en utilisant 19 paramètres MFCC et LPCC avec leur delta et l'énergie.

La figure V.10 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) et 40(19LPCC+19delta+2energie), ceci est observé pour les deux type des modèles. 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 les approches GMM-UBM et GMM-SVM.

A partir de ces résultats, nous constatons que le taux de reconnaissance varie considérablement en fonction de la méthode utilisée. Pour la même méthode, le taux de vérification varie avec les paramètres choisis. Nous constatons aussi que la méthode GMM-SVM donne un meilleur taux de reconnaissance, par rapport à la méthode GMM-UBM, dans les mêmes conditions d'enregistrement et pour le même nombre de locuteurs.

D'après ces résultats, et pour la meilleure configuration que nous devrons utiliser dans la suite des expériences nous 'utilisons 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 reconnaissance du locuteur distribuée. Dans ce but, on applique le G.729 à la base de données ARADIGIT8K pour obtenir la base transcodée G.729 (G729-ARADIGIT8K). L'évaluation des performances de reconnaissance du locuteur distribué est basée sur le modèle GMM-UBM et GMM-SVM. Les résultats obtenus sont illustrés dans la figure V.11 :

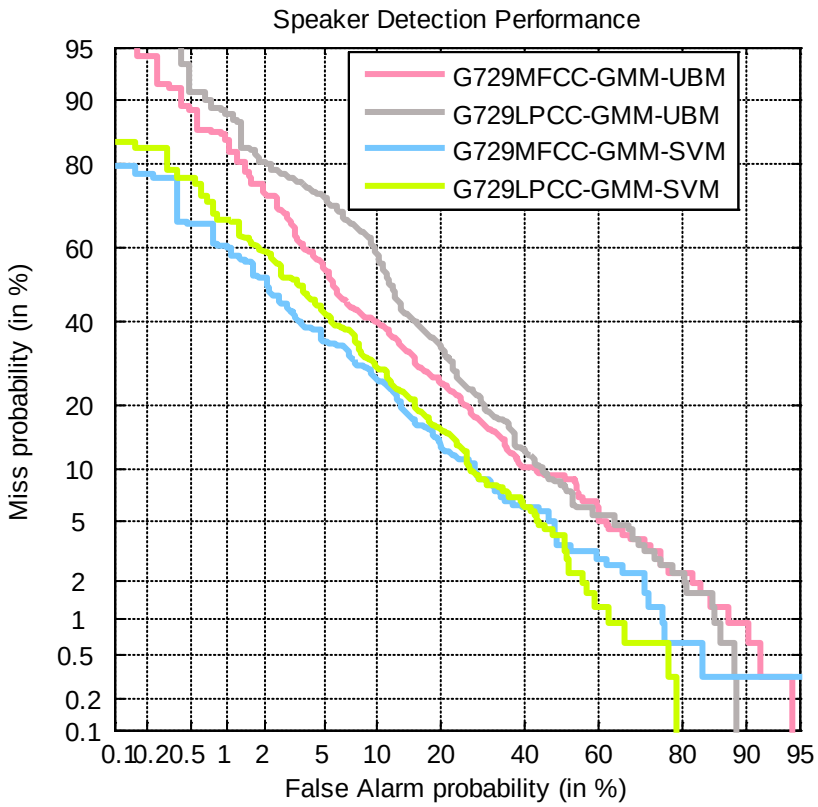


Figure V.11: Les performances d'un système RAL distribuée en utilisant le codec G.729 avec modélisation GMM-UBM et GMM-SVM.

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 65% pour les LPCC basé GMM-UBM, à 78% avec l'utilisation des MFCC basé GMM-UBM, à 77% pour les LPCC basé GMM-SVM et à 82% pour les MFCC basé GMM-SVM. Malgré la dégradation imposée par le G.729 sur les performances de la reconnaissance distribuée, le modèle GMM-SVM présente toujours les meilleurs performances.

La figure V.12 illustre la comparaison entres les performances de la reconnaissance basée sur la base de données G.729-ARADIGIT8K over IP et celles obtenues en utilisant la base ARADIGIT8K sans codage.

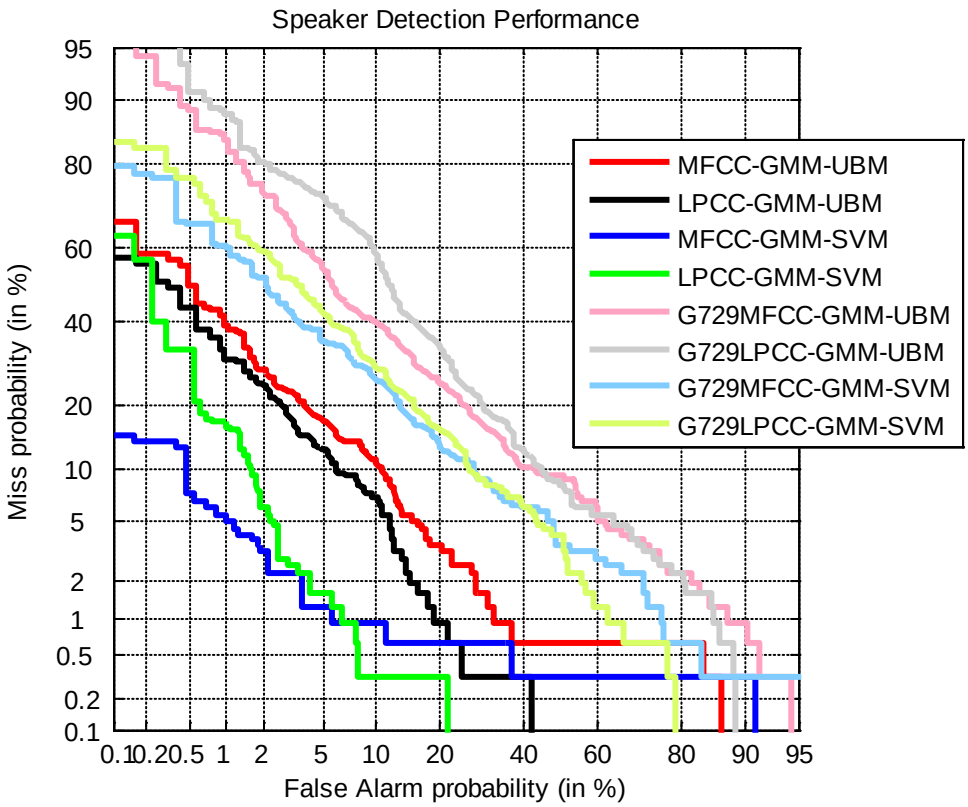


Figure V.12: Les performances d'un système RAL normal et distribué basé sur le codec G.729 en utilisant le modèle GMM-UBM et GMM-SVM.

Les courbes DET montrent une dégradation significative des taux de reconnaissance. Elle s'élève à plus de 10 % pour les deux systèmes GMM-UBM et GMM-SVM en utilisant les MFCC et LPCC extraite à partir de la base G.729-ARADIGIT8K par rapport aux résultats basés sur ARADIGIT8K. Le meilleur résultat obtenu correspond au modèle GMM-SVM basé sur les MFCC extraits à partir de la base ARADIGIT8K sans codec.

Dans cette expérience, nous avons comparé nos résultats de la reconnaissance du locuteur avec une reconnaissance distribuée utilisant la reconstruction du signal (parole synthétisée) au niveau du serveur avant la reconnaissance. 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 reconnaissance 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 [175].

V.6.4 Influence du G.729 bit-stream sur ARADIGIT8K

Cette expérience exploite les coefficients MFCC et LPCC dérivés directement du LSP basé G.729 bit-stream au niveau du serveur pour éviter la reconstruction du signal issue du G.729 encodeur. L'objectif est de présenter l'avantage de faire l'économie d'une reconstruction du signal très couteuse en temps machine, ce qui pourrait influer sur la qualité de la DSR, sous peine de dégrader la qualité de service (QoS).

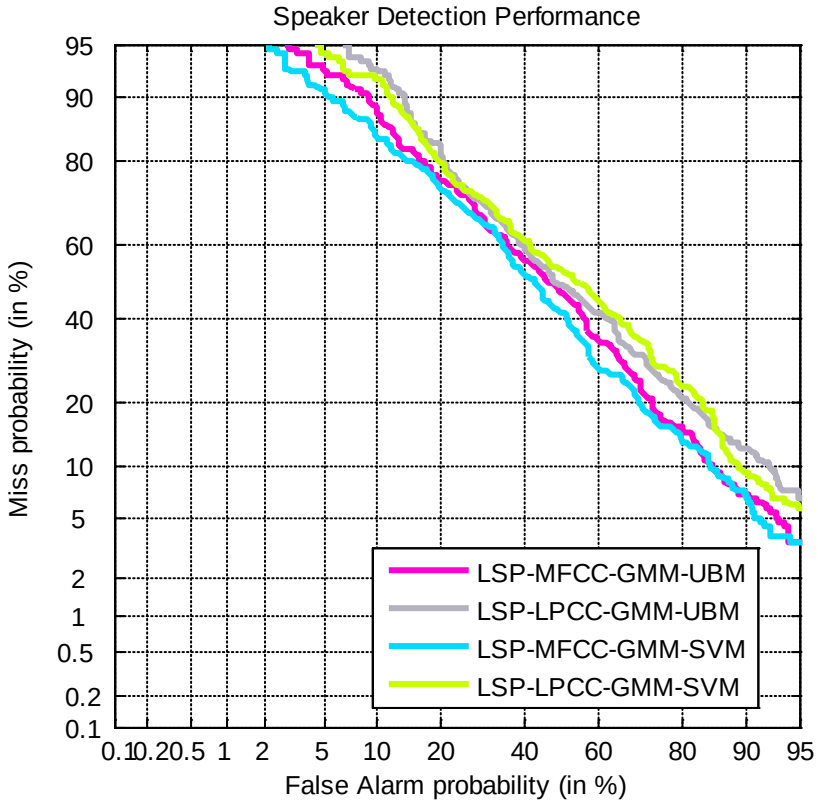


Figure V.13: Les performances d'un système RAL distribué en utilisant les LSP du G.729 bit-stream à base GMM-UBM et GMM-SVM.

La figure V.13 donne les performances de reconnaissance distribuée des modèles GMM-SVM et GMM-UBM basés MFCC et LPCC extraites du LSP-G.729-bit-stream. Les

résultats obtenus, bien qu'avec des taux de reconnaissance encore faibles sont très encourageants.

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 DSR. 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.

Dans l'objectif d'avoir une séparabilité entre les distributions client et imposteur, ainsi d'améliorer les résultats de reconnaissance obtenus précédemment, nous avons investigué la normalisation Min-max des scores présentés au chapitre 5. La figure ci-dessous présente les courbes DET des résultats de normalisation des scores.

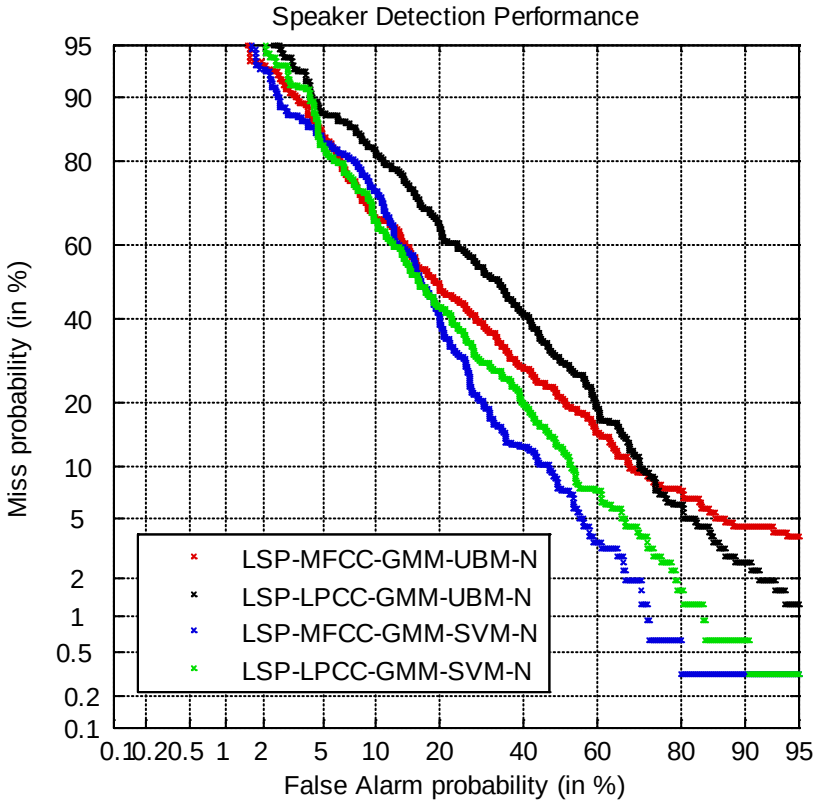


Figure V.14: Les performances d'un système RAL distribué en utilisant les LSP du G.729 bit-stream basé sur la normalisation des scores issus du GMM-UBM et GMM-SVM.

D'après les résultats obtenus et présentés par la figure V.14, on observe que les performances de vérification augmentent. Ces résultats sont valables pour les deux types de

paramètres basés sur les deux modèles GMM-UBM et GMM-SVM. Les résultats obtenus par GMM-SVM sont meilleurs que les résultats de GMM-UBM. Pour le reste des résultats, les scores normalisés sont nommés GMM-UBM-N et GMM-SVM-N. La figure V.15 illustre une comparaison entres les résultats des scores obtenus avec et sans normalisation.

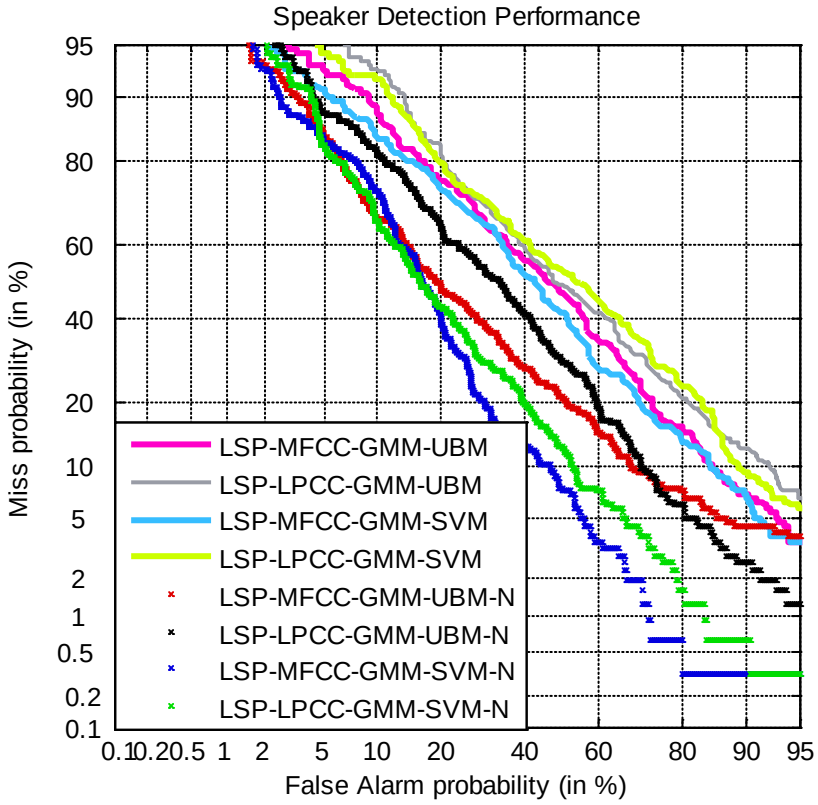


Figure V.15: Les performances d'un système RAL distribué en utilisant les LSP du G.729 bit-stream basé sur les scores avec et sans normalisation issus du GMM-UBM et GMM-SVM.

Sur la figure V.15, nous avons représenté les variations du taux de reconnaissance avec et sans normalisation des scores issus de la modélisation GMM-UBM et GMM-SVM, en fonction des paramètres MFCC et LPCC extraites a partir des LSP basé G.729-bit-stream. Les taux de reconnaissance croissent significativement avec la normalisation, pour les deux type de modélisation GMM-UBM et GMM-SVM, même jusqu'à 70% pour GMM-SVM-N basé MFCC. Ces résultats confirment le caractère discriminant de la normalisation. Pour améliorer ces résultats nous avons proposé une méthode de fusion originale permettant

d'améliorer les performances. Cette méthode, présentée au chapitre 5, exploite la régression robuste pour calculer le poids des scores normalisés.

En revanche, les exigences de performance sont impactées par notre choix de la méthode de fusion. En effet, selon l'application DSR étudiée, qui repose sur le modèle GMM-UBM et GMM-SVM, les performances attendues ne sont pas les mêmes, surtout pour les deux modèles GMM-UBM et GMM-SVM, basés sur deux types de paramètres. Les scores obtenus et le temps de reconnaissance sont plus liés au choix des modèles et de leur nombre qu'au choix de la méthode de fusion proprement dite. En effet, la fusion des scores doit être examinée séparément pour chaque système à combiner. L'étude des performances individuelles peut nous aider à choisir le modèle adéquat; et ainsi, de connaître préalablement à l'étape de fusion, les différentes performances individuelles de chaque modèle. On ne combinera pas de la même façon deux systèmes avec des performances à peu près équivalentes et deux systèmes avec des performances très différentes, car dans ce dernier cas, on exigera que le bon système ait plus de poids dans la décision que le moins bon.

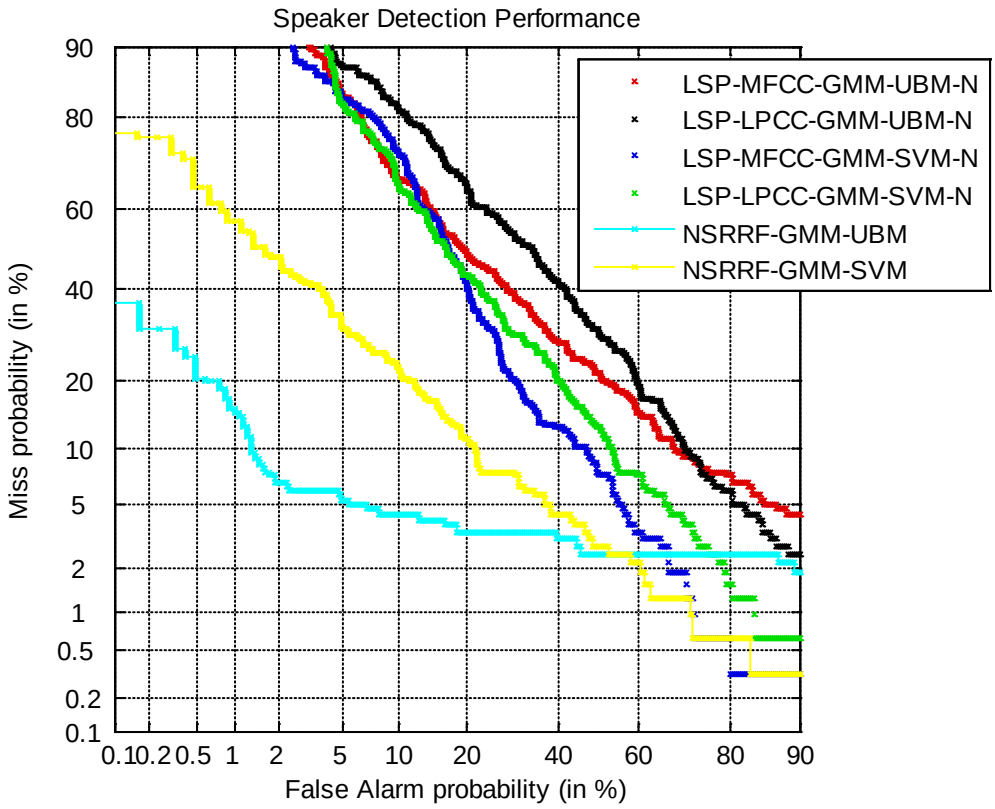


Figure V.16: Les performances d'un système RAL distribué en utilisant les LSP du G.729 bit-stream basé sur la fusion des scores normalisé issus du GMM-UBM et GMM-SVM.

La figure V.16 donne les performances de reconnaissance distribuée des scores normalisés fusionnés des modèles GMM-SVM-N et GMM-UBM-N. Les résultats sont satisfaisants. Le taux de reconnaissance de GMM-UBM-N basé MFCC fusionné GMM-UBM-N basé LPCC nommée NSRRF-GMM-UBM (Normalised Score based Robust Regression Fusion) est de 95%. Le taux de reconnaissance de GMM-SVM-N basé MFCC fusionné GMM-SVM-N basé LPCC nommée NSRRF-GMM-SVM est de 85%.

Nous avons montré que la fusion des scores normalisés de l'approche GMM-UBM-N basée LPCC fusionnée avec GMM-UBM-N basé MFCC et l'approche GMM-SVM-N basée LPCC fusionnée GMM-SVM-N basée MFCC, lors des expériences réalisées, surpassent les performances du système GMM-UBM-N et GMM-SVM-N seules. La fusion devrait apporter une information absente du processus de vérification du locuteur indépendant du texte. Une complémentarité des informations fournies par les deux scores, devrait ainsi bénéficier à la fusion. Ainsi, nous proposons dans ce travail la fusion des scores NSRRF-GMM-UBM et les scores du NSRRF-GMM-SVM.

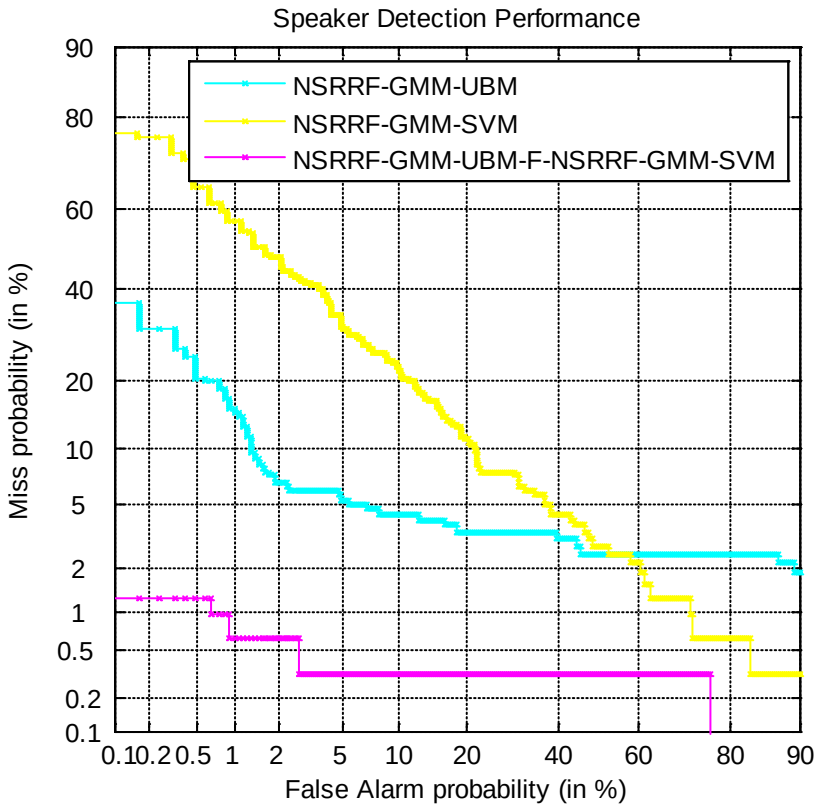


Figure V.17: Les performances d'un système RAL distribué à base de fusion de données par la régression robuste: étude comparative entre NSRRF-GMM-UBM et NSRRF-GMM-SVM.

La figure V.17 illustre les performances des deux systèmes NSRRF-GMM-SVM et NSRRF-GMM-UBM fusionnés, La combinaison des deux systèmes à l'aide de la fusion proposée montre un gain significatif comparé aux systèmes NSRRF-GMM-UBM et NSRRF-GMM-SVM. La courbe DET donne un taux de reconnaissance correcte jusqu'à de 99%, des scores basés sur le système NSRRF-GMM-SVM fusionné avec NSRRF-GMM-UBM. Les résultats obtenu sont très satisfaisants, et ont permis de confirmer les bonnes qualités de la fusion proposée.

V.7 Conclusion

Ce chapitre a porté sur le développement et la mise en œuvre complète d'un système DSR basé sur l'architecture client/serveur. Une mise en œuvre du côté du client avec son unité d'encodeur G729 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 plateforme ALIZE. 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. Dans notre approche, et pour faire l'économie d'une reconstruction et éviter d'influer sur la qualité de la DSR, les paramètres pertinents sont dérivés directement du LSP basé G.729 bit-stream. Dans l'objectif d'améliorer les performances du système distribué en exploitant ces paramètres, nous avons proposé une méthode efficace de fusion des scores basés sur la régression robuste permettant d'augmenter les performances et d'apporter un gain significatif. Les résultats obtenus par la DSR basée sur l'architecture client-serveur utilisant le codec G.729 exploitant la fusion proposé des scores normalisés sont prometteurs, comparativement avec ceux obtenus avec la DSR utilisant le codec G.729 sans fusion. Ces résultats dépassent aussi les performances d'un système de reconnaissance conventionnel. D'après les résultats obtenus, nous pouvons conclure que la méthode originale de fusion proposée est une méthode efficace de fusion des scores et bien adaptée pour la reconnaissance de locuteur distribuée.

CONCLUSION GENERALE ET PERSPECTIVES

CONCLUSION GENERALE ET PERSPECTIVES

Ce travail de recherche s'inscrit dans le cadre d'un axe émergeant à savoir la reconnaissance automatique du locuteur, en mode indépendant du texte, utilisant les réseaux de communications avec la VoIP. Dans cette thèse, nous avons proposé des techniques de traitements vocaux dédiés à vérification de locuteurs dans un système distribuée (DSR) 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. Les travaux présentés dans cette thèse s'inscrivent dans le cadre de cette tâche et sont orientés autour de quatre axes principaux :

i) La modélisation générative GMM-UBM, est à la base de nos travaux. Nous l'avons utilisée comme système de référence tout au long de cette thèse. Pour cela, nous avons mené une étude approfondie des comportements de l'approche GMM-UBM dans le contexte de la reconnaissance du locuteur distribué.

ii) Cependant, ce modèle génératif n'est pas sans dérives. Ceci a conduit récemment à l'émergence d'approches discriminantes basées sur les SVM, qui donnent généralement de bien meilleurs résultats, Exploitant cette propriété, nous présentons dans cette thèse la version GMM-SVM combinant la méthode discriminante et générative. Les résultats que nous avons obtenus montrent que les performances du système hybride SVM-GMM sont meilleurs que celles du système de référence basé sur la technique GMM-UBM avec normalisation des scores.

iii) La contribution originale de ce travail repose sur la fusion des scores normalisés basée sur la régression robuste en vue d'améliorer les performances de vérification automatique du locuteur distribué. L'approche proposée s'appuie, dans un premier temps, sur la mesure de performance des scores issus des approches génératives et discriminantes étudiées dans le cadre de cette thèse, puis dans un deuxième temps à normaliser et fusionner les scores en exploitant les propriétés de la régression robuste. La normalisation Min-max des scores est une méthode performante et simple à mettre en œuvre. Son apport est significatif comme il est montré dans les expériences réalisées le long de ce travail.

iv) L'objectif principal est d'améliorer les performances de la reconnaissance du locuteur à travers la VoIP. Ainsi, 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 la plateforme ALIZE. 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 ; i) le G.729 décodeur va reconstruire le signal de parole (resynthesized speech) puis vérification de locuteur, ii) la vérification sera effectuée à l'aide de paramètres directement extraits du bit-stream faisant l'économie d'une reconstruction du signal vocal (reconnaissance dans le domaine compressé).

Dans cette thèse, nous avons adressé quelques axes importants du domaine de la reconnaissance du locuteur avec la VoIP, à savoir : Exploitation du bit-stream issu du G.729, investigation du modèle hybride GMM-SVM, la normalisation et la fusion des scores. Néanmoins, il reste d'autres axes qui peuvent faire l'objet de nos futurs travaux, parmi lesquels :

- Le premier objectif de nos 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 ;
- Etendre les travaux à la fusion des paramètres;
- Utilisation d'autre type de codecs de parole tel que le G.722.

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] S. Voran, 1997. *Listener rating of speech passbands.* IEEE Workshop on Speech Coding For Telecommunications Processing, pp. 81–82.

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

[8] I. Karlsson, T. Banziger, J. Dankovicova , 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, vol. 64, n°. 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] S. Meignier, 2002. *Indexation en locuteurs de documents sonores: Segmentation d'un document et Appariement d'une collection*. Thèse de doctorat, Université d'Avignon et des Pays de Vaucluse.

[24] J.F. Bonastre, P. Delacourt, C. Fredouille, T. Merlin et C.J. Wellekens, 2000. *A speaker tracking system based on speaker turn detection for NIST evaluations.* International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istamboul (Turquie).

[25] S. Meignier, J.F. Bonastre et I. Magrin-Chagnolleau, 2002. *Speaker utterances tying among speaker segmented audio documents using hierichical classification towards speaker indexing of audio databases*. International Conference on Spoken Language Processing (ICSLP), pp.573–576, Denver (Etats Unis).

[26] 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.

[27] G. Madre, 2004. *Application de la transformée en nombres entiers à l'étude et au développement d'un codeur de parole pour transmission sur réseaux IP*. Thèse doctorat. Université de Bretagne Occidentale.

[28] T. Tremain, 1982. *The government standard linear predictive coding algorithm: LPC-10*. Speech Technology, vol. 1, n°. 2, pp. 40–49.

[29] J. Mariani, 2002. *Analyse, synthese et codage de la parole-traitement automatique de la parole 1*. Livre p.21-64, Hermes Science Publications Paris.

[30] B. S. Atal, 1974. *Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification.* Journal of the Acoustical Society of America (JASA), vol. 55, pp. 1304–1312.

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

[32] L.R. Rabiner et R.W. Schafer, 2007. *Introduction to digital speech processing*. Now Publishers Inc., Hanover, MA, USA.

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

[34] J. D. MARKEL et A. H. Gray JR, 1976. *Linear prediction of speech*. Communication and Cybernetics. Berlin Heidelberg New York : Springer-Verlag.

[35] S. B. Davis et P. Mermelstein, 1980. *Comparison of parametric representations for monosyllabic word recognition in continuous spoken sentences*. IEEE Transactions on Acoustics, Speech and Signal Processing 28, pp. 357–366.

[36] F. Bimbot, J.F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meigner, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, et D. A. Reynolds, 2004. *A tutorial on text-independent speaker verification.* EURASIP Journal on Applied Signal Processing 4, pp. 430–451.

[37] O. Le-Blouch, 2009. *Décodage acoustico-phonétique et applications à l'indexation audio automatique*. Thèse de doctorat. Université de Toulouse.

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

[39] A. V. Oppenheim et R. W. Schafer, 1968. *Homomorphic analysis of speech*. IEEE Transactions on Audio and Electroacoustics, vol. 16, n°. 2, pp. 221–226.

[40] H. Hermansky, B. A. Hanson, et H. Wakita, 1985. *Perceptually based linear Predictive analysis of speech*. IEEE international conference, vol. 10, pp 509–512.

[41] N. Morgan, H. Bourlard, Michael, et H. Hermansky, 1991. *Continuous speech recognition using PLP analysis with multilayer perceptions*. IEEE international conference, pp 49–52.

[42] H. Hermansky, N. Morgan, 1994. *Integrating RASTA-PLP into speech recognition*. IEEE international conference, speech and audio processing, vol. 1, pp 421–424.

[43] H. Hermansky, et N. Morgan, 1994. *RASTA Processing of speech*. IEEE Transaction on speech and audio processing, vol. 2, n° .4. pp. 587–589

[44] S. Furui, 1981. *Cepstral analysis technique for automatic speaker verification*. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29 n°. 2, pp. 254–272.

[45] F. K. P. Soong et A. E. Rosenberg, 1988. *On the use of instantaneous and transitional spectral information in speaker recognition*. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 36, n°. 6, pp. 871–879.

[46] 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, vol. 26, n°. 6, pp. 575–582.

[47] 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), vol. 28, n°. 6, pp. 623–635.

[48] 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).

[49] 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),

[50] F.Bimbot, I.Magrin Chagnolleau et L. Mathan, 1995. *Second-order statistical measures for text-independent speaker identification*. Speech Communication, tome 17(1-2), pp. 177–192.

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

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

[53] J. L. Gauvain et C. H. Lee, 1994. *Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains.* IEEE Transactions on Speech and Audio Processing, vol. 2, n°. 2, pp. 291–298.

[54] 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.

[55] S. E. Frederickson et L. Tarassenko, 1994. *Radial basis functions for speaker identification.* Workshop on Automatic Speaker Recognition, Identification, Verification, pp. 107–110, Martigny (Suisse).

[56] 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).

[57] I. Magrin-Chagnolleau, J. Wilke et F. Bimbot, 1996. *Further investigation on ar-vector models for text-independent speaker identification*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 401–404, Atlanta (Etats Unis).

[58] 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).

[59] 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).

[60] S. Furui, 1978. *Research on Individuality Information in Speech Waves*. Thèse de Doctorat, Université de Tokyo.

[61] 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.

[62] R. Auckenthaler, M. Carey, et H. Lloyd-Thomas, 2000. *Score normalization for text-independent speaker verification system. Digital Signal Processing (DSP)*. A review journal - Special issue on NIST 1999 speaker recognition workshop, vol 10. n°. 1-3, pp. 42–54.

[63] A. Preti, 2008. *Surveillance de réseaux professionnels de communication par la reconnaissance du locuteur*. Thèse doctorat de l'Université d'Avignon.

[64] E. DUBOIS, 2008. *Convergence dans les réseaux satellite*. Thèse de doctorat, Université de Toulouse.

[65] ITU-T. Recommendation, 2004.*General Overview of NGN (Y.2001)*.

[66] ITU-T. Recommendation, 2004. *General Principles and General Reference Model of Next Generation Network (Y.2011)*.

[67] Rapport de l'ETSI-NGN Starter Groupe, compte-rendu de l'assemblée GA38.

[68] ITU-T. Recommendation, 2006. *Functional requirements and architecture of the NGN release 1 (Y.2012)*.

[69] M. A. Chalouf, 2009. *Offre de service dans les réseaux de nouvelle génération : Négociation sécurisée d'un niveau de service de bout en bout couvrant la qualité de service et la sécurité*. Thèse doctorat. Université de Bordeaux I.

[70] N. Fourty, 2008. *Contribution à l'ingénierie du réseau sans fil WiMAX pour des applications audio d'aide au handicap et aux personnes âgées.* Thèse doctorat, Université de Toulouse II.

[71] ITU-T, 1988. *Pulse code modulation (PCM) of voice frequencies.*

[72] ITU-T, 2007. *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code excited linear prediction (CS-ACELP)*.

[73] ITU-T, 2006. *Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s*.

[74] ETSI, 2000. *Full rate speech; Transcoding (GSM 06.10 version 8.1.1)*.

[75] ETSI, 1998. *Half rate speech; Part 2: Half rate speech transcoding (GSM 06.20 version 4.3.1)*.

[76] AMR speech Codec, 2009. *General description (Release 9), 3GPP Technical Specification 3GPP TS 26.071 V9.0.0.*

[77] Internet Low Bit Rate Codec (iLBC), 2004. *Global IP Sound Request for Comments RFC 3951*.

[78] The Speex Codec Manual Version, 2007. *1.2 Beta 3, Xiph.org Foundation Std.*

[79] SILK Speech Codec, 2010. *Skype Technologies S.A. Internet-Draft.*

[80] UIT-T, *H.323,* 2006. *Systemes de communication multimedia en mode paquet,* UIT-T H.323.

[81] T. Guillet, 2010. *Sécurité de la téléphonie sur IP*. Thèse doctorat. Ecole de Télécom Pari-Tech (France).

[82] F. Boulos, 2010. *Transmission d'images et de vidéos sur réseaux à pertes de paquets: mécanismes de protection et optimisation de la qualité perçue*. Thèse doctorat. Université de Nantes.

[83] F. Kurose and K. W. Ross, 2001. *Computer networking A Top-Down Approach*

featuring the Internet. Pearson Addison Wesley. ISBN 0-201-47711-4.

[84] ITU-T, 2003. *One-way transmission time (G.114 Recommendation)*.

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

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

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

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

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

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

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

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

[93] R. Meraihi, 2007. *Gestion de la qualité de service et contrôle de topologie dans les réseaux ad hoc*. Thèse doctorat, Ecole nationale supérieure des télécommunications.

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

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

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

[97] V. Riboud, 2002. *Introduction aux réseaux et programmation en C des protocoles TCP et UDP*. Livre.

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

[99] D. A. Reynolds, 1995. *Speaker identification and verification using gaussian mixture speaker models*. Speech Communication. vol. 17, n°. 1-2, pp. 91–108.

[100] M. J. Carey et E. S. Parris, 1992. *Speaker verification using connected words*. Dans Proceedings of Institute of Acoustics. vol. 14, n°.6, 96–100.

[101] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, et D. A. Reynolds, 2004. *A tutorial on text-independent speaker verification.* EURASIP Journalon Applied Signal Processing 4, pp. 430–451.

[102] V. Wan et W. M. Campbell, 2000. *Support vector machines for speaker verification and identification*. Neural Networks for Signal Processing, vol. 2, pp. 775–784.

[103] C. Burges, 1998. *A Tutorial on support vector machines for pattern recognition*. Data Mining and Knowledge Discovery.

[104] W. M. Campbell, D. E. Sturim, D. E. Sturim, D. A. Reynolds et D. A. Reynolds, 2006. *Support vector machines using GMM supervectors for speaker verification*. Signal Processing Letters, IEEE Transactions , vol. 13, n. 5, pp. 308–311.

[105] D. Matrouf, J. F. Bonastre, C. Fredouille, A. Larcher, S. Mezaache, M. McLaren, et F. Huenupan, 2008. *LIA GMM-SVM system description: NIST SRE08.* NIST Speaker Recognition Evaluation Workshop, Montreal (Canada).

[106] 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, pp. 291–298.

[107] 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.

[108] A. P. Dempster, N. M. Laird, D. B. Rubin, 1977. *Maximum-likelihood from incomplete data via the EM algorithm.* Journal of Acoustical Society of America (JASA), vol. 39, pp. 1-38.

[109] Y. Mami, 2003. *Reconnaissance de locuteurs par localisation dans un espace de locuteur de référence*. Thèse de doctorat, Ecole nationale supérieure des télécommunications paris.

[110] G. McLachlan, 1988. *Mixture Models*. New York : Marcel Dakker.

[111] D.A.Reynolds, et R.C.Rose, 1995. *Robust text idependent speaker identification using gaussian mixture speaker models*. IEEE Transactions on Speech and Audio Processing (SAP), pp.72–83.

[112] J. Holmes et N. Sedgwick, 1986. *Noise compensation for speech recognition using probabilistics models.* IEEE International Conference on Acoustics, Speech, and

Signal Processing, ICASSP.

[113] R. Hathaway, 1985. *A constrained formulation of maximum-likelihood estimation for normal mixture distributions*. Ann. Stat., vol. 13, n. 2, pp. 795-800.

[114] J. F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, et J. Mason, 2008. *ALIZE/SpkDet: a state-of-the-art open source software for speaker recognition*. Odyssey – The Speaker and Language Recognition Workshop.

[115] V. N. Vapnik, 1998. *Statistical Learning Theory*. Wiley.

[116] M. Schmidt, et H. Gish, 1996. *Speaker identification via support vector classifiers*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 105–108.

[117] W. Campbell, 2002. *Generalized linear discriminant sequence kernels for speaker recognition*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. I–161–I–164.

[118] V. Wan et S. Renals, 2002. *Evaluation of kernel methods for speaker verification and identification*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP) , vol. 1, pp. I–669–I–672.

[119] P. Staroniewicz, et W. Majewski, 2004. *SVM based textdependent speaker identification for large set of voices*. European Signal Processing Conference Processing (EUSIPCO), pp.333–336.

[120] W. Campbell, J. Campbell, D. Reynolds, E. Singer et P. Torres-Carrasquillo, 2006. *Support vector machines for speaker and language recognition*. Computer Speech and Language, vol. 20, n°. 2-3, pp. 210–229.

[121] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman et A. Stolcke, 2005. *Modeling prosodic feature sequences for speaker recognition*. Speech Communication, vol. 46, n. 3-4, pp. 455–472.

[122] W. Campbell, J. Campbell, D. Reynolds, D. Jones, et T. Leek, 2004. *Phonetic speaker recognition with support vector machines*. Neural Information Processing Systems 16, pp.1377–1384. .

[123] H. Shimodaira, K. Noma, M. Nakai, et S, Sagayama, 2001. Support vector machine with dynamic time-alignment kernel for speech recognition. EUROSPEECH processing, pp.1841–1844.

[124] V. Wan, et J. Carmichael, 2005. *Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data*. INTERSPEECH processing, pp. 3321–3324.

[125] J. Louradour, 2007. *Noyaux de séquences pour la vérification du locuteur par machines à vecteurs de support*. Thèse doctorat, Université Toulouse 3 Paul

Sabatier.

[126] S. Bengio, et J. Mariethoz, 2001. *Learning the decision function for speaker verification*. ICASSP processing, vol. 1, pp. 425–428.

[127] S. Fine, J. Navratil, et R. Gopinath, 2001. *Enhancing GMM scores using SVM "hints"*. EUROSPEECH processing, pp. 1757–1760.

[128] J. Kharroubi, D. Petrovska-Delacretaz, et G. Chollet. 2001. *Combining GMM's with suport vector machines for text-independent speaker verification*. EUROSPEECH processing, pp. 1761–1764.

[129] Q. Le, and S. Bengio, 2003. *Client dependent GMM-SVM models for speaker verification*. Artificial Neural Networks and Neural Information Processing ICANN/ICONIP, volume 2714 of Lecture Notes in Computer Science, pp. 443–451. Springer Berlin / Heidelberg.

[130] M. Liu, B. Dai, Y. Xie, et Z. Yao, 2006. *Improved GMM-UBM/SVM for speaker verification*. In Proc. of ICASSP, vol. 1, pp. I–925–I–928.

[131] N. Krause et R. Gazit, 2006. *SVM-based speaker classification in the GMM models space*. IEEE Odyssey - Speaker and Language Recognition Workshop processing.

[132] S. Fine, J. Navratil, et R. Gopinath, 2001. *A hybrid GMM-SVM approach to speaker identification*. ICASSP processing, vol. 1, pp. 417–420.

[133] P. Moreno, et P. Ho, 2003. *A new SVM approach to speaker identification and verification using probabilistic distance kernels*. EUROSPEECH processing, vol. 3, pp. 2965–2968.

[134] P. Ho, P. Moreno, 2004. *SVM kernel adaptation in speaker classification and verification*. INTERSPEECH processing, pp. 1413–1416.

[135] N. Dehak, et G. Chollet, 2006. *Support vector GMMs for speaker verification*. IEEE Odyssey - The Speaker and Language Recognition Workshop processing.

[136] V. Wan, et S. Renals, 2005. *Speaker verification using sequence discriminant support vector machines*. IEEE Transactions on Speech and Audio Processing, vol. 13, n. 2, pp. 203–210.

[137] W. Campbell, D. Sturim et D. Reynolds 2006. *Support vector machines using GMM supervectors for speaker verification*. IEEE Signal Processing Letters, vol. 13, n. 5, pp. 308–311.

[138] W. Campbell, D. Sturim, D. Reynolds, et A. Solomonoff, 2006. *Svm based speaker verification using a gmm supervector kernel and nap variability compensation*. ICASSP processing, vol. 1, pp. I–97–I–100.

[139] D. Yessad et A. Amrouche, 2012. *SVM based GMM supervector speaker*

recognition using LP residual signal. 5th International Conference on Image and Signal Processing (ICISP 2012). Agadir, Marocco. LNCS 7340, ISSN : 0302-9743 Springer. pp. 579-586.

[140] M. Bin, H. Meng, et M. Man-Wai, 2007. *Effects of device mismatch, language mismatch and environmental mismatch on speaker verification*. International Conference on Acoustics, Speech, and Signal Processing (ICASSP),

[141] R. Auckenthaler et J. S. Mason, 2000. *Score normalisation for text-independent speaker verification systems*, Digital Signal Processing Journal.

[142] R. Auckenthaler, M. Carey et H. Lloyd-Thomas, 2000. *Score Normalization for Text-Independent Speaker Verification System*. Digital Signal Processing, vol. 1, n°.10, pp. 42–54.

[143] A. Higgins, L. Bahler et J. Porter, 1991. *Speaker verification using randomized phrase prompting. Dans les actes de Digital Signal Processing,* vol. 1, pp. 89–106.

[144] K.-P. Li et J. E. Porter, 1998. *Normalizations and selection of speech segments for speaker recognition scoring*. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New York (USA), vol. 1, pp. 595–598.

[145] [T. Matsui et S. Furui, 1993. *Concatenated phoneme models for text-variable speaker recognition*. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Minneapolis (USA), vol. 2, 391–394.

[146] D. A. Reynolds, 1996. *The effects of handset variability on speaker recognition performance : experiments on the Switchboard corpus*. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1.

[147] C. Fredouille, J.-F. Bonastre, et T. Merlin, 1999. *Similarity normalization method based on world model and a posteriori probability for speaker verification*. European Conference on Speech Communication and Technology (Eurospeech), Budapest (Hungary), vol. 2, pp.983–986.

[148] A. Rosenberg, 1992. *The use of cohort normalized scores for speaker verification*. Second International Conference on Spoken Language Processing (ISCLP).

[149] L. Kuncheva, 2002. *A theoretical study on six classifier fusion strategies*.IEEE Transaction. Pattern Anal. Mach. Intell., vol. 24, n. 2 pp. 281-286.

[150] S. Y. Sohn et H. Shin, 2007. *Experimental study for the comparison of classifier combination methods*. Pattern Recognition, vol. 40, n. 1, pp. 33-40.

[151] J. Kittler, M. Hatef, R. P. W. Duin et J. Matas, 1998. *On combining classifiers*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, n. 3, pp. 226-239.

[152] R. O. Duda, P. E. Hart, et D. G. Stork, 2000. *Pattern Classification (2nd Edition)*. Wiley-Interscience.

[153] K. Chen, L. Wang, et H. Chi, 1997. *Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification*. International Journal of Pattern Recognition and Artificial Intelligence, vol. 11, n. 3, pp. 417–446.

[154] C. Fredouille, J.F. Bonastre, et T. Merlin, 2000. *AMIRAL: a block-segmental multirecognizer architecture for automatic speaker recognition*. Digital Signal Processing, vol. 10, n. 1-3, pp. 172–197.

[155] R. Ramachandran, K. Farrell, R. Ramachandran, et R. Mammone, 2002. *Speaker recognition–general classifier approaches and data fusion methods*. Pattern Recognition, vol 35, n. 12, pp.2801–2821.

[156] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, et J. Ortega-Garcia, 2003. *Support vector machine fusion of idiolectal and acoustic speaker information in Spanish conversational speech*. International Conference on Multimedia and Expo Processing, vol. 3, pp. 205–208.

[157] J. Campbell, D. Reynolds, et R. Dunn, 2003. *Fusing high- and low-level features for speaker recognition*. EUROSPEECH Processing, pp. 2665–2668.

[158] T. Kinnunen, V. Hautamaki, et P. Franti, 2004. *Fusion of spectral feature sets for accurate speaker identification*. SPECOM Processing, pp. 361–365.

[159] W. Campbell, D. Reynolds, et J. Campbell, 2004b. *Fusing discriminative and generative methods for speaker recognition: experiments on switchboard and NFI/TNO field data*. Odyssey Processing - The Speaker and Language Recognition Workshop, pp. 41–44.

[160] N. Scheffer, et J. F. Bonastre, 2006. *Fusing generative and discriminative UBM-based systems for speaker verification*. International workshop on MMUA (MultiModal User Authentification) Processing.

[161] D. Mashao, et M. Skosan, 2006. *Combining classifier decisions for robust speaker identification*. Pattern Recognition, vol. 39, n. 1, pp. 147–155.

[162] F. Huenupan, N. Yoma, C. Molina, C. Garreton, 2007. *Speaker verification with multiple classifier fusion using Bayes based confidence measure*. INTERSPEECH Processing, pp.2041–2044.

[163] Y. Solewicz et M. Koppel, 2007. *Using post-classifiers to enhance fusion of low- and high-level speaker recognition*. IEEE Transactions on Audio, Speech and Language Processing, vol. 15, n. 7, pp. 2063–2071.

[164] L. Ferrer, M. Graciarena, A. Zymnis et E. Shriberg, 2008. *System combination using auxiliary information for speaker verification*. Acoustics, Speech and Signal Processing (ICASSP), pp. 4853–4856.

[165] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones et B. Xiang, 2003.

The SuperSID project : Exploiting high-level information for high-accuracy speaker recognition. Acoustics, Speech and Signal Processing (ICASSP), vol. 4, pp. IV–784–IV–787.

[166] D. Yessad, et A. Amrouche, 2013. Fusion Strategies Distributed speaker recognition using residual signal based G729 resynthesized speech. 16Th International Conference on Information Fusion (ICIF 2013). 9-12 July 2013, Askeri Museum-Istanbul Turkey, pp. 432–437.

[167] N. Brummer, et J. du Preez, 2006. *Application-independent evaluation of speaker detection*. Computer Speech and Language, vol. 20, n. 2-3, pp. 230–275.

[168] A. Jain, K. Nandakumar, A. Ross, 2005. Score normalization in multimodal biometric systems. Pattern Recognition, vol. 38, pp. 2270–2285.

[169] 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.

[170] W. H. DuMouchel, F. L. O'Brien, 1989. *Integrating a Robust Option into a Multiple Regression Computing Environment*. Computer Science and Statistics: Proceedings of the 21st Symposium on the Interface, Alexandria, VA, American Statistical Association.

[171] P. J. Huber, 1964. *Robust estimation of a location parameter*. Annals of Mathematical Statistics, pp. 3–73.

[172] P. J. Huber, 1981. *Robust Statistics.*Wiley.

[173] J. O. Street, R. J. Carroll, D. Ruppert, 1988. *A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares*. The American Statistician, vol. 42, n. 2, pp. 152–154.

[174] A. Amrouche, M. Debyeche, A. Taleb Ahmed, J. M. Rouvaen, M. C. E Yagoub, 2010. Efficient system for speech recognition in adverse conditions using nonparametric regression. Engineering applications on artificial intelligence, vol 23, n°1, Elsevier, pp. 85–94.

[175] D. Yessad, A. Amrouche, M. Debyeche, 2011. *Influence of G729 speech coding on automatic speaker recognition in VoIP applications*. The 2011 International Workshop on Computing and Communications (CC-11). Jeju, Korea. In conjunction with CSA 2011, ISSN: 1876-1100, pp.745–751.

MES PUBLICATIONS

Contribution de l'auteur

Publication internationale

1) Yessad D., and Amrouche, A. (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. http://link.springer.com/article/10.1007/s10772-013-9204-6

Proceeding édités (avec comité de lecture et ISSN)

1) Yessad D., and Amrouche, A. (2012).SVM based GMM supervector speaker recognition using LP residual signal, 5th International Conference on Image and Signal Processing (ICISP 2012). 28-30 June 2012, Agadir, Marocco. LNCS 7340, ISSN : 0302-9743 Springer. pp. 579-586.
2) Yessad, D., Amrouche, A., Debyeche, M. (2011). Influence of G729 speech coding on automatic speaker recognition in VoIP applications. The 2011 International Workshop on Computing and Communications (CC-11). 12-15 December, Jeju, Korea. In conjunction with CSA 2011, ISSN: 1876-1100, pp.745-751.
3) Yessad D., Amrouche, A., Debyeche, M. (2011). Micro-Doppler classification for ground surveillance radar using speech recognition tools. 16th Ibero American Conference on Pattern Recognition CIARP'11. 15-18 November. Pucón, Chile. In Lectures Notes on Computer Sciences, Cesar San Martin and S.-W. Kim (Eds.). LNCS 7042, ISSN: 0302-9743, pp. 280-287.
4) Yessad D., Amrouche, A., Debyeche, M. (2011). SVM and Greedy GMM applied on target identification. International Conference on Neural Information Processing, 13-17 November, Shanghai, China. In Lectures Notes on Computer Sciences. B.-L. L. Lu, L. Zhang, and Kwok (Eds.): ICONIP 2011, PartII, LNCS 7063, pp. 292-299.

Conférences internationales avec comité de lecture et proceeding

1) Yessad, D., and Amrouche, A. (2013). Fusion Strategies Distributed speaker recognition using residual signal based G729 resynthesized speech. 16Th International Conference on Information Fusion (ICIF 2013). 9-12 July 2013, Askeri Museum-Istanbul Turkey, pp. 432-437.
2) Yessad D., and Amrouche, A. (2013). Automatic speaker recognition using G729 resyntesized speech over IP. International Conference on Tecommunication and Signal, Image, Vision and their applications (SIVA'2013), 18-20 November 2013, Guelma, Algeria.
3) Yessad D., and Amrouche, A. (2012). G729 Coded parameters under matched and mismatched conditions for distributed speaker recognition, International Congress on Telecommunication and Application'12 University of A.MIRA (ICTAI12), 11-12 April 2012, Bejaia, Algeria.

ANNEXES

Annexe A

G.729 codec de la VoIP

Cette annexe présente les critères relatifs au codage audio, puis expose les différentes approches exploitées en codage de la parole, enfin étudie le plus clairement possible les fonctions principales et les algorithmes qui permettent l'implémentation du G.729

A.1 Critères liés au codage

Le codage et la compression des signaux audio ont connu un développement considérable pendant cette dernière décennie. Le choisie d'un codec est en relation directe avec la compression du débit binaire nécessaire pour transmettre le signal audio et l'optimisation de la qualité perçue du signal restitué par le décodeur. Les différents codecs de parole sont caractérisés par différents paramètres comme la largeur de la bande passante, le débit binaire, la qualité de la parole restituée, le retard et la complexité de codage et finalement les standards.

A.1.1 Largeur de bande passante

La largeur de la bande passante (bande fréquentielle) et la dynamique du signal audio représentent la gamme de qualité à laquelle en destinant les codecs qui peuvent être classés suivant cette gamme en quatre catégories: Les codecs à bande étroite (NB: Narrow Band) ; à large bande (WB: Wide Band) ; à bonde radio FM et à bande HI-FI qualité CD ROM. Le tableau A.1 résume les caractéristiques principales des différentes catégories des codages audio.

	Bande passante	Fréquence d'échantillonnage
Bande téléphonique	300-3400 Hz	8 kHz
Bande élargie	50-7000 Hz	16 kHz
Bande FM	20-1500 Hz	32 kHz
Bande HI-FI	20-2000 Hz	44.1 kHz

Tableau A.1: Gamme de qualité des codecs audio.

A.1.2 Débit de transmission

Le débit de transmission (débit binaire) est la composante la plus importante puisqu'il détermine le nombre de bits par seconde nécessaire pour la représentation de l'information, ainsi le taux de compression fourni par l'algorithme de codage. Le débit de transmission est également fonction de la gamme de qualité considérée puisque celle-ci détermine la fréquence d'échantillonnage. L'objectif d'un algorithme de codage est de réduire ce débit en maintenant la bonne qualité du signal. La largeur de bande disponible dans le système de communication détermine la limite supérieure du débit envisageable du codeur de parole. Lors de la conception d'un système de codage, un choix sera effectué parmi les codeurs à débit de transmission fixe ou variable.

A.1.3 Qualité de parole

La qualité du signal restitué est une considération importante dans tout codage de parole. Les recherches sur les différents types de codage essaient toujours de trouver un bon compromis entre la qualité du signal synthétisé et le débit de transmission. Pour un débit fixé, la qualité de la parole restituée après décodage est évaluée par deux types de mesures, objective et subjective :

- Mesure subjective: appelée MOS (Mean Opinion Score). Il s'agit des tests d'écoute du signal de la parole reconstruit dans les conditions désirées, ces tests sont évalués par des auditeurs qui jugeront subjectivement la qualité globale et l'intelligibilité de la parole. Pour ce genre d'étude, un grand nombre de personnes est nécessaire pour effectuer une analyse statistique de leur opinion moyenne MOS. Cette analyse est basée sur une échelle numérique allant de 5 à 1 pour définir la qualité de la parole correspondant à: excellent, bon, moyen, médiocre et mauvais. La qualité standard téléphonique correspond au niveau 4 du MOS.
- Mesure objective : Cette mesure exploite des fonctions ou des critères mathématiques pour comparer la parole restituée et originale telle que le rapport signal sur bruit (SNR : Signal to Noise Ratio).

A.1.4 Délai de codage

Le délai de codage (retard de codage) est très important dans les transmissions en temps réel. Le délai global est engendré par le temps de traitement du codage et décodage mais aussi par le délai de transmission qui correspond au temps de la traversé du canal et aux différents temps d'attente liés à l'application choisie.

A.1.5 Complexité de calcul et d'implémentation

Baisser le débit de transmission en maintenant une bonne qualité du signal se fait généralement au détriment de la complexité des algorithmes mis en place. Pour une implantation temps réel exigée sur des processeurs numériques du signal (DSP : Digital Signal Processor), les algorithmes ne doivent pas être trop complexes et leurs exigences ne pas dépasser les capacités des processeurs DSP, limitées au niveau de leur mémoire RAM et de leur vitesse MIPS. La complexité algorithmique est souvent évaluée en termes de millions d'instructions par secondes (MIPS: Million d'Instructions par Seconde) et par la taille de mémoire vive (RAM: Random Access Memory) et de mémoire morte (ROM: Read Only Memory) utilisées.

A.1.6 Standardisation

Parmi les organismes internationaux qui produisent les normes pour les codeurs de parole on peut citer :

- l'ITU International Telecommunication Union pour la téléphonie sur les réseaux fixes et paquets, les communications personnelles ou les communications avec les mobiles par satellites.
- l'ETSI European Telecommunication Standards Institute ou le TIA Telecommunication Industry Association pour la téléphonie mobile cellulaire grand-public ou privée respectivement en Europe et aux États-Unis,
- l'OTAN Organisation du Traité de l'Atlantique Nord (en anglais NATO North Atlantique Treaty Organization) ou le DoD Department of Defense aux État-Unis pour les applications militaires ou les applications sécuritaires ;
- INMARSAT International Maritime Satellite Corporation ou Auptus/Aussat Australien Satellite pour la téléphonie mobile maritime par satellite.

Par ailleurs, l'ISO/IEC International Organization for Standardization / International Electrotechnical Commission normalise des codeurs pour la compression des signaux audio

pouvant intégrer des codeurs de parole normalisés et non normalisés, comme Internet Low Bit rate Coder (iLBC) ou speekx utilisés dans certaines applications voix sur IP grand public.

A.2 Différentes approches en codage de parole

Les méthodes de codage de la parole sont nombreuses, en distinguant généralement trois grandes classes : les codeurs de forme d'onde qui sont surtout performants pour des débits élevés (au-delà des 16 kbit/s), les codeurs paramétriques, appelés vocodeurs sont plutôt destinés aux très bas débits (au-dessous de 2.4 kbit/s) et aux bas débits (2.4 à 8 kbit/s), et les codeurs hybrides avec des débits nominaux intermédiaires (8 à 16 kbit/s), bien qu'ils puissent aussi être utilisés pour de bas débits. Le choix d'une méthode va dépendre surtout de l'application visée et des contraintes sur le débit. La majorité des codeurs utilisés dans les réseaux de communications vocales sont à bande étroite. La bande fréquentielle de la parole codée est limitée entre 300 et 3400 Hz et la fréquence d'échantillonnage du signal d'entrée est alors à 8 KHz.

Dans ce travail on s'intéresse aux codeurs hybrides qui sont pour la plupart dérivés de l'algorithme de codage CELP (Code Excited Linear Prediction) introduit par Schroeder et Atal. Cet algorithme effectue une quantification vectorielle prédictive du signal temporel et utilise une approche d'analyse par synthèse. Il s'appuie sur une prédiction linéaire à court terme pour le codage de l'enveloppe spectrale et sur une prédiction linéaire à long terme pour la modélisation des excitations voisées. Le CELP et ses dérivés sont à la base de la majorité des standards de codeurs de parole utilisés dans les réseaux de téléphonie mobiles et les communications vocales sur Internet en voix sur IP. Les codeurs qui utilisent l'algorithme CELP délivrent un débit moyen compris typiquement entre 4 et 16 Kb/s, avec une bonne qualité de codage allant de 3 à 4 en terme de note MOS. Leur complexité est élevée comparée aux codeurs de forme d'onde.

Récemment, les codeurs CELP ont suscité beaucoup d'attention et servent de base à la plupart des algorithmes de codage de la parole. Les premiers travaux ont appris des codes CELP sur un algorithme spécial ACELP (Algebraic Code-Excited Linear-Prediction), en utilisant un dictionnaire stochastique algébrique. Le VSELP (Vector Sum Excited Linear Prediction) proposé par Motorola et utilisé par le codeur HR Half Rate du GSM. Un codeur LD-CELP (Low Delay-CELP) ciblant un faible délai de codage/décodage a été donné par la recommandation G.728 de l'ITU. Le codeur G.729, basé sur l'algorithme CS-ACELP

(Conjugate Structure Adaptive Code Excited Linear Predictive) à 8 kbits/s, a été normalisé. CS-ACELP utilisant un dictionnaire algébrique a été proposé.

Par la suite, un codeur ACELP particulier sera au centre de notre intérêt, il s'agit du codec CS-ACELP correspondant au G.729 normalisé par l'UIT-T en 1996, constitue de bon exemple pour un fonctionnement dans un environnement avec les contraintes de bandes passantes, pertes, faible délai et complexité. Il est recommandé dans la norme H.323 pour la téléphonie sur IP.

A.3 Codec G.729 (CS-ACELP)

Le G.729 fait partie de la famille ACELP, il se fait en deux algorithmes: l'encodeur et le décodeur. Au niveau d'encodeur, le signal d'entrée est prétraité puis analysé dans le but d'extraire les paramètres du modèle de prédiction CELP qui sont: Les coefficients du filtre de prédiction linéaire, le pitch et le code d'excitation. Ces coefficients une fois extraites sont transmis comme bit-stream au décodeur. Au niveau du décodeur les paramètres des flux binaire servent à récupérer les coefficients d'excitation et du filtre de synthèse. La filtration du flux d'excitation dans le filtre de synthèse à court terme donne en sortie le signal reconstitué appelé aussi signal synthétisé au bien signal transcodé G.729.

A.3.1 Codeur G.729

Le codeur G.729 opère sur des trames de 10 ms que l'on découpe ensuite en deux sous-trames de 5 ms. Une trame future de 5ms est utilisée lors de l'analyse LP. L'algorithme d'encodeur peut être décomposé en trois parties: la prédiction linéaire, l'analyse du pitch et la recherche de code d'excitation, comme le montre la figure A.1

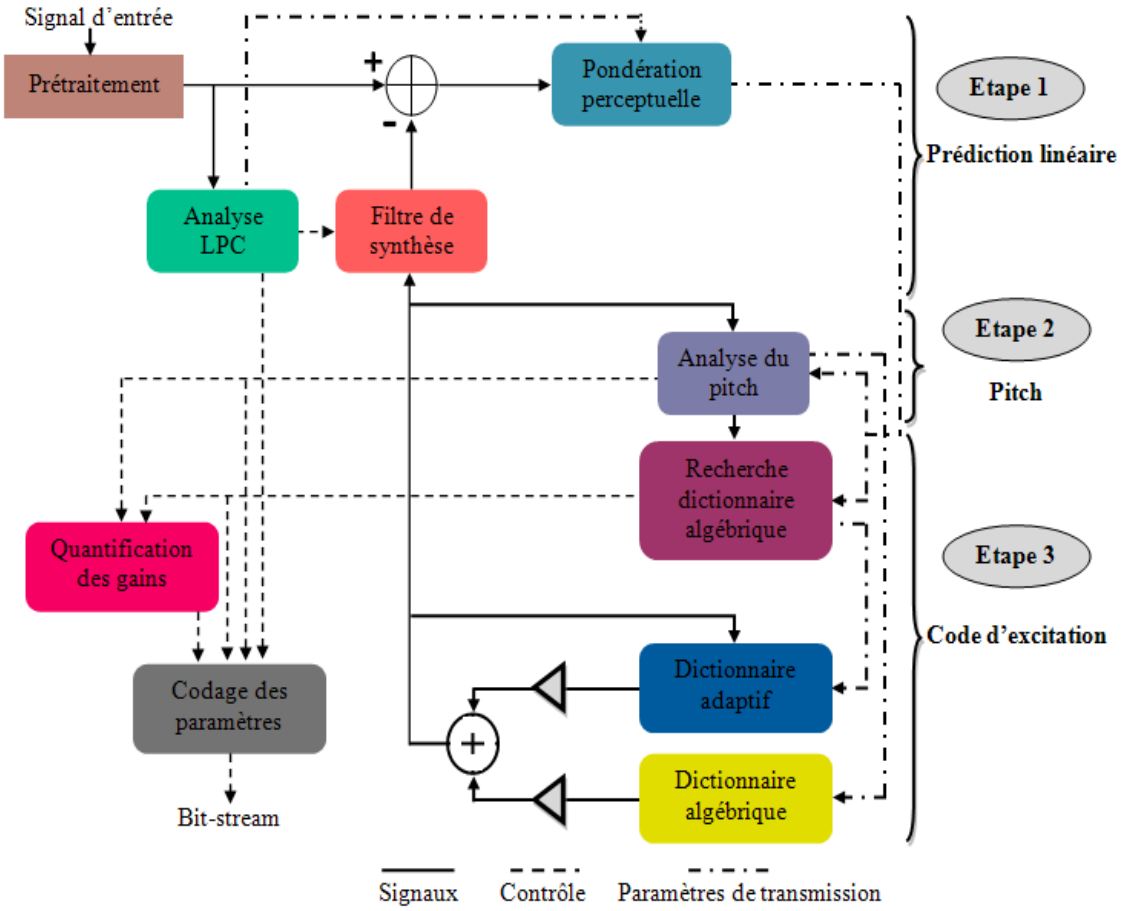


Figure A.1: Diagramme de codeur G.729.

A.3.1.1 Prédiction linéaire (LP)

Dans la recommandation G.729, le codeur est conçu pour fonctionner avec un signal échantillonné à 8000 Hz et quantifié sur 16 bits. Le signal de parole d'entrée, échantillonné et quantifié, subit une normalisation dans un filtre passe-haut de prétraitement H(z), dont la fréquence de coupure du filtre est égale à 140 Hz.

$$H(z) = \frac{0.46363718 - 0.92724705\,z^{-1} + 0.46363718\,z^{-2}}{1 - 1.9059465\,z^{-1} + 0.9114024\,z^{-2}} \qquad \text{(A.1)}$$

Le signal en sortie du filtre de prétraitement noté $s(n)$, sera utilisé dans toutes les opérations ultérieures du codeur. Le calcul des coefficients de prédiction est effectué toutes les 10 ms.

Le G.729 détermine les paramètres de prédiction linéaire à court terme par un filtre tout pôle d'ordre 10 basé sur l'analyse LPC (Linear Predictive Coding). Le calcul des coefficients de prédiction linéaire (LP) est effectué toutes les 10 ms. Avant d'effectuer l'analyse de prédiction linéaire à court terme, le signal discret $s(n)$ est fenêtré par une fenêtre d'analyse de prédiction linéaire (ω_{lp}) comme le montre la figure ci-dessous.

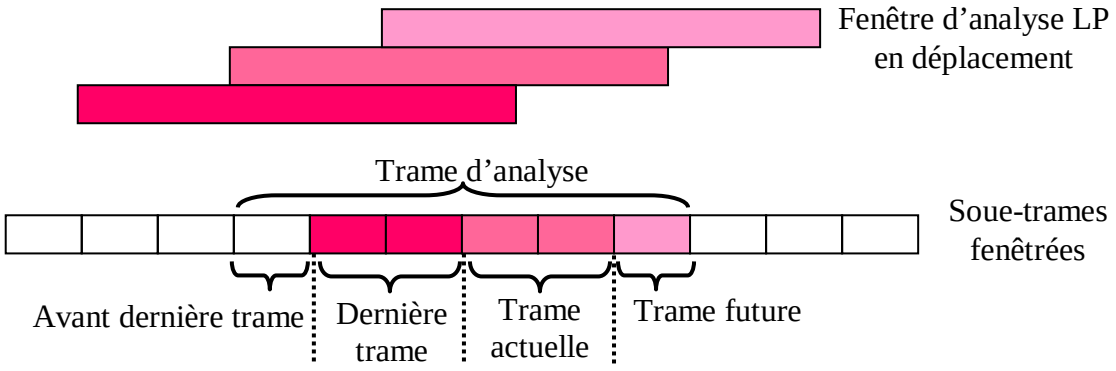


Figure A.2: Procédure de fenêtrage en analyse LP.

La fenêtre (ω_{lp}) de largeur $N_\omega = 240$ échantillons (comprenant à 30 ms), s'applique aux 120 échantillons (15 ms) issus des deux trames vocales passées, aux 80 échantillons (10 ms) issus de la trame vocale courante et aux 40 échantillons (5 ms) de la trame vocale future, cette fenêtre d'analyse LP est représentée par l'équation suivante:

$$\omega_{lp}(n) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{399}\right) & n = 0, \dots, 199 \\ \cos\left(\frac{2\pi(n-200)}{159}\right) & n = 200, \dots, 239 \end{cases} \quad \text{(A.2)}$$

La figure A.3 montre que la fenêtre d'analyse LP (ω_{lp}) du codeur G.729, est asymétrique et se décompose en deux parties; la première partie est une demi-fenêtre de Hamming et la deuxième est un quart de période d'une fonction Cosinus.

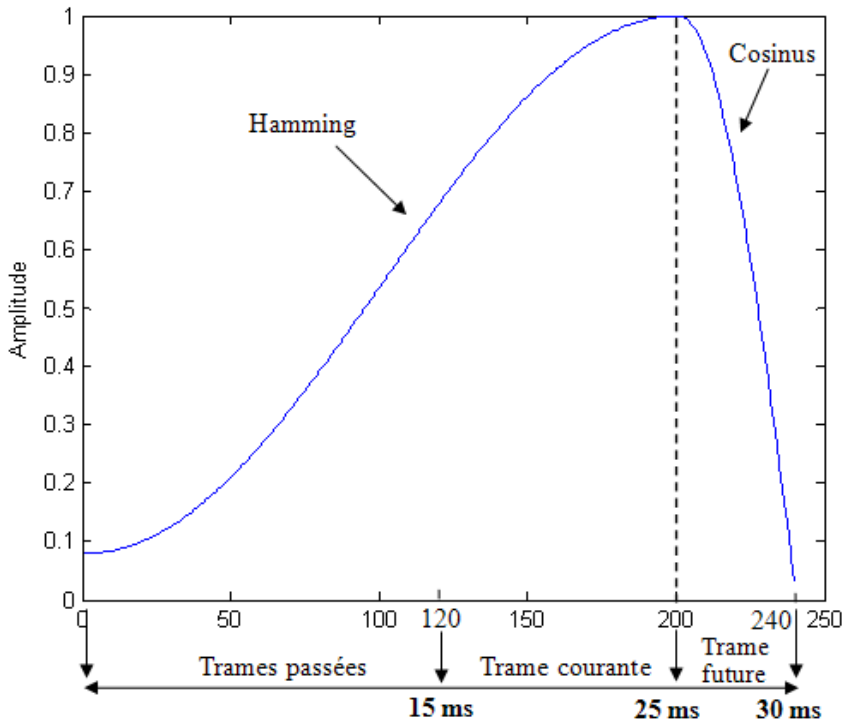


Figure A.3: Fenêtre d'analyse LP.

L'analyse LP comporte un déplacement de 5 ms, c'est-à-dire qu'il faut 40 échantillons issus de la prochaine trame vocale. Cela se traduit par un délai algorithmique supplémentaire de 5 ms au niveau du codeur G.729.

Le signal vocal issu de la fenêtre d'analyse (ω_{lp}) est:

$$s(n)' = \omega_{lp}(n)s(n) \qquad n = 0 \cdots 239 \tag{A.3}$$

Ce signal une fois obtenu il est exploité pour calculer les coefficients d'autocorrélation ainsi pour extraire les coefficients $\{a_k\}$, $k = 0, \ldots, p$ de filtre de Prédiction Linéaire (LP) du dixième ordre ($p = 10$) en utilisant l'algorithme de Levinson-Durbin. Les paramètres du filtre sont ensuite convertis en lignes de raies spectrales (LSP) aux fins de la quantification et de l'interpolation.

A.3.1.2 Code du signal d'excitation

Les paramètres d'excitation (par dictionnaire codé fixe et par dictionnaire codé adaptatif) sont déterminés à chaque sous-trame de 5 ms (40 échantillons), à travers le choix du signal d'excitation au moyen d'une procédure de recherche par analyse et synthèse dans

laquelle l'erreur entre le signal vocal original et le signal vocal reconstitué est minimisée en fonction d'une mesure de distorsion pondérée par la perception. A cette fin, le signal d'erreur passe par un filtre de pondération perceptive dont les coefficients sont déduits du filtre de prédiction linéaire avant quantification. Les coefficients du filtre de prédiction linéaire, quantifiés et non quantifiés, sont utilisés pour la deuxième sous-trame, alors que la première utilise une interpolation des coefficients (quantifiés que non quantifiés) du filtre de prédiction linéaire.

Dans le cas les codeurs CELP, pour déterminer l'excitation $u(n)$ une recherche exhaustive dans les dictionnaires adaptatif et stochastique est effectuée. $u(n)$ est une combinaison linéaire de la participation des dictionnaires :

$$u(n) = \beta v(n) + Gc(n), \qquad avec \quad n = 0, \dots, L-1 \tag{A.4}$$

Où L est le nombre d'échantillons contenus dans une sous trame, les formes d'onde $\{c, v\}$ et les gains associés $\{G, \beta\}$ des dictionnaires fixe et adaptatif représentent les paramètres d'excitation possible qui cherchent à minimiser l'erreur quadratique moyenne entre les sous trames de parole originales et celles reconstruites par le filtre de synthèse LP (figure A.4). Le dictionnaire adaptatif tend à fournir une bonne approximation de la forme d'onde d'excitation et plus particulièrement pour les segments continus de parole voisée. Alors que le dictionnaire stochastique a pour objectif de modéliser les composants aléatoires restants et les segments non-voisés du signal d'excitation.

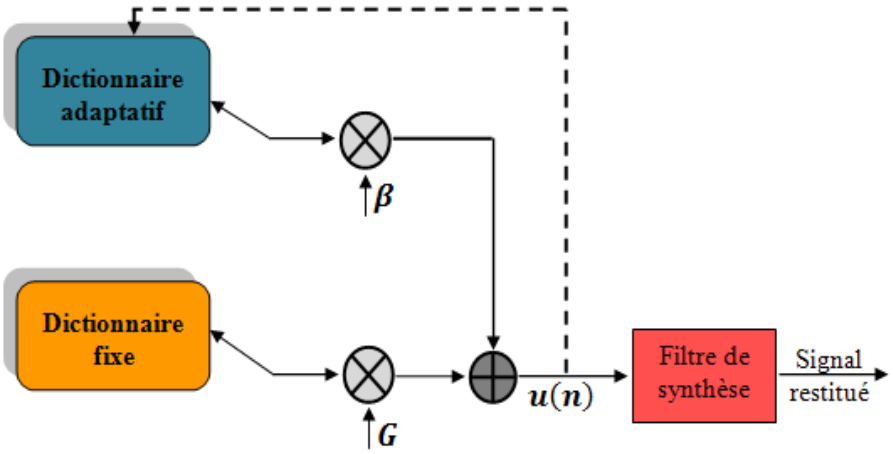


Figure A.4 : Synthèse de codeur CELP avec dictionnaires fixe et adaptatif.

Le codeur G.729 (CS-ACELP) comporte deux dictionnaires. Chacun de ces dictionnaires a une taille de 40 échantillons soit 5 ms.

- Un dictionnaire adaptatif: Le mot choisit est une portion du passé du signal de synthèse, le décalage étant égal au délai de pitch. Ce dictionnaire représente la contribution du pitch
- Un dictionnaire fixe et algébrique: codage de l'excitation à bas débit. On utilise un dictionnaire ISPP (Interleaved Single-Pulse Permutations) à 4 pulses. Chaque sous-trame de 10 échantillons est décomposée en 5 tracks numérotées de 0 à 4 comme le montre la figure **A.5**. On positionne un pulse dans chacune des track 0 à 2 et un pulse dans une des deux track 3 et 4. Les modules des pulses sont indépendants et peuvent prendre comme valeurs +1 ou -1.

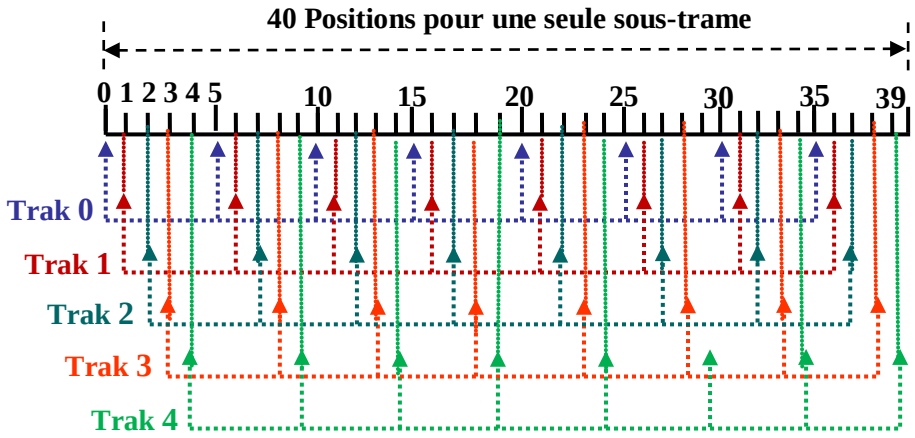


Figure A.5: Positions des pulses pour dictionnaire fixe.

A.3.1.3 L'analyse de pitch

La recherche du pitch est faite de façon très minutieuse, en deux étapes:

- Détermination du pitch en boucle ouverte: Cette recherche vise à donner une première approximation du pitch. Elle se fait en boucle ouverte sur la trame courante, pondérée par un filtre perceptuel. La valeur cherchée est le décalage qui maximise la corrélation entre le signal et ses versions décalées.
- Détermination du pitch en boucle fermée: On établit pour chaque sous-trame (5ms) une recherche autour de l'évaluation en boucle ouverte. On détermine la nouvelle valeur du délai en minimisant l'erreur quadratique dans un domaine perceptuel entre le

signal cible (calculé par filtrage de l'énergie résiduelle du codage prédictif linéaire dans le filtre de synthèse pondérée) et le signal reconstruit à partir du passé.

A.3.2 Paramètres transmis comme bit-stream

Pour chaque trame de 80 échantillons le codeur transmet 80 bits. Le bit-stream (train de bits) comprend l'affectation des positions binaires aux paramètres suivants: coefficients lignes de raies spectrales (LSP) numérisés sur 18 éléments, délai de pitch, code d'excitation ainsi que les gains pour les dictionnaires. L'affectation des bits dans une trame issue du codeur CS-ACELP sous forme du bit-stream est donnée par le tableau ci-dessous.

Paramètre	Mot de code	Sous-trame 1	Sous-trame 2	Total par trame
LSP	L0, L1, L2, L3			18
Délai de pitch	P1, P2	8	5	13
Délai de pitch: Parité	P0	1		1
Index du dictionnaire fixe	C1, C2	13	13	26
Signe du dictionnaire fixe	S1, S2	4	4	8
Gaine de dictionnaire (étap1)	GA1, GA2	3	3	6
Gaine de dictionnaire (étap2)	GB1, GB2	4	4	8
Total				80

Tableau A.2: Trame de 10 ms du bit-stream transmis par le codeur G.729.

A.3.3 Décodeur G.729

Le diagramme du décodeur décrit par la figure A.6 représente la procédure de décompression appliquée sur une trame vocale de 10 ms du bit-stream transmis par le codeur. Pour décoder le flux binaire on commence d'abord par la décompression des indices *L*0, *L*1, *L*2 et *L*3 qui servent à reconstruire les coefficients LSP, puis on récupère le code vectoriel du dictionnaire adaptatif, code vectoriel du dictionnaire fixe et les gains.

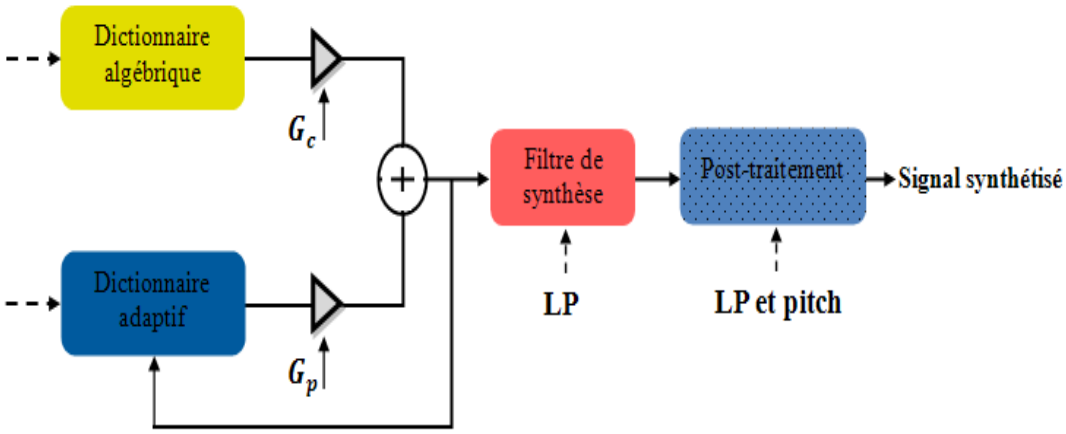


Figure A.6: Diagramme de décodeur G.729.

Donc les paramètres du bit-stream une fois décodés, sont exploités pour récupérer les paramètres d'excitation et du filtre de synthèse. Le flux d'excitation est filtré dans le filtre de synthèse à court terme. Ce filtre fait appel à une prédiction linéaire (LP) du dixième ordre. Ainsi les coefficients LP sont utilisés pour synthétiser le signal vocal dans la sous-trame. Le filtre à long terme ou de synthèse tonale, est mis en œuvre par le dictionnaire codé adaptatif. Les étapes suivantes sont répétées pour chaque sous-trame:

- Décodage du code vectoriel de dictionnaire adaptatif;
- Décodage du code vectoriel de dictionnaire fixe;
- Décodage des gains par dictionnaire adaptatif et par dictionnaire fixe;
- Calcul du signal vocal reconstitué ;
- Amélioration du signal vocal reconstitué par une opération de post-traitement, qui comprend un post-filtre adaptatif utilisant le montage en cascade de trois filtres (filtre de compensation et la sortie des deux filtres de synthèse à court et à long terme), suivi d'un filtre passe-haut et d'un échantillonneur-normalisateur.