Indexation de documents audio : Cas des grands volumes de données

Jamal Rougui

UNIVERSITE DE NANTES

ÉCOLE DOCTORALE

SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DES MATERIAUX

Année : 2008

Thèse de Doctorat de l'Université de Nantes

Spécialité : INFORMATIQUE

Présentée et soutenue publiquement par

Jamal-Eddine ROUGUI

le 16 Juillet 2008

à l'Université Mohammed V-Agdal, Faculté des Sciences, Rabat, Maroc

TITRE
Indexation de documents audio : Cas des grands volumes de données

Jury

Président	:	D. Aboutajdine	Professeur à la faculté des sciences Rabat.
Rapporteurs	:	EL. Mouaddib	Professeur à l'Université de Picardie Jules verne, Amiens.
		D. Mammass	Professeur à Faculté des Sciences Ibn zouhr, Agadir.
Examinateurs	:	N. Mouaddib	Professeur à l'Université de Nantes.
		K. Daoudi,	Chargé de recherche 1ère classe au CNRS, Toulouse.
		M.B García	Professeur à l'Université de Deusto, Bilbao, Espagne.
		M. Gelgon	Maître de conférences HDR à l'Université de Nantes.
		M. Rziza	Professeur Assistant à la Faculté des Sciences, Rabat.

Directeur de Thèse : Noureddine Mouaddib
Laboratoire : LINA (UMR CNRS 6241), équipe ATLAS-GRIM
Co-encadrant : Marc Gelgon
Laboratoire : LINA (UMR CNRS 6241), équipe ATLAS-GRIM
Composante de rattachement du directeur de thèse :

N° ED 0366-372

Cotutelle de thèse dans le cadre de la coopération Franco-Marocaine dans le domaine des STIC entre la Faculté ds Sciences de l'Université Mohammed V Rabat-Agdal et L'Ecole Polytechnique de l'Université de Nantes.

— Jamal-Eddine ROUGUI, Mémoire de thèse.

Résumé

Cette thèse est consacrée à l'élaboration et l'évaluation des techniques visant à renforcer la robustesse des systèmes d'indexation de documents audio au sens du locuteur. L'indexation audio au sens du locuteur consiste à reconnaître l'identité des locuteurs ainsi que leurs interventions dans un flux continu audio ou dans une base de données d'archives audio, ne contenant que la parole. Dans ce cadre nous avons choisi de structurer les documents audio (restreints à des journaux radiodiffusés) selon une classification en locuteurs. La technique utilisée repose sur l'extraction des coefficients mel-cepstrales, suivi par l'apprentissage statistique de modèles de mélange de gaussiennes (MMG) et sur la détection des changements de locuteur au moyen de test d'hypothèse Bayésien. Le processus est incrémental : au fur et à mesure que de nouveaux locuteurs sont détectés, ils sont identifiés à ceux de la base de données ou bien, le cas échéant, de nouvelles entrées sont créées dans la base.

Comme toute structure de données adaptée au problème incrémental, notre système d'indexation permet d'effectuer la mise à jour des modèles MMG de locuteur à l'aide de l'algorithme fusion des MMG. Cet algorithme à été conçu à la fois pour créer une structure ascendante en regroupant deux à deux les modèles GMM jugés similaires.

Enfin, à travers de deux études utilisant des structures arborescentes binaire ou n'aire, une réflexion est conduite afin de trouver une structure ordonnée et adaptée au problème incrémental. Quelques pistes de réflexions sur l'apport de l'analyse vidéo sont discutées et les besoins futurs sont explorés.

Mots-clés : Reconnaissance automatique de locuteurs, bases de données multimédias, structuration audiovisuelle, classification hiérarchique, modèle de mélange de gaussiennes, divergence de Kullback-Leibler, architecture arborescente, structure incrémentale, Archivage audio.

Text-independent speaker technologies for Audio indexing and retrieval in the case of large data

Abstract

This thesis is devoted to techniques for speaker-based recognition systems to scale up to large amounts of data and speaker models. We have chosen to partition audio documents (news broadcast) according to speakers. The mel-cepstral acoustic characteristics of each speaker are model through a probabilistic Gaussian mixture model. First, speaker change detection in the stream is carried out by Bayesian hypothesis testing. The scheme is incremental : as new speakers are detected, they are either identified in the database or new entries are created in the database. First, we have examined some issues related to building a tree structure exploiting a similarity between speaker models. Several contributions were made.

First, a proposal for organising a set of speaker models, based on an elementary model grouping. Then, we used an approximation of Kullback-Leibler divergence for this purpose. Finally, through two studies using binary of nary tree structures, we discuss the way of a version suitable for incremental processing. Finally, perspectives are drawn regarding joint audio/video analysis and future needs are analyzed.

Keywords: Speaker recognition, multimedia databases, audiovisual structuring, hierarchical classification, Gaussian mixture, Kullback-Leibler divergence, incremental processing.

Discipline :Informatique

N° : ED 0366-372

Remerciements

Ce travail a été effectué au Laboratoire $GSCM - LRIT$ (Group Signaux,Communications and Multimedia), à la faculté des sciences, dirigé par Monsieur le Professeur Driss ABOUTAJDINE, qui me fait l'honneur de présider cette commission d'examen. Je l'en remercie sincèrement.

Cette thèse s'est, ainsi déroulée dans le cadre de co-tutelle de thèse entre l'Université Mohammed V-Agdal, à la Faculté des Sciences Rabat et l'Université de Nantes à l'Ecole Polytechnique de Nantes. Le travail est effectué de part égale, coté France, au Laboratoire d'Informatique de Nantes Altantique (LINA INRIA équipe ATLAS-GRIM, LINA FRE CNRS 2729) sous la direction de M. Noureddine Mouaddib et co-encadré par M. Marc Gelgon et coté Maroc au Laboratoire de Recheche en Informatique et Télécommunication LRIT-GSCM sous la direction de M. Driss Aboutajdine et co-encadré par M. Mohammed Rziza.

Que Monsieur Driss ABOUTAJDDINE, Directeur de ma thèse, reçoive toute l'expression de ma reconnaissance pour m'avoir proposé ce sujet, et, pour tout son dynamisme et ses compétence scientifiques qui m'ont permis de mener à bien cette étude.

Je remercie très vivement mon directeur de thèse côté français, Professeur Noureddine Mouaddib de l'École Polytechnique de Nantes, pour son encadrement à la fois sérieux et amical.

Je ne peux pas commencer mes remerciements sans évoquer Monsieur Driss ABOUTAJDINE, Directeur de ma thèse, qu'il reçoive toute l'expression de ma reconnaissance pour m'avoir proposé ce sujet de recherche, et, pour tout son dynamisme et ses compétences scientifiques qui m'ont permis de mener à bien cette étude.

Je tiens à remercier très chaleureusement le Professeur HDR Marc Gelgon et le Professeur Mohammed Rziza également pour la qualité de leur encadrement, aussi bien sur le plan scientifique que sur la dimension humaine, de par leurs orientations scientifiques, leur écoute, leur disponibilité et leur soutien et avec qui j'ai pris beaucoup de plaisir à travailler.

Mes remerciements s'adressent également aux responsables et tous membres de la Comité Scientifique de la coopération franco-marocaine dans le domaine des STIC (programme géré par l'INRIA du côté français), grâce aux moyens considérables fournis pour le bon déroulement de ma thèse.

Mes remerciements vont également aux membres de mon jury :

- Rapporteurs :
  - El Moustapha Mouaddib, Professeur à l'Université de Picardie Jules verne, Amiens.
  - Driss Mamass, Professeur à la Faculté des Sciences, Université Ibn Zohr, Agadir.
  - M. Khalid Daoudi, Chargé 1ère de recherche CNRS, IRIT Toulouse.
- Examinateurs :
  - Maria Begoña García Zapirain, Professeur à l'Université de Deusto, Bilbao, Espagne.
  - Marc Gelgon, Maître de conférences HDR à l'Ecole Polytechnique de l'Université de Nantes.
  - Mohammed Rziza, Professeur Assistant à la Faculté des Sciences de Rabat.

d'avoir accepté d'évaluer mon travail et m'ont prodigué de bons conseils que j'ai essayés de suivre.

Au cours de la préparation de ma thèse, j'ai bénéficié d'une bourse d'excellence octroyée par le Centre National pour la Recherche Scientifique et Technique (CNRST) et ce dans le cadre du programme de bourses de recherche initié par le Ministère de l'Education Nationale, de l'Enseignement Supérieur, de la Formation des Cadres et de la Recherche Scientifique. A cet effet, je remercie tous les responsables du CNRS pour m'avoir permis d'effectuer ce travail dans de bonnes conditions matérielles.

Un grand merci a tous les personnels du laboratoire d'informatique de Nantes Atlantique, spécialement M. Christian Attiogbe et Mme Christine Brunet pour leurs accueil et disponibilité, ainsi tous les membres du service technique.

Je suis également très reconnaissant envers mes amis de LINA M. Frédéric Jouault et Marcos Didonet Del Fabro pour les fructueuses discussions que nous avons eues.

Une pensée particulière aux chers collègues du laboratoire, Najlae, Siham, , Sanaa, Manal, Sanaa, Fadwa, Leila, Aouatif, Naoual, Youssef, My Ahmed, Hicham, Rachid, Khalid. En fin merci a tous les membres du Laboratoire LRIT.

Une thèse de doctorat est un travail bien trop personnel pour pouvoir se passer du soutien de ses proches. Je tiens tout particulièrement à remercier mes parents ainsi que mon frère Karim mes soeurs Fadwa et Aicha et toute ma famille, en particulier ma tante Latifa et son mari BenMhamed pour leur soutien et leur aide. Enfin je termine par remercier mes amis d'enfance et mon parcours scolaire Nassim, Mounir, Mahmoud, Driss.

Sommaire

— Pages liminaires —

— Corps du document —

— Pages annexées —

CHAPITRE 1

Introduction générale

La recherche d'information

L'informatique a permis le développement d'outils pour traiter l'information et établir la représentation des documents au moment de leur indexation. La recherche d'information est un domaine historiquement lié aux Sciences de l'information et à la bibliothéconomie Abrégée en RI ou IR (Information Retrieval en anglais). Cependant, la recherche d'information consiste à établir des représentations des documents dans le but d'en récupérer des informations, à travers la construction d'index. A l'heure actuelle la recherche d'information est un champ transdisciplinaire, qui peut être étudié par plusieurs disciplines, approche qui devrait permettre de trouver des solutions pour améliorer son efficacité. Au sens large, la recherche d'information inclut deux aspects : l'indexation des corpus et la recherche par contenu des informations sémantiques.

Compte tenu de l'accroissement gigantesque du volume de données à traiter, la tâche d'indexation devient extrêmement fastidieuse, et l'automatisation semble désormais indispensable. Dans nos équipes [1,2] cette problématique est bien mise en évidence, en particulier l'aspect incrémental de traitement d'un flux audio. Cette problématique est aussi prise en compte dans la gestion de l'information issue des données volumineuses de type multimédia.

Plusieurs axes d'investigation sont au centre de nos intérêts. Les besoins tels que la gestion, la manipulation et le stockage des données multimédia, doivent être synthétisés afin d'être exploités.

En outre, le secteur professionnel et industriel de plus en plus d'attente que soit au niveau applicatif ou au niveau de formation. Cependant, nous retrouvons ces fortes demandes des méthodologies de recherche d'information sur Internet, annuaires, moteurs de recherche, portails, sites et outils de recherche thématiques.

Hors les techniques traditionnelles de recherche d'information, la recherche par contenu dans les documents multimédia et notamment un flux continu audio ou encore dans une base volumineuse sollicitera des techniques plus ingénieuses basées sur un modèle de recherche d'information spécifique.

L'indexation

Une définition plus classique, non contradictoire d'indexation audio, est d'identification par une localisation de séquences pertinentes ou de thèmes majeurs au sein d'un document par une analyse de son contenu.

L'indexation a pour but majeur de faciliter l'accès et la recherche au sein des documents dans une grande base à l'aide des mots clé, grâce aux descripteurs qui offrent un résumé sur le contexte d'une entité qui fait partie d'une

[1] LRIT Laboratoire de Recherche en Informatique et Télécommunications : http ://www.fsr.ac.ma/GSCM/

[2] ATLAS-GRIM Groupe de Recherche d'Information Multimédia : lina.atlanstic.net/fr/equipes/team7/index.html

collection d'information. Le but principal de l'indexation est de classer ultérieurement le contenu des documents sous des ensembles partageant les mêmes propriétés.

L'objectif est, à l'aide de ses index, le système d'indexation doit de classer ultérieurement le document parmi un ensemble de documents d'une collection donnée, d'extraire le contexte de cet index au sein du document lui-même. Ce type d'indexation a pour but l'optimisation de l'accès aux données dans de grandes bases. Jusqu'à présent, les méthodes d'indexation en audio sont principalement manuelles : un opérateur humain doit lire, écouter et/ou regarder le document numérique dans le but de sélectionner par des informations de valeurs sémantiques recherchées.

La recherche d'information et la navigation dans les documents multimédia à besoin de nouvelles techniques de recherche de haut niveau. La recherche par contenu dans un document audio image ou vidéo, nécessite d'autre technique que la recherche semi-structurée utilisée actuellement dans le Web et les différents moteurs de recherche dans le Web.

L'indexation au sens du locuteur

Si l'on se réfère à la norme MPEG7, indexer un document sonore signifie rechercher aussi bien des composantes de bas niveau dites primaires comme la parole, la musique, les sons clés que des descripteurs de plus haut niveau tels l'identité des locuteurs et leurs interventions.

Un document sonore, c'est-à-dire la bande sonore d'un document multimédia ou enregistrement d'émission radiophonique, est un document particulièrement difficile à indexer, car l'extraction de l'information élémentaire se heurte à l'extrême diversité des sources acoustiques. Les segments acoustiques sont de nature très diverses de par leur production et leur enregistrement : l'environnement d'enregistrement peut être propre ou bruité, canal de transmission, musique, parole (monologue ou dialogue). Pour cela, le signal audio doit subir un certain nombre de prétraitement afin d'optimiser l'extraction des informations pertinentes.

Ce travail a pour objectif l'indexation au sens du locuteur des documents audio ne contenant que de la parole dans le cas de grande quantité de données (archives ou un flux en continu d'émission radiophonique, flux radio ou bande sonore). Le processus d'acquisition et d'identification est incrémental. Un exemple simple consiste à de détecter les changements des locuteurs équivalents à des tours de parole dans un dialogue. Cependant, cette étude se concentre sur trois problèmes : la détection de changement de locuteurs, identification du locuteur et l'organisation des modèles de locuteurs.

Problématiques

Les technologies de navigation et de recherche d'information connaissent une évolution rapide afin de répondre aux besoins diversifiés des applications multimédia. A cet effet, un fort appel a été lancé par des spécialiste pour guider les utilisateurs à une meilleure navigation. En effet, l'analyse du contenu enrichie des méta-données au delà de ce qu'on peut faire à la main, ainsi, la recherche d'information peut évoquer des ressources tel que (archives

des stations radio, télévision) ou encore le flux en continu des données audio vidéo transmises en flux continu (broadcast).

L'objectif principal est de structurer les documents audio selon une classification en locuteurs. L'apprentissage statistique de modèles de mélange de gaussiennes et la détection de changement de locuteurs au moyen de test d'hypothèses bayésiennes représentent les outils de ce travail. Cependant, le schéma d'acquisition et de traitement est incrémental. Aujourd'hui, une grande partie des sources de parole numérique sont en générale transmis sous différentes type tels que : flux d'émissions télé, radio sur Internet, contenus auto-produits ou encore le Podcasting etc... Les masses de données sont absolument énormes ce qui rends difficile aux experts d'automatiser leurs indexation, d'où de nouveaux défis sont alors définis pour le passage à l'échelle incrémentale.

L'indexation des documents audio permet alors d'identifier automatiquement les interventions des locuteurs d'une manière non-supervisée. Différentes technologies sont alors fortement sollicitées dans la mise en marche de l'étude d'exploration souhaitée. La reconnaissance automatique du locuteur (RAL) constitue l'axe principal de notre système d'indexation. Plusieurs études ont été élaborées dans ce domaine afin d'améliorer la performance de l'identification d'un côté et de réduire le temps de recherche d'un autre côté. En revanche, plusieurs contraintes font toujours face à la réalisation d'une technique offrant une réponse satisfaisante dans un cas plus pratique. En outre, nous ne pouvons pas évoquer d'indexation des documents audio sans oublier un domaine far celui de l'analyse du signal audio, un domaine qui s'intéresse à améliorer les traitements et adaptation du signal audio pour l'acquisition, débruitage, compression, transcription, séparation de source et la détection d'événement sonores dans un document audio. Toute une panoplie d'axes de recherche sont en cours de progression avec le même centre d'intérêt celui de rattraper le développement rapide du volume de données et les applications appropriées. Dans la suite de cette présentation des domaines technologiques qui se superposent au domaine d'indexation au sens du locuteur, la reconnaissance automatique de locuteur représente le troisième domaine technologique qualifié comme un paquetage des outils de modélisation. Le langage de modèle utilise des techniques de modélisation déterministes, stochastiques ou probabilistes.

Contexte de travail

Le présent travail entre dans le cadre de co-tutelle de thèse entre l'Université Mohammed V-Agdal et l'Université de Nantes. Le travail est effectué de part égale dans les laboratoires Laboratoire d'Informatique de Nantes Atlantique (LINA) le laboratoire de Recherche en Informatique et Télécommunications, Groupe de Signal et Communications Multimédia (LRIT-GSCM), dans le cadre du coopération Franco-marocaine STIC. Les travaux de recherche ont été dirigés par :

- Côté Maroc :
  - Prof. Driss Aboutajdine, Professeur à l'Université Mohammed V-Rabat FSR.
  - Prof. Mohammed Rziza, Professeur Assistant à l'Université Mohammed V-Rabat FSR
- Côté France :
  - Prof. Noureddine Mouaddib, Professeur à l'Ecole Polytechnique de Nantes.
  - MC. Marc Gelgon, Maître de conférence HDR à l'Ecole Polytechnique de Nantes.

Organisation du mémoire

Ce document est organisé en deux parties.

La première partie permet de situer le cadre de notre travail à l'aide d'une présentation des fondements théoriques nécessaires. Cette partie est composée de deux chapitres intitulée une présentation de l'indexation audio. Le chapitre 1, présente l'indexation audio comme technologie de reconnaissance automatique du locuteur ainsi que certains aspects applicatifs qui sont au coeur des problématiques traitées dans cette thèse. Le système proposé est basé sur des outils de paramétrisation cepstrales et de représentation à l'aide du modèle de mélange de gaussiennes (MMG) les mieux adaptée dans le cas traitement et l'analyse de signal en mode indépendant du texte prononcé. Le chapitre 2, présente l'état de l'art des techniques de segmentation progressive en locuteur des documents audio. Cependant, l'approche probabiliste est présentée de façon formelle. La prospection de l'état de l'art a permis d'identifier certaines faiblesses des techniques actuelles et de définir les points particuliers que nous allons traiter.

Parmi les inconvénients d'un tel de système est sa forte dépendance vis-à-vis des conditions d'enregistrement et notamment la taille de données utilisées. Ce manque de robustesse ne permet pas une utilisation aveugle sur tout type de document. C'est pourquoi nous proposons un nouveau système d'indexation audio au sens de locuteur par des améliorations sur les trois niveaux de base du système ; à savoir : Une segmentation non supervisé en locuteur est réalisée à l'aide de test d'hypothèses bayésiennes, une représentation des segments de locuteurs qui reste fiable et efficace lors du passage à l'échelle incrémentale et enfin par une organisation hiérarchique des modèles de locuteurs afin de réduire le temps de recherche.

La seconde partie est consacrée au développement de nouvelles techniques d'indexation au sens du locuteur basées sur l'utilisation d'une mesure de similarité entre modèles de mélange de gaussiennes. Nous abordons dans le chapitre 3, par un panorama des techniques de mesure de similarité entre modèles de mélange de gaussiennes développées jusqu'ici en RAL, ce qui permet d'étudier et valider l'approche d'organisation hiérarchique via une première expérience de regroupement des modèles de mélange de gaussiennes sous une structure arborescente.

Le quatrième chapitre présente la contribution théorique basées sur une nouvelle technique d'organisation hiérarchique des modèles de mélange de gaussiennes. Les principales contributions dans ce travail de thèse consiste à réduire le temps de requête et la complexité de recherche de l'identité des locuteurs dans les documents audio, notamment, dans le cas de grands volumes de données. Plusieurs propositions sont alors présentées dans les chapitres $4, 5, 6$ et 7 voir ci-dessous les différentes contributions selon l'ordre chronologique dans le manuscrit.

Des propositions alternatives aux techniques de recherche exhaustive actuellement utilisé dans le domaine de la RAL présentée dans l'état de l'art ont été élaborées dans l'ordre suivant :

Chapitre 4 : Première contribution basée sur des techniques de mesure de similarité (i.e la divergence de Kullback-Leibler (KL), ou encore la vraisemblance des données) a donnée lieu à une approche de regroupement ascendant de modèles de mélange de gaussiennes en utilisant les données d'entraînement.

Chapitre 5 :
- (a) Développement d'un algorithme de fusion de modèles de mélange de gaussiennes pour optimiser le coût de construction d'un arbre de recherche de modèle de locuteurs par lot de modèles de mélange de gaussiennes via un arbre binaire de recherche ascendant [69].
- (b) Création d'un algorithme incrémental d'organisation des modèles de mélange de gaussiennes de locuteurs basé sur la mesure de divergence de KL proposant un mécanisme de mise à jour des modèles de mélange de gaussiennes à l'aide des segments temporels du locuteurs [70].

Chapitre 6 : Une approche originale et efficace d'organisation hiérarchique des modèles de mélange de gaussiennes de

locuteurs a priori est basée sur des techniques de classification (dendogram-based ou K-mean like), à partir d'une expression modifiée de la divergence de KL de noeud père et noeud fils des critères d'optimisation sont dérivés. Le schéma proposé est évalué sur des données réelles [67].

Chapitre 7 : Dans un travail récent, nous avons proposé une nouvelle approche permet de définir des régions temporelles de similarité à l'aide de structure en treillis. Deux majeures contributions sont réalisées :

(a) une transformation non linéaire de l'approximation de Monte-Carlo de la divergence de KL avec réduction de nombre de point sigma (i.e. 'point sigma' est représenté par le vecteur de la covariance) des composantes gaussiennes des deux modèles de mélange de gaussiennes comparés donne lieu à une nouvelle approximation de la divergence de KL robuste et discriminante entre modèles de mélanges de gaussiennes de locuteurs.

(b) un algorithme de création d'une structure arborescente adaptée au problème incrémental à l'aide de treillis, la structure permet de définir sur des régions temporelles selon l'ordre de similitude des modèles de mélanges précédemment inscrits partageant une mesure de similarité très proche entre le modèle en cours de traitement [68].

Enfin, un récapitulatif de tous les avantages et les performances des méthodes proposées suivi des perspectives comme conclusion générale de ce travail.

PARTIE I

Présentation de l'indexation audio

CHAPITRE 2

Positionnement et aspects applicatifs de l'indexation audio

Ce chapitre introduit les notions de base permettant de situer le cadre de travail présenté dans cette thèse. Il est consacré à la description de l'environnement applicatif des systèmes de traitement de signal avancé (reconnaissance, identification et vérification, etc.). En particulier, la description par locuteur du document audio et les difficultés rencontrés en pratique. Ces difficultés proviennent essentiellement de l'adaptation des algorithmes d'identification de locuteur au problème incrémental dans le cas d'un flux en continu ou des archives audios. En suite, nous précisons la place et le rôle de l'indexation au sens du locuteur parmi les techniques de la reconnaissance automatique de locuteur.

2.1 Indexation multimédia

Le but de l'indexation multimédia des documents sonores est de créer, stocker et gérer des bases de données multimédia notamment des documents sonores sous leurs différents forme de diffusion. À l'heure actuelle, les méthodes d'indexation en audio sont principalement manuelles : un opérateur humain doit lire, écouter et/ou regarder le document numérique de façon à sélectionner les informations sémantiques à indexer. Cependant, la norme MPEG-7 [56, 57], permet de semi-structurer les documents audio par des composantes de bas niveau dites "primaires" comme la parole, la musique, les sons clés (jingles, mots-clés...) ainsi que par des descripteurs de plus haut niveau tels les sons dans in environnement intelligent (émotion, événements, ...)

Les documents audio contiennent deux types d'informations : des informations concernant le sens (bruit, music, parole, code sonore, etc ...) et des informations concernant la source (excitation des poumons). Si la tâche

associée à l'extraction du sens est difficile, la tâche concernant la détection de la source n'en demeure pas moins. Actuellement, avec la présence de milliers de documents audio-visuels et audio diffusés à travers des chaînes de télévision et de chaînes de radio, une quantité immense d'informations provenant des différents coins du monde demande à être traitée et filtrée en vue de l'extraction des informations propres à une personnalité bien déterminée. Le but de ce travail est de structurer les documents audio au sens de locuteurs pour un SGBD audio avancé. Pour cela, une nouvelle tâche de la reconnaissance automatique de locuteur a vue le jour en relation avec l'essor du multimédia dans notre société. L'objectif sera alors de rendre l'information souhaitée (propre à un locuteur donné) aisément accessible.

2.1.1 Bases de données multimédias

Depuis son apparition, le rôle du multimédia dans la diffusion de l'information n'a cessé de croître, pour en être aujourd'hui le principal support. La problématique d'une diffusion massive et de qualité n'a toujours pas été résolue. Nous disposons d'un héritage volumineux de documents multimédias (vidéo, audio, image), qui bien sûr ne fait qu'augmenter. Pour donner un exemple, l'INA (Institut National de l'Audiovisuel en France) dispose d'un patrimoine de documents audiovisuels constitué à ce jour de plus de 1.455.000 documents de productions audiovisuelles nationales, de 625.000 documents d'actualités et de 1.000.000 de documents régionaux.

Nous avons donc de grandes quantités d'informations multimédias en notre possession et se pose alors le problème d'organiser ces informations permettant de retrouver une information particulière dans tous ces documents. Face au nombre de ceux-ci, une recherche exhaustive consiste à explorer chaque document devient une tâche excrément fastidieuse. Il est donc nécessaire de disposer d'outils d'archivage et des structures de données qui permettent de rechercher efficacement une information donnée au sein des masses de documents multimédias.

L'utilisation des méta-données pour la description manuelle du contenu est actuellement le seul moyen d'intégrer les données audio dans un SGBD [75]. Les SGBD Multimédia doivent offrir des services pour sauvegarder, jouer, rechercher, manipuler, et réaliser des fonctionnalités de gestion et d'organisation des applications adaptées aux utilisateurs. Certains SGBD commercialisés proposent des modules tel que le DB v6 $(QBIC)$ et ORACLE $8i$ InterMedia. Ces modules offrent un minimum d'intégration de ce type de documents. Des outils d'exploration et de navigation via des requêtes peuvent être exprimés sur des images sous des critères de histogramme de couleur, de texture et de contour. En revanche, ni l'audio ni la vidéo ne sont encore traitées. En outre, deux grands volets de difficultés se posent, d'une part la complexité du pré-traitement du document audio par son organisation en "locuteur", "non locuteur", "musique", "silence", "bruit", et d'autre part, la fiabilité du système dans lors de passage

à l'échelle incrémentale. Cependant, dans le cas des données audio l'extraction automatique des méta-données à partir du signal audio demeure très difficile [40].

Une exploitation sérieuse de cette masse d'informations est pratiquement non réalisable sans un système logiciel intelligent. La future génération de systèmes reposera sur la possibilité de stocker, d'accéder et de raisonner sur de gros volumes d'objets multimédias.

2.1.2 Principe général de système d'indexation en locuteurs

Le principe général de notre système d'indexation est illustré par (figure 2.1). D'abord, une segmentation du signal audio est effectuée sur des fenêtres d'analyse de (20ms) avec un recouvrement des segments de 50%. Une deuxième étape consiste à extraire des descripteurs acoustiques du signal sous forme des vecteurs *Mel Frequency Cepstral Coefficients* (MFCC). Les descripteurs MFCC offrent une information pertinente sur le locuteur ce qui permet de réaliser en premier temps une segmentation temporelle du signal en locuteur. Au fur et à mesure que le système extrait des descripteurs, ces derniers seront identifiés dans une base de données à l'aide d'un module étiquetage pour chaque segment. Enfin, la gestion et l'analyse de ces données utilise des services de requêtes adaptées à un SGBD audio. Un modèle de base de données (relationnel, Objet, XSLT, XQuery) doit bien être conçu pour offrir une grande flexibilité afin d'exprimer des requêtes approfondies qui répondent aux besoins des utilisateurs.

Dans la première étape, la segmentation audio permet d'organiser le document selon plusieurs descriptions primaires ou encore sémantiques (parole, music, homme, femme, enfant, etc.). Un prétraitement de document permet de de regrouper les segments selon des types et genres d'évenement détectés et. Notons que, l'indexation au sens du locuteurs est effectuée sur les segments ne contenant que de la parole. Dans notre travail, le système d'indexation est basé sur les hypothèses suivantes :

1. pas d'information a priori : ni sur les locuteurs, ni sur le langage.
2. le nombre de locuteurs est inconnu.
3. que des données de parole : ni musique, ni publicité.
4. les personnes ne parlent pas simultanément.
5. avec (flux en continu) / sans de contrainte de temps réel (cas d'archives audio).

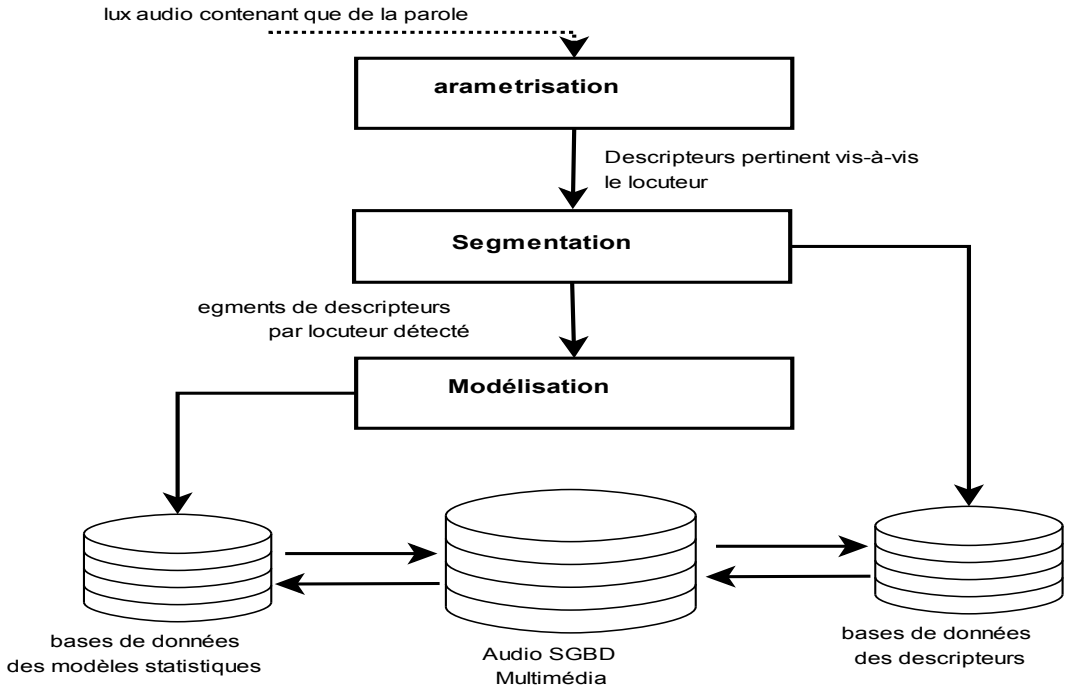


Figure 2.1 – Synoptique du système d'indexation audio au sens de locuteur

2.1.3 Typologies des systèmes de RAL

La RAL est un domaine de traitement automatique de la parole qui regroupe toutes les tâches liées à l'extraction et à l'exploitation d'information concernant l'identité des locuteurs dans un enregistrement audio. Historiquement, la vérification et l'identification du locuteur sont les premières tâches de la RAL qui sont apparues, liées à des besoins de sécurité. De nouvelles tâches ont récemment vu le jour en relation avec l'essor du multimédia dans notre société. Il s'agit de tâches d'extraction et d'exploitation d'informations relatives aux locuteurs dans des bases de données audios ou multimédias. Notons enfin l'utilisation grandissante de la technique de la RAL dans les tâches de transcriptions orthographique de documents audios. Les informations liées aux locuteurs peuvent en effet aider amplement à améliorer les performances des systèmes de reconnaissance de la parole (voir figure. 2.2).

Dans la présentation de ce qui suit, nous avons distingué trois tâches principales de la RAL en fonction du type d'information qu'elle permet d'obtenir, c'est-à-dire en fonction de la nature de la sortie fournie par le système :

- la détection du locuteur qui fournit une sortie binaire de type détection/non-détection ;
- l'identification du locuteur dont la sortie est un identifiant de locuteur ;
- la description en locuteur de documents audios qui fournit une liste descriptive des interventions des locuteurs intervenants dans le document.

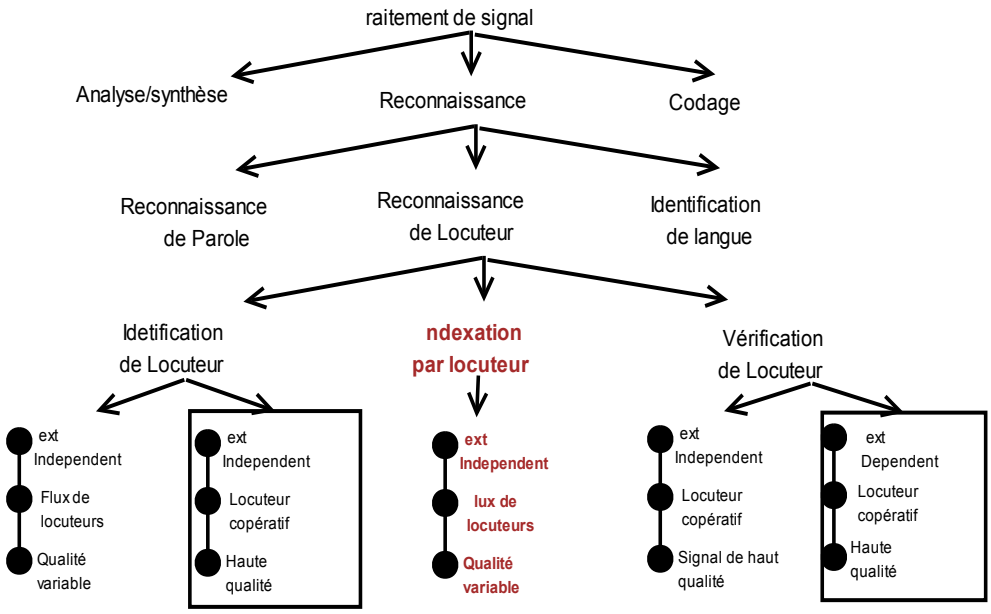


Figure 2.2 – Indexation par locuteurs comme tâche émergente de la RAL

Parmi ces tâches de base, nous présentons des variantes liées aux conditions applicatives, ou des sous-tâches intermédiaires. Nous décrivons les différents types de systèmes d'un point de vu fonctionnel, en identifiant les entrées/sorties et les données disponibles pour effectuer la tâche considérée.

2.2 Traitement "avancé" du signal audio

Les domaines d'application centrés sur le signal sonore comportent trois volets : la caractérisation du locuteur (notamment pour la vérification vocale d'identité et le suivi de locuteur dans les enregistrements sonores), le traitement "avancé" des signaux sonores (détection de classes de sons, séparation de sources et de voies) ainsi que certains aspects de reconnaissance de parole. Les fondements scientifiques de nos activités s'inscrivent dans le cadre des mathématiques appliquées, du traitement du signal, de la modélisation statistique, de l'estimation statistique et la théorie de la décision. Nous nous appuyons sur les outils de traitement du signal au niveau de la représentation du signal, de sa paramétrisation et de sa décomposition. Les approches probabilistes interviennent au niveau de la modélisation acoustique et de la classification à l'aide de tests d'hypothèses.

2.2.1 L'authentification biométrique

Les technologies de l'authentification ou l'identification biométrique regroupent un ensemble de procédés dont le but est d'identifier automatiquement une personne à partir de la mesure directe de l'une des ses caractéristiques physiques ou comportementales. Alors que les mots de passe, les clefs ou les cartes sont facilement oubliés, perdus ou volés. L'identification biométrique permet de s'en affranchir et de sécuriser, sans ce type de contrainte, l'accès à un service, aux locaux ou aux données protégées. Cette caractéristique fait de la biométrie l'une des technologies privilégiées pour sécuriser les applications pour lesquelles le client n'est pas physiquement en contact avec son prestataire comme dans [39] où elle permet de sécuriser l'accès à certains sites sur internet.

Apparues il y a une trentaine d'années au sein de la société *"Shearson Hamil"* à "Wall Street" sous la forme d'un système vérifiant la taille des doigts des employés, les techniques automatique d'authentification biométrique n'ont, dès lors, cessé de se diversifier et de se perfectionner.

D'une manière générale, l'évolution des modes de consommation de la société avec un exemple, la forte augmentation des transactions et achats en ligne et la généralisation de l'électronique personnelle, a considérablement accru l'intérêt suscité par la biométrie.

2.2.2 La vérification automatique de locuteur

La vérification automatique de locuteur (VAL) est une technologie de sécurisation bien adaptée aux applications Homme-Machine afin de garantir la simplicité et la transparence vis-à-vis de l'utilisateur. De plus, elle apparaît actuellement comme le moyen le plus adéquat pour sécuriser les transactions ou échanges de données sur le réseau téléphonique et sur internet. Le signal de parole est fortement corrélé avec certains attributs physiologiques et comportementaux du locuteur. Leurs influences se retrouvent dans la densité spectrale de puissance du signal à court terme (caractéristique du conduit vocal, de la source glottique et du timbre de la voix). La vérification du locuteur repose sur tous ces attributs qui définissent la variabilité interlocuteur.

2.2.3 Difficultés rencontrées en authentification de locuteur dans un système d'indexation audio

La VAL repose sur la variabilité interlocuteur du signal de parole. Cependant, le signal tel qu'il est transmis au système de vérification contient trois autres sources de variabilité difficilement dissociables de cette première :

1. la variabilité intra-locuteur, liée aux changements et à l'évolution de la voix d'un locuteur,
2. la variabilité introduite par les modifications des caractéristiques du matériel d'acquisition et de transmission du signal de parole,

3. la variabilité due au changement d'environnement et par exemple à la présence de bruit dans le signal de parole.

L'obtention d'une représentation robuste à la variabilité intra-locuteur est cruciale pour la réalisation d'un système d'identification automatique de locuteur notamment dans un système de VAL. Les deux autres facteurs de variabilité sont des causes majeures de la dégradation des performances et parmis les principaux défies des systèmes de VAL pour réduire leurs influences autant que possible.

2.2.3.1 Variabilité intra-locuteur

Deux signaux de parole produits par la même personne prononçant le même énoncé sont toujours différents. Le signal de parole est par nature un processus aléatoire et il est impossible pour un individu de produire plusieurs fois exactement le même signal. De plus, la voix d'un locuteur est fortement influencée par son état physique et émotionnel. Tous les changements observables dans la voix d'une même personne définissent la variabilité intra-locuteur. Ils ont pour origine différents facteurs qui peuvent être :

- occasionnels : l'état pathologique (rhume, maladie des dents, etc.), émotionnel (stress, angoisse) ou de fatigue d'une personne qui modifie temporairement sa voix.
- à moyen terme : le comportement d'un individu se modifie lorsqu'il s'habitue au système, ainsi s'il s'applique et parle distinctement lors des premiers accès, sa parole évolue et devient plus naturelle au cours des accès suivants. L'évolution à moyen terme de la voix des utilisateurs est un des effets du mode de fonctionnement stable/instable des systèmes d'authentification biométrique.
- à long terme : la voix évolue avec l'âge.

Même si les variations à court terme sont très préjudiciables aux systèmes de vérification du locuteur, des travaux ont permis de mettre en évidence une dégradation croissante des performances en fonction de temps nécessaire entre la phase d'apprentissage et celle de la décision via un test. La mise à jour régulière par un apprentissage incrémental des modèles de locuteurs avec de nouvelles données permet d'assurer une amélioration des performances obtenues [28].

2.2.3.2 Variabilité matérielle d'acquisition et canal de transmission

De nombreux travaux expérimentaux comme par exemple [84] et surtout l'étude des performances sous différentes configurations de test des nombreux systèmes présentés aux évaluations [65] ont permis de mettre en évidence l'influence des variations de la chaîne d'acquisition entre la phase d'apprentissage et celle de test. Les va-

riations sont à l'origine d'une forte dégradation des performances obtenues. Ceci est particulièrement vrai lorsque le signal est fortement perturbé par le canal de transmission, comme c'est le cas avec la parole téléphonique (limitation de la bande utile et distorsion dues au combiné et au canal de transmission). Ainsi la compensation et l'annulation des effets de cette variabilité sont l'un des enjeux fondamentaux de la recherche actuelle.

2.2.4 La RAL comme technique de traitement automatique du locuteur

Dans la section précédente, nous avons inscrit l'indexation audio au sens du locuteur comme une tâche émergente de la reconnaissance automatique du locuteur. Ceci nous a permis de situer la RAL par rapport à des technologies concurrentes de conception d'applications pour l'indexation automatique du locuteur. Dans cette partie, nous présenterons l'indexation audio dans le contexte de la technologie de la RAL et plus particulièrement celui du suivi de locuteur et de segmentation par locuteur. La motivation principale pour situer l'indexation audio dans ce contexte est liée aux fortes interactions de réalisation entre les systèmes d'indexation audio et ceux des différentes applications du traitement automatique de louteur.

2.2.5 Indexation sonore

Un document sonore est l'enregistrement d'un signal acoustique obtenu à partir de plusieurs sources de production sonore. Il est donc constitué de nombreuses composantes, dont les plus communes sont la parole et la musique. Ces deux composantes sont dites primaires, et il faut leur rajouter les composants liés aux multiples sources de bruit potentielles. Avant d'aller plus loin, il convient de préciser certaines définitions, notamment sur la parole et la musique afin de lever toute ambiguïté [61].

Parole :

Le signal de parole appartient à la classe des signaux acoustiques produits par des vibrations des couches d'air. Les variations de ce signal reflètent les fluctuations de la pression de l'air [11]. La parole est une suite de sons produits soit par des vibrations des cordes vocales (source quasi périodique de voisement), soit par une turbulence créée par l'air s'écoulant dans le conduit vocal, lors du relâchement d'une occlusion ou d'une forte constriction de ce conduit (sources de bruit non voisées) [15].

Musique :

Les particularités de la musique, qui la différencient de toutes autres sonorités, ne résident pas seulement dans des différences culturelles, mais dans des propriétés physiologiques très spécifiques du système auditif de l'homme. Ainsi, définir la musique est très difficile car celle-ci peut être produite et perçue de différentes manières.

2.2.6 La caractérisation automatique du locuteur

La caractérisation automatique du locuteur est une tâche du traitement du signal, sous-ensemble des disciplines de l'interaction Homme-Machine regroupant l'ensemble des technologies ayant pour objet l'étude du signal de la parole, la reconnaissance de la parole et la caractérisation automatique du locuteur.

Le but de la caractérisation automatique du locuteur est d'extraire du signal de la parole toutes sortes d'informations relatives à l'individu l'ayant prononcé. La nature des caractéristiques recherchées est très variée (identité, pathologie, origine géographique, etc.) et dépend du type d'application visée. La reconnaissance automatique du locuteur est une discipline dérivée de la caractérisation automatique du locuteur. Elle regroupe l'ensemble des applications dont le but est de déterminer l'identité d'une personne à partir de sa voix.

Rappelons qu'en domaine de la RAL, on considère généralement que le signal de parole est une source véhiculant trois informations différentes : la première correspond à l'information linguistique, la seconde à l'information sur le support de transmission du signal et la troisième à l'information propre pour caractériser le locuteur. Le système de RAL va chercher à isoler et à interpréter les différentes informations sources pour identifier le locuteur.

2.2.7 Classification des systèmes de reconnaissance automatique du locuteurs

Une classification essentielle d'un système RAL est basée sur la dépendance au texte. Pour définir d'autres critères de la classification, nous pourrons par exemple fixer le type d'environnement opérationnel ou matériel d'acquisition du signal de parole. Cependant, ces caractéristiques de conception sont générales à tout système d'indexation et identification de locuteur.

Deux types de traitements sont adoptés par des systèmes de reconnaissance de locuteurs : nous distinguons ceux dépendants du texte et ceux indépendants du texte. En mode dépendant texte, la reconnaissance d'une personne est réalisée sur un signal de parole dont le contenu linguistique (mot de passe, phrase, code) est connu des systèmes. Les différentes configurations possibles sont :

- systèmes à messages fixés : la vérification de l'identité du client est alors précédée d'une étape de reconnaissance de la parole. La personne doit, selon les cas, prononcer un message qu'elle aura préalablement choisi

[38]. Dans le dernier cas, le message sera imposé par le système [37]. Dans le cas précédent, le message peut être différent à chaque nouvel accès. La motivation de cette approche est de se protéger des imposteurs disposant d'un enregistrement de la voix d'un client.

- système à unités segmentales fixées : lors d'un accès au système, le client doit prononcer un signal de parole contenant soit une séquence de mots (ex : chiffres [53]), soit des traits phonétiques connus du système [46].
- systèmes indépendant du texte, Le système de reconnaissance n'impose alors aucune contrainte sur le contenu linguistique du signal de parole.

L'information apportée par la connaissance *a priori* du contenu linguistique permet généralement d'augmenter les performances. Cependant cette amélioration est obtenue au prix d'une baisse de l'ergonomie du système : l'utilisateur doit se souvenir d'un mot de passe, soit être en mesure de lire un message prédéfini.

2.3 Description des applications de la RAL

Après l'identification automatique et la vérification automatique comme applications liées à la préoccupation de la RAL (i.e l'authentification automatique humaine), sont apparus de nouveaux objectifs plus complexes liés à l'extraction et à la gestion d'information dans les bases de données multimédia. En effet, les deux principaux domaines d'applications du RAL sont actuellement liés d'une part à la sécurisation par authentification du client et d'autre part à l'indexation pour l'aide à la navigation dans les documents audio.

La suite de cette section présente brièvement les différentes tâches du RAL autres que la vérification du locuteur et donne pour chacune d'elles un schéma de leur principe de fonctionnement ainsi que les exemples d'applications.

2.3.1 Identification automatique du locuteur

L'identification automatique de locuteur consiste à déterminer, parmi une population de N locuteurs connus, celui qui a prononcé un message donné. Lors d'un accès à un système d' identification automatique de locuteur, le signal de parole fourni à l'entrée du système est comparé à la référence caractéristique de chacun des locuteurs connus et l'identité retournée est celle dont la référence est la plus proche du signal de test. Le signal est la seule entrée du système d'identification automatique de locuteur. Dans un système d'identification du locuteur sur un ensemble fermé, le locuteur est supposé être l'un des N locuteurs du système. Dans un système d'identification du locuteur sur un ensemble ouvert, le système peut décider qu'aucune des N identités connues n'est celle du locuteur. Il doit pour cela disposer d'un modèle de rejet.

Les performances obtenues par les systèmes d'identification automatique de locuteur sont directement liées

au nombre de N locuteurs du système. La figure (2.3) représente un schéma illustrant le fonctionnement d'un système d'identification automatique de locuteur.

Application :

En ensemble fermé, les applications d'un système d'identification automatique de locuteur sont peu nombreuses. L'identification automatique du locuteur peut cependant être utilisée de manière très efficace pour simplifier l'accès des membres d'une population d'individus à des données ou à des services personnalisés (mise en place automatique de paramètres d'utilisation, etc.). En ensemble ouvert, les applications de l'identification automatique de locuteur sont essentiellement liées à des problèmes de sécurisation comme la protection de l'accès à des sites sensibles.

2.3.2 Détection automatique de locuteur

La détection automatique des locuteurs consiste à déterminer la présence ou nom d'un locuteur donné sur un enregistrement audio. Si l'on fait l'hypothèse que le signal sonore est mono-locuteur, cette tâche est équivalente à la vérification automatique du locuteur. Comme dans le cas de la VAL ; l'identité recherchée ainsi que le signal de parole constituent les deux entrées des systèmes de détection automatique du locuteur (voir figure 2.4).

Les figures (2.4, 2.5, 2.6), note la présence de différents locuteurs dans signal d'entrée est symbolisée par différents niveaux de gris.

Application :

Les applications de la détection des locuteurs sont toujours liées en principe à la sécurisation (authentification de l'interlocuteur dans une communication téléphonique pour la validation de transactions, etc.). Cependant, d'autres applications du domaine de l'indexation de documents multimédia telles la recherche d'informations dans un document audio numérisé ou la navigation dans les données sonores sont actuellement étudiées. Ainsi, les futurs moteurs de recherche permettront sans doute de retrouver des fichiers audio contenant la voix d'un individu donné.

2.3.3 Description en locuteur de documents audio

La tâche dite **description en locuteur** a pour but de structurer un enregistrement audio en fonction des locuteurs intervenants sur cet enregistrement. L'unique entrée du système est l'entrée 'signal' où arrive le flux audio et

la sortie est une description des interventions des locuteurs, c'est-à-dire une liste d'intervalles temporels étiquetés en fonction des différentes interventions.

D'un point de vue général, la tâche de description en locuteur comprend deux sous tâches : une tâche de **segmentation en locuteur** du flux audio, c'est-à-dire une détermination des tours de parole, et une tâche de reconnaissance appliquée sur segment. Suivant les algorithmes utilisés, ces deux sous tâches peuvent être traitées conjointement ou séquentiellement, voir même itérativement.

Deux variantes de la tâche de description en locuteur existent en fonction des informations *a priori* dont on dispose pour structurer le document. Dans la première configuration, la description est effectuée à partir d'un ensemble de références caractéristiques des locuteurs connus. Cette tâche est nommée **de suivi de locuteur**.

Une entrée de sélection du ou des locuteurs à suivre peut éventuellement être utilisée.

Dans la deuxième configuration, la tâche de description en locuteur doit se faire sans aucune connaissance a priori : on ne possède pas de références caractéristiques des locuteurs et on ne connaît pas non plus le nombre de locuteurs intervenants dans l'enregistrement. Le but est alors de segmenter le document en tours de parole et d'étiqueter ces tours de parole en fonction des différents locuteurs intervenants. Nous identifierons cette tâche sous l'appellation d'organisation en locuteurs (en anglais *Speaker Diarization*) d'un document audio. La sortie du système est une liste descriptive qui représente temporellement les interventions des différents locuteurs "supposés" qui sont alors identifiés de façon arbitraire et répertoriés dans l'ensemble du document.

Les applications de la description en locuteur sont principalement liées au domaine du multimédia, qui sont en plein essor actuellement. Elle peut servir à indexer un document sonore ou audiovisuel afin de permettre une navigation rapide à l'intérieur de celui-ci pour retrouver les interventions de tel ou tel locuteur. La description en locuteur peut également servir d'aide à la transcription manuelle d'enregistrements audio, par exemple des comptes-rendus de réunion. Dans ce cas, la liste descriptive fournie par le système permet au transcripteur d'identifier plus rapidement les différents intervenants au cours de la réunion. Enfin, les systèmes de transcription orthographique automatique de documents audio peuvent utiliser la description en locuteur pour enrichir la transcription ou pour adapter le système de reconnaissance de parole au locuteur.

Les schémas du fonctionnement d'un système d'indexation et de suivi de locuteurs sont présentés respectivement dans (figures 2.5 et 2.6).

Applications :

Le domaine d'application de ces deux tâches est principalement le développement d'un SGBD audio avancé et un moteur de recherche rapide par contenu audio. Citons par exemple [24] : la recherche d'information dans des séquences d'émissions télévisées ou radiophoniques, l'estimation du temps de parole de chaque intervenant lors d'un débat et la recherche des interventions d'une personne dans des archives.

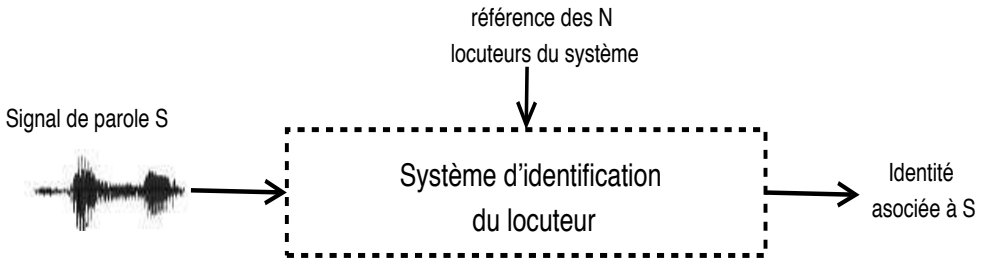


Figure 2.3 – Schéma d'un système d'identification du locuteur

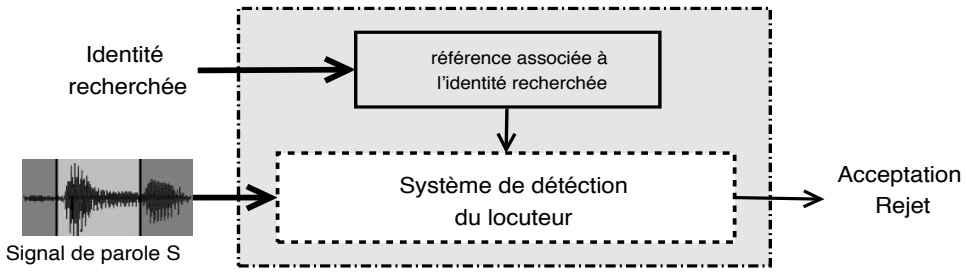


Figure 2.4 – Schéma d'un système de détection du locuteur

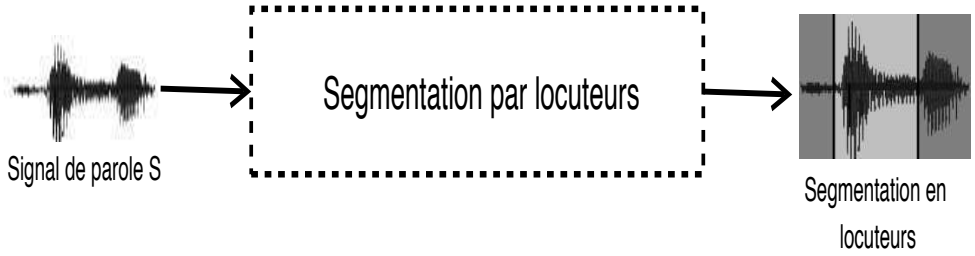


Figure 2.5 – Schéma d'un système d'indexation par locuteur

On pourra trouver des études détaillées de la tâche de description en locuteur et de ses sous-tâches dans les

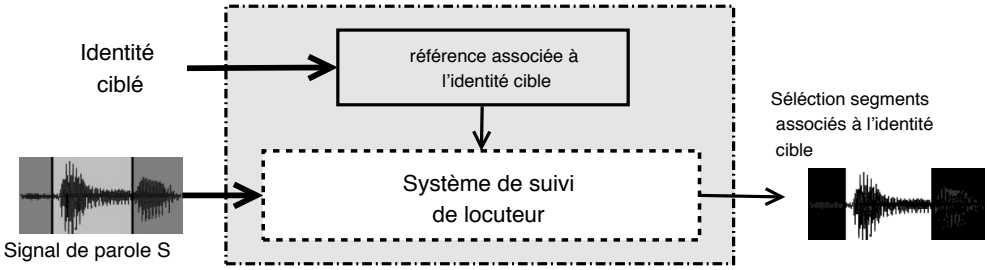


Figure 2.6 – Schéma d'un système de suivi de locuteurs

travaux de thèse de Sylvain Meignier [51] (organisation en locuteurs et segmentation), de Mouhamadou Seck [74] (segmentation et suivi de classes de sons) et de Perrine Delacourt [24] (segmentation et regroupement pour lorganisation en locuteur).

2.4 Les méthodes émergentes des systèmes de reconnaissance du locuteur

Auparavant, les approches probabilistes ont dominé l'état de l'art des systèmes de reconnaissance du locuteur. Elles proposent un cadre puissant pour prendre en compte un certain nombre de variabilités contenues dans le signal de parole. En outre, elles permettent de définir de façon précise une mesure de similarité entre un ensemble de données de test et une référence caractéristique, représentée dans ce cas par un modèle probabiliste du locuteur. Le succès des méthodes probabilistes a engendré une multitude de travaux. En mode indépendant du texte pronocé, la technique de modélisation prédominante est basée sur des modèles de mélange de gaussiennes (MMG, MMG-UBM, MMG super factor). De nombreuses techniques destinées à renforcer la robustesse des systèmes basés sur cette approche ont été développées, laissant parfois peu de place au développement d'autres méthodes.

Citons cependant certaines méthodes alternatives, comme par exemple les réseaux de neurones probabilistes [29] ou les classifieurs linéaires polynomiaux [18], qui ont obtenu de bonnes performances en RAL.

Plus récemment, les approches par "*Support Vector Machine* : machine à vecteur de support" (SVM) ont fait une percée parmi les méthodes les plus performantes en RAL. Les travaux de thèse de Vincent WAN [85] notamment ont marqué un pas dans l'avancement de ces techniques en développant des systèmes SVM rivalisant avec les systèmes basés sur l'approche probabiliste. Depuis, les SVM ont été utilisés dans le cadre des évaluations NIST [16]. De nouvelles méthodes visant à renforcer la robustesse de ce type de systèmes ont été développées [79], amenant les performances des systèmes SVM à un niveau équivalent. En outre, le caractère fondamentalement différent de la méthode de classification par SVM (modèles discriminants) par rapport aux approches probabi-

listes (modèles génératifs) a permis d'obtenir des gains significatifs des performances lorsque ces deux types d'approches sont utilisés conjointement, dans le cadre d'une fusion de plusieurs systèmes [17].

2.5 Problématiques soulevées et orientation du travail

L'indexation au sens de locuteur est basée sur la performance des techniques de la reconnaissance automatique de locuteurs, bien qu'il y a encore une forte demande d'améliorer d'une part cette performance à cause de la variabilité intrinsèque de la voix d'un locuteur et d'autre part des conditions d'acquisition du signal de parole. Du fait de cette limitation, les références caractéristiques construites au moment de l'apprentissage ne sont représentatives que d'une partie restreinte des conditions d'acquisition et de transmission du signal de parole. En outre, le problème de traitement des données comme la voix dans le cas d'un flux en continu ou dans le cas de grands volumes de données est une tâche très coûteuse en terme de temps et de calcul. En d'autres termes, le système d'indexation par locuteur où l'identification est un processus incrémental qui consiste à effectuer des comparaisons linéaires entre tout les modèles existants contre le modèle requête.

Les systèmes de gestion de bases de données multimédia, n'offrent qu'un minima d'intégration des documents tel que l'audio, l'image et la vidéo. Il est donc primordial d'étudier et de proposer des solutions afin d'adapter le système au problème incrémental tout en gardant des bonnes performances en terme d'indexation. Cela exige un travail en aval afin de choisir et adapter les techniques performantes en terme de reconnaissance automatique de locuteurs, et d'autre part un travail en amont pour proposer des approches novatrices dans le but de pouvoir stocker et de retrouver efficacement des descripteurs dans un système de gestion de la base de données multimédia.

Trois types de stratégies peuvent être envisagées à cet effet :

1. Améliorer la performance de la représentation de locuteur, il faut rechercher des méthodes et des outils qui à la fois améliorent la performance de la caractérisation et aussi permettent d'offrir plus de souplesse en terme de technique de mesure de similarité comme outils de construction et d'exploration dans une structure optimisée et adaptée au problème incrémental.
2. Améliorer la robustesse de la modélisation, pour cela, il faut choisir judicieusement de nouveaux descripteurs (modèles), afin de décrire correctement les données possédant de bonnes propriétés d'indexation.
3. Proposer une structure d'organisation des modèles de locuteurs afin de faciliter l'accès et la recherche à moindre coût face au problème incrémental.

Ces objectifs ont pour but d'améliorer la robustesse des systèmes d'indexation au sens de locuteurs lors du passage à l'échelle incrémentale et face aux difficultés d'acquisition et de représentation avec des quantités variantes et insuffisantes des données acoustiques.

Dans cette thèse nous traitons les problèmes de robustesse par les stratégies 1 et 3 mentionnées ci-dessus. La mise en oeuvre des techniques correspondantes s'intègre dans le cadre théorique de l'approche probabiliste pour l'indexation audio en se basant sur des outils de base tirés du domaine de la RAL décrite dans le chapitre suivant.

CHAPITRE 3

Segmentation audio : Détection des changements de locuteurs

L'objectif de ce chapitre est de donner des repères formels et méthodologiques sur la détection de ruptures dans un signal aléatoire. Nous introduisons d'abord le formalisme nécessaire, puis nous définissons le type d'approche étudié. Ensuite, une section est consacrée à la méthode de segmentation basée sur le test d'hypothèses bayésiennes. Enfin une présentation du Critère d'Inférence Bayésien (BIC) récemment utilisée afin d'améliorer la performance du regroupement des segments de locuteur suite à une détection de changement de locuteur à l'aide de test d'hypothèses bayésiennes.

3.1 Introduction

La détection et l'estimation des changements dans un signal consistent à localiser les instants de passage liés à des changements des caractéristiques du signal à un autre (changement de locuteur, changement de type : parole, musique, silence, etc.) [74]. Il peut s'agir par exemple, de détecter les moments de passage d'un locuteur à un autre dans une conversation, ou de localiser les transitions de parole à la musique dans un document audiovisuel, ces changements peuvent être transitoires ou instantanés.

Nous plaçons ce problème dans le cadre statistique, où les changements instantanés sont alors appelés *ruptures*. La détection de rupture est au carrefour entre plusieurs thématiques de la statistique inférentielle : le test d'hypothèses, la théorie de l'estimation, le contrôle de qualité, etc. Le contrôle de la qualité de produits manufacturés est

une application à l'origine de la détection de ruptures [76]. Les années 50 constituent une étape importante dans l'évolution de la formalisation du problème. En l'occurrence on peut citer [35] et [60] en relation avec les travaux de $Schwarz$ sur la détection automatique de rupture.

La détection de ruptures a toujours suscité un intérêt croissant auprès de la communauté scientifique dans celui des mathématiques appliquées et le domaine de traitement de signal. Cependant, la bibliographie sur le sujet est relativement vaste. Les ouvrages [13] et [4] reflètent respectivement l'aspect théorique et applicatif et constituent des références incontournables sur les méthodes de détection et d'estimation de ruptures. On peut mentionner l'existence de travaux spécialisés dans certains types de problèmes de détection de rupture, comme [19, 80] sur les approches bayésiennes, et [27, 23] qui proposent une étude sur les propriétés asymptotiques de test d'existence d'une rupture. En outre, la conception d'algorithmes statistiques de détection et de diagnostic de changements dans les signaux et systèmes dynamiques avec comme application, la segmentation automatique de signaux en vue de la reconnaissance consiste le principal thème de $M.Basseville$ [1] et met ces travaux [5] au centre d'interêt de notre travail.

Deux types de méthodes de détection de ruptures peuvent être distinguées : les méthodes séquentielles (généralement orientées surveillance) et les méthodes non-séquetielles (plus spécifiquement dédiées à l'analyse *a posteriori*). La différence fondamentale entre ces deux familles réside dans la contrainte du délai de détection. Il doit être aussi court que possible pour les méthodes séquentielles, d'où la nécessité d'une décision à chaque observation. Au contraire, les méthodes non-séquentielles (dans un but descriptif par exemple) ne sont pas forcément contraintes par l'urgence de la détection. Par conséquent, le traitement est effectué après la phase complète d'observations. Ces méthodes non-séquentielles sont évaluées par leurs taux de non-détection (dans le cas où le système ne détecte aucun changement) et leur taux de fausse alarme (le système détecte un changement sans avoir lieu réellement), l'appréciation des performances des méthodes séquentielles doit également tenir compte du délai de détection.

Notons toutefois que la frontière entre ces deux types de méthodes n'est pas aussi claire, dans la mesure où des méthodes séquentielles peuvent être utilisées sur des problèmes non-séquentiels, et vice-versa. C'est le cas notamment dant [12, 26, 1], où des méthodes de détection séquentielle de ruptures sont appliquées à la segmentation de signaux de parole.

Le champ d'application ouvert par la détection de ruptures s'est considérablement élargi. Les sujets rattachés à la surveillance sont de natures très diversifiées. En effet, celle-ci peut porter sur les infrastructures industrielles. Les réseaux de télécommunications, des grandes structures (ponts, immeubles, sous-sol d'une région, ...) pouvant être

[1] Ouvrages et publications de Michèle Basseville :

soumises à des vibrations ou aléas naturels, etc. Les signaux biomédicaux (électro-encéphalogrammes, électrocardiogrammes) ont fait l'objet de plusieurs applications de la détection de ruptures. Les applications aux signaux biomédicaux peuvent être orientées surveillance, quand il s'agit de veiller sur l'état physiologique d'un individu. Mais elles peuvent aussi être orientées analyse, lorsque le signal est segmenté en zones de stationnarité dans un but descriptif.

D'autres applications existent dans des domaines en pleine expansion en traitement de la parole, de l'image et de toutes les données multimédia en général. La détection de rupture peut être utilisée comme méthode de segmentation temporelle, elle consiste une étape qui précède la phase de classification et d'étiquetage du flux audio en segments. En outre, des travaux récents [6, 20, 24] se focalisent sur le problème de suivi de locuteur à l'aide des techniques de détection de rupture et la segmentation temporelle à l'aide de test d'hypothèses bayésiennes.

L'objectif de la première section est de donner des repères formels et méthodologiques sur la détection de rupture dans un signal aléatoire. Nous introduisons d'abord le formalisme nécessaire, puis nous définissons l'approche étudiée.

3.2 Détection et estimation de rupture : techniques existantes

3.2.1 Formulation de la détection et estimation de ruptures

Soit $Y = Y_1, ..., Y_n, ...$ un signal multidimensionnel à temps discret, à valeur dans $\Re^d$. Relativement à la tâche abordée dans ce travail, ce signal est celui des vecteurs de paramètres acoustiques, extraits des trames d'un signal sonore. On suppose qu'il existe des instants $r_1 \prec r_2 \prec ....$ tels que le signal est stationnaire sur chaque intervalle de temps $[r_i + 1, r_{i+1}]$, avec la convention $r_0 = 0$. L'homogénéité dans le segment, l'hypothèse à laquelle il a été fait allusion dans l'introduction, est ainsi modélisée par la stationnarité du signal. Les instants r_i sont ceux de changement de distribution stationnaire du signal, et sont appelés instant de ruptures. On notera y_i une observation de la variable Y_i.

La détection et l'estimation de ruptures consistent à localiser les éventuels instants de ruptures sur une séquence d'observations $y_1, ..., y_n$ du signal Y. Les critères de classification des problèmes de détection et estimation de ruptures sont multiples. Pour ne retenir que ceux qui nous semblent essentiels, on peut citer :

- Le caractère séquentiel (au fur et à mesure qu'on observe le signal) ou non-séquentiel (après la phase d'observation du signal) du traitement.
- La connaissance *a priori* sur les distributions des segments :

  - distributions des segments appartenant à une famille finie connue,
  - distributions des segments appartenant à une famille paramétrée, décrite par un paramètre inconnu,
  - aucune information *a priori* sur les distributions des segments.
- La connaissance *a priori* sur la distribution des instants de ruptures :
  - instants de ruptures aléatoires,
  - instants déterministes,

Dans ce travail, le problème de la détection et de l'estimation des ruptures est placé dans le cas où le traitement peut être séquentiel ou non, et les distributions des segments sont décrites par une famille paramétrés de densités $P(.|\theta); \theta \in \Theta \subset \Re^d$. Dans nos expériences, la famille des modèles de mélange de gaussiennes est utilisée. Dans certains cas, l'ensemble des distributions des segments est réduit à un nombre fini de distributions. Par exemple, la segmentation en locuteurs est réalisée à l'aide d'une représentation par simple gaussienne de segments de données acoustiques, puis par mélange de gaussiennes pour la classification.

Les sections (3.2.2 et 3.2.3) présentent respectivement quelques méthodes non-séquentielles de détection de ruptures particulièrement connues en Traitement du Signal.

3.2.2 Méthodes non-séquetielles

Une méthode non-séquentielle consiste en une détection et estimation des éventuelles ruptures après une phase complète d'observation du signal. On peut distinguer deux types de méthodes : d'une part les approches globales qui tentent d'estimer conjointement et de manière directe les paramètres de ruptures $(k, r_1, ...r_k)$ à partir des observations $y_1, ..., y_n$; et d'autre part les méthodes locales qui se proposent de se ramener localement au cas d'une seule éventuelle rupture et de traiter le signal au fur et à mesure. Les sous sections suivantes présentent des exemples de méthodes locales et globales.

3.2.2.1 Une approche non-séquentielle globale

Le signal $\{Y_t, t \geq 1\}$ et celui indicateur des instants de rupture représentent respectivement la partie observable et la partie cachée d'un signal plus complet. Les approches globales utilisent un modèle de la variable complète et tentent d'estimer les instants de rupture à partir des observations $y_1, ..., y_n$; du signal accessible.

Une approche globale permet de résoudre le problème de la détection et d'estimation de rupture, elle consiste, dans un premier temps, à effectuer un test d'hypothèse d'absence contre celle de présence de rupture. Dans un

second temps, si la décision du test est en faveur d'une présence de rupture, on entame une phase d'estimation des instants de rupture. Ci-après, les énoncés exacts des hypothèses de présence de rupture H_0, et d'absence de rupture H_1 :

- H_0 : Il existe $\theta_0^* \in \Theta$ tel que $Y_1, ..., Y_n$ suivent la même distribution de parmaètres θ_0^*.
- H_1 : Il existe un entier $k^* \geq 1, k^*$ instants $1 \leq r_1^* \prec ... \prec r_k^* \prec n$, et $(k^* + 1)$ paramètres de distribution $\theta_1^*, ..., \theta_{k+1}^*, \theta_i^* \neq \theta_{i+1}^*$, tels que sur chaque intervalle $[r_{i-1}^*, r_i^*]$, les variables $Y_{r_{i-1}^*+1}, ..., Y_{r_i^*}$ suivent la même distribution de paramètres θ_i^*. On adoptera la convention $r_0^* = 0$ et $r_{k^*+1}^* = n$.

L'étoile $(*)$ signifie qu'il s'agit de vrais paramètres de distribution, k^* représente le nombre de ruptures, r_i^* désigne un instant de rupture, et θ^* correspond à un paramètre de distribution d'un segment.

Les vrais paramètres n'étant pas fixés, les hypothèses H_0 et H_1 sont des types composites, justifiant généralement l'utilisation d'un test du rapport de vraisemblance. Deux cas sont alors envisagés, selon que l'on dispose ou non d'une distribution *a priori* sur les instants de ruptures.

En l'absence de distribution *a priori* sur les instants de ruptures

Le test du rapport de vraisemblance généralisé de H_0 contre H_1 est basé sur la statistique :

$$\frac{\sup_{(k,r_1,...,r_k)\in\tau,(\theta_1,...,\theta_{k+1})\in\Theta^{k+1}} p(y_1, ..., y_n|k, r_1, ..., r_k, \theta_1, ..., \theta_{k+1})}{\sup_{\theta\in\Theta} p(y_1, ..., y_n|\theta)} \tag{3.1}$$

où k représente le nombre du ruptures, r_i désigne le i^{me} instant de rupture, τ est l'ensemble des paramètres de ruptures $(k, r_1, ..., r_k)$ décrivant H_1, et θ_i correspond au paramètre de distribution du i^{me} segment. Le dénominateur de la statistique du rapport de vraisemblance généralisé (3.1) représente le maximum de vraisemblance sous H_0 (sachant qu'il n y a pas de rupture), et le numérateur est le maximum de vraisemblance sous l'hypothèse H_1 (sachant qu'il y a au moins une rupture). Dans la pratique les paramètres $(k, r_1, ..., r_k)$ qui réalisent le maximum du numérateur, sont pris comme estimateurs des paramètres du rupture (lorsque le test décide qu'il y a une rupture).

En général, on se place dans le cas d'une famille de distribution des segments pour laquelle les estimateurs du maximum de vraisemblance sont accessibles. Ce dernier peut alors être relativement obtenu facilement sous H_0.

Le maximum de vraisemblance sous H_1 est plus difficile à calculer. D'abord, il est clair que si τ est l'ensemble des paramètres de rupture possible, plus k est grand, plus on dispose de paramètres et plus le maximum de vraisemblance a tendance à augmenter, et l'optimum est atteint pour le plus grand nombre de ruptures $k = n - 1$ et

$r_i = i$. D'ou l'intérêt de restreindre le domaine des paramètres de ruptures à un ensemble τ. Il peut être construit en introduisant une contrainte (discrète) sur k, ou (indirectement) sur la longueur des segments (écarts $(r_{i+1} - r_i)$). Ces contraintes peuvent également se justifier par le souci d'avoir un nombre suffisant d'observations sur un segment, pour une bonne estimation du paramètre de distribution du segment.

En présence d'une distribution *a priori* sur les instants de ruptures

Les contraintes sur les paramètres de rupture $(k, r_1, ..., r_k)$ peuvent également s'exprimer par le biais d'une distribution *a priori*, ainsi le problème peut se palcer dans un cadre Bayésien. Si $p(k, r_1, ..., r_k|\theta_1, ..., \theta_k)$ dénote la densité *a priori* de $(k, r_1, ..., r_k)$ sous H_1, le rapport de vraisemblance généralisé la formule (3.1) devient :

$$\frac{\sup_{(k,r_1,...,r_k)\in\tau,(\theta_1,...,\theta_{k+1})\in\Theta^{k+1}} \{p(y_1, ..., y_n|k, r_1, ..., r_k, \theta_1, ..., \theta_{k+1}) * p(k, r_1, ..., r_k, \theta_1, ..., \theta_{k+1})\}}{\sup_{\theta\in\Theta} p(y_1, ..., y_n|\theta)} \quad (3.2)$$

Pour plus de détail sur les approches bayésiennes voir [19, 80].

En plus des difficultés mentionnées dans le cas d'absence de distribution sur les paramètres de ruptures, la maximisation de la vraisemblance sous H_1 est un problème d'optimisation avec des paramètre discrets $(k, r_1, ..., r_k)$ sur un ensemble dont la taille peut croître très rapidement en fonction du nombre d'observation n et selon la structure de τ. On peut par exemple utiliser des algorithmes stochastiques comme le recuit simulé [2, 3]. En revanche, ces techniques d'optimisation sont très gourmandes en temps de calcul. Cela compromet l'application et l'évaluation à grande échelle de ces méthodes.

Il existe un cas de figure très utilisé en indexation de documents sonores, pour lequel la maximisation de la vraisemblance sous H_1 est réalisée d'une manière . Il s'agit du cas où le signal $Y_t, t \geq 1$ est modélisé par la technique de Chaîne de Markov Cachée (ou partiellement observée). Un segment correspond au séjour dans un des états, et les distributions des segments sont décrites par une famille finie de K densités $\{p(.\theta), \theta \in \Theta = \underline{\theta_1}, ..., \underline{\theta_k} \subset \Re^m\}$.

La modélisation par un modèle de Markov Caché commence à être utilisé en indexation de signaux sonores, quand le signal est constitué de plages sonores appartenant à des classes de sons connues, le segment représentant un intervalle de séjour dans une des classes sonores.

Á ce titre, on peut faire référence au récent travaux de [30] pour la segmentation de journaux télévisés en plage de parole, de musique et de silence, et ceux de [78, 52] pour la segmentation d'une conversation en plages (intervalle de temps) propre à un seul locuteur.

Dans le cas d'un Modèle de Markov Caché, la densité des paramètres de rupture $(k, r_1, ..., r_{k+1})$ connaissant la succession d'états $(\theta_1, ..., \theta_{k+1})$, noté $p(k, r_1, ..., r_{k+1}|\theta_1, ..., \theta_{k+1})$ est donnée par :

$$p(k, r_1, ..., r_{k+1}|\theta_1, ..., \theta_{k+1}) = p(p_{1,1})^{r-1} * \Pi_{j=2}^{k+1} p_{j,j-1}(p_{j,j})^{r_j - r_{j-1} - 1} \tag{3.3}$$

où p_i est la probabilité *a priori* de l'état associé à la distribution de paramètre θ_i, et $p_{i,j}$ est la probabilité de transition de la distribution de paramètre θ_i à celle de paramètre θ_j. Dans ce cas, les paramètres de rupture qui maximisent la vraisemblance sous H_1, peuvent être obtenus au moyen de l'algorithme de Viterbi [83].

Dans le cas d'indexation par locuteur, les changements de rôle de parole des intérvenants dans le document audio provoque des ruptures, ces rupture sont appelées changement de locuteurs, seul une simple gaussiennes peut servir pour la représentation des segments de test d'hypothèses (voir section 3.3.3).

3.2.2.2 Une approche non-séquentielle locale

Pour remédier aux difficultés de mise en oeuvre des approches globales, surtout quand la famille de distributions de segments est complexe, une approche consiste à se ramener à un traitement raisonnable. Généralement, on teste l'existence d'une rupture à chaque instant. La détection en un instant t peut être basée, aussi bien sur des observations au voisinage de t, que sur toute la séquence d'observations $y_1, ..., y_n$.

La plupart des méthodes commencent par calculer, à chaque instant t, un *indice* (statistique) de rupture, puis à en extraire un *critère* de décision. La décision de présence d'une rupture en un instant est prise par la comparaison du critère à un seuil donné. Ce principe de calcul en deux temps (indice puis critère) est à la base de nombreuses méthodes de segmentation de signaux. Des travaux en segmentation de documents sonores peuvent être trouvés dans [54, 6, 25]. Des résultats théoriques sur cette approche sont présentés dans [8], traitant des ruptures correspondant à un changement de moyenne dans une famille gaussienne de distribution des segments.

La propriété fondamentale requise pour un indice de rupture est qu'elle prenne ses plus grandes valeurs aux instants de rupture. On utilise alors généralement comme indice, une statistique de test d'existence d'une rupture à chaque instant, sur une fenêtre de contexte autour de l'instant considéré et ne contenaient au plus qu'une seule rupture ; par exemple une fenêtre de contexte, supposée de taille égale de part et d'autre de l'instant t, c'est-à-dire w observation $y_{t-w+1}, ..., y_t$. pour *le passé* et w autres observations $y_{t+1}, ..., y_{t+w}$. pour *le futur*, comme le schématise la figure (3.1).

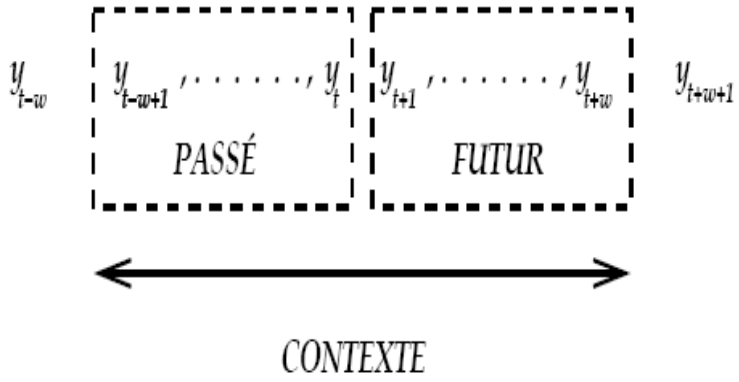


Figure 3.1 – Position des fenêtres de contexte passé et futur d'un instant t pour le calcul d'une statistique de détection de rupture

Une des statistiques les plus utilisées est le rapport de vraisemblance généralisé du test d'existence de rupture en un instant t fixé. Cette statistique a pour expression en t :

$$\frac{\sup_{(\theta_-)\in\Theta} p(y_{t-w+1}, ..., y_t|\theta_-)p(y_{t+1}, ..., y_{t+w}|\theta_+)}{\sup_{\theta_0\in\Theta} p(y_{t-w+1}, ..., y_{t+w}|\theta_0)} \tag{3.4}$$

La figure (3.2) illustre un exemple de rapport de vraisemblance généralisé [74]. Les maxima locaux de la statistique sont au voisinage des vrais instants de ruptures, qui se produisent toutes les 5 secondes. Cependant, cet exemple met clairement en évidence le fait que l'extraction de ces maxima locaux ne peut pas être obtenue par un seuillage direct de l'indice.

Généralement, la statistique de détection de rupture présente une allure très irrégulière, avec de nombreux maxima locaux rapprochés. La détection de changement de la satatistique se résume à détecter les pics observables de la courbe donnée par le rapport de vrassemblance des données. Par la suite, le critère de décision doit être choisi dans le but de réduire le taux des fausses alarmes et le taux de fausses détections.

Plusieurs méthodes d'extraction de critère peuvent être trouvées dans la littérature. Cependant, elles sont souvent empiriques et dépendantes de seuils et/ou de lissage intermédiaires plus ou moins explicitement formulés.

Le critère de décision proposé par [74] permet d'éviter les seuils et les lissages intermédiaires. L'approche qui y est décrite pour l'extraction du critère de décision, est basée sur la notion de hauteur d'un maximum, empruntée à l'optimisation stochastique. Cette méthode est évaluée sur une tâche de détection de rupture sur un signal obtenu

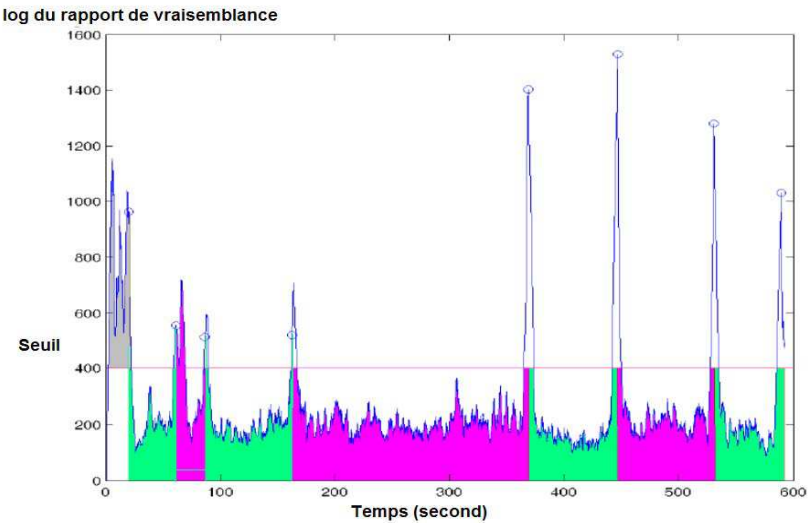


Figure 3.2 – Logarithme du rapport de vraisemblance généralisé instantané, sur une fenêtre de contexte glissante autour de l'instant courant. Les ruptures sont générées en concaténant des morceaux de parole de 5 secondes. Les fenêtres du passé et du futur ont une durée $2sec$ à $4sec$. Les vecteurs acoustiques utilisés sont constitués des 26 paramètres des Mel Frequency Cepstral Coefficients (MFCC), et les modèles du passé et du futur sont gaussiens

par concaténation de segments réels de parole et de musique.

3.2.3 Méthodes séquentielles

Une méthode séquentielle de détection de ruptures est une méthode qui, à chaque instant d'observation, rend une décision d'absence ou de présence d'une rupture avec éventuellement un certain retard, ou délai intrinsèque. Les méthodes séquentielles sont plus adaptées à certains problèmes de surveillance, bien qu'elles s'appliquent aussi à d'autres problèmes de détection de ruptures. Le cadre théorique de ces méthodes est celui des tests séquentiels présenté par exemple dans [77, 4].

La méthode séquentielle du traitement n'entre pas dans le cadre de recherche de ce travail. Nous allons mentionner le principe de deux algorithmes séquentiels de détection de ruptures, très connus dans la littérature : l'algorithme de Brandt [12] et l'algorithme de la divergence [1, 4, 26].

Soit $\widehat{r}_0$ le dernier instant de rupture estimé. Ces algorithmes testent à partir de l'instant $\widehat{r}_0$ l'existence d'une rupture à chaque nouvel instant d'observation. Si le test décide à l'insatnt t qu'il y a eu une rupture estimé, la

même procédure de détection d'une rupture est réalisée à partir de $\widehat{r}_1$, et ainsi de suite.

3.3 Méthode proposée

3.3.1 Segmentation progressive à l'aide de test d'hypothèses Bayésiennes

Le problème du partitionnement non-supervisé en segments homogènes au sens du locuteur fait partie de la classe générale des problèmes de classification, pour lesquels ont été proposées des approches statistiques, issues de la théorie de l'information et la théorie neuronale. Le problème complet est divisé en trois sous-problèmes indépendants, à résoudre conjointement : L'estimation des modèles de locuteurs, le partitionnement en locuteurs, et la détermination du nombre de locuteurs. Les approches itératives de type *restauration-maximisation* ont montré de bonnes performances, et leurs évolutions récentes abordent le traitement efficace de la détermination du nombre de classes.

Dans l'application que nous visons, le coût de calcul doit être une fonction linéaire de la longueur du document, où la structure des documents doit être extraite en temps streaming plutôt que sur un lot de documents s (archives). Les prochains paragraphes détaillent la procédure proposée. La solution comporte deux phases :

1. Détecter un changement éventuel de locuteur ;
2. Etiquetage : un nouveau segment homogène au sens du locuteur est généré
   - Examiner si ces données correspondent à un locuteur déjà connu de la base de données,
   - Sinon : créer une nouvelle entrée.

3.3.2 Notions et définitions (Test de rapport de vraisemblance)

Le rapport de vraisemblance entre un paramètre θ_1 de l'hypothèse H_1 et un paramètre θ_0 de l'hypothèse H_0 joue un rôle important dans la théorie statistique des tests. Il a pour expression :

$$\frac{p(y_1, ..., y_n|\theta_1)}{p(y_1, ..., y_n|\theta_0)} \tag{3.5}$$

Les exemples 1 et 2 énoncent des tests basés sur le rapport de vraisemblance. Le premier porte sur des hypothèses simples, et le second est une généralisation du premier aux hypothèses composites.

Exemple 1 (Test du rapport de vraisemblance)

Le test du rapport de vraisemblance (pour des hypothèses) de $H_0 = \theta^* = \theta_0$ contre $H_1 = \theta^* = \theta_1$, a pour statistique le rapport de vraisemblance est défini pour un seuil λ par sa région de rejet

$$\left\{ \frac{p(y_1, ..., y_n|\theta_1)}{p(y_1, ..., y_n|\theta_0)} \succ \lambda \right\} \tag{3.6}$$

Exemple 2 (Test du rapport de vraisemblance généralisé)

Le test du rapport de vraisemblance, défini ci-dessus pour des hypothèses simples, se généralise aux cas d'hypothèses composites $H_0 = \theta^* \in \Theta_0$ et $H_1 = \theta^* \in \Theta_1$, où Θ_0 et Θ_1 sont des sous-ensembles de Θ. Ce test est basé sur la statistique du rapport de vraisemblance généralisé qui a pour expression :

$$\frac{\sup_{\theta_1 \in \Theta_1} p(y_1, ..., y_n|\theta_1)}{\sup_{\theta_0 \in \Theta_0} p(y_1, ..., y_n|\theta_0)} \tag{3.7}$$

Le rapport de vraisemblance généralisé est le rapport entre le maximum de vraisemblance sous H_1 et celui sous H_0. C'est la statistique la plus utilisée dans le cas d'hypothèse composite. Le test basé sur cette statistique a pour région de rejet relatif à un seuil λ

$$\left\{ \frac{\sup_{\theta_1 \in \Theta_1} p(y_1, ..., y_n|\theta_1)}{\sup_{\theta_0 \in \Theta_0} p(y_1, ..., y_n|\theta_0)} \succ \lambda \right\} \tag{3.8}$$

Cette section a permis de définir quelques notions de base très utilisées en théorie statistique des tests compte tenu du rôle important que jouent les tests statistiques dans la phase de détection de changement de locuteur.

3.3.3 Détection de changement de locuteur à l'aide de test d'hypothèses Bayésien

Il s'agit de mettre en compétition deux hypothèses H_0 et H_1 et comparer la vraisemblance de ces deux probabilités. Le but est de calculer le rapport des deux probabilités et d'estimer le modèle le plus probable qui vérifie une des hypothèses supposées. Le test d'hypothèses Bayésien est utilisé pour la détection de changement de locuteur. Pour cela, on désigne par D_i tel que $i \in \{1, 2\}$, les données MFCC d'une fenêtre de taille (2 seconds) extraits juste avant D_1 et après D_2 l'instant en cours t (voir figure 3.3), notons aussi par $D_{1\cup 2}$ le bloc de l'union.

A la différence des tests d'hypothèses usuels (Neyman-Pearson), l'approche bayésienne est employée dans ce travail pour la comparaison des modèles [19]. Si l'objectif est le même, le point de vue paraît cependant plus

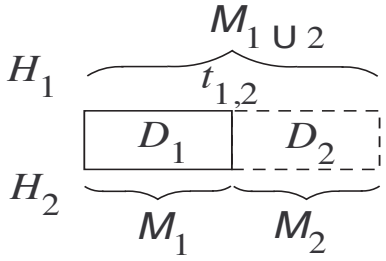


Figure 3.3 – Test une hypothèse contre deux hypothèses de fenêtres de données D_1, D_2 et $D_1 \cup D_2$

conforme à l'intuition, puisqu'il s'agit de comparer les probabilités *a posteriori* des deux hypothèses en concurrence sur la lumière des données. Notons par M_m le modèle de locuteur à estimer sur l'ensemble de données D_m, où la notation regroupe les trois cas $m = 1, 2$ et $1 \cup 2$ (voir figure. 3.3). Ce rapport des probabilités *a posteriori* (connu sous le nom de facteur de Bayes) s'exprime comme suit :

$$\frac{p(H_1|D_1, D_2)}{p(H_0|D_1, D_2)} = \frac{p(D_1|M_1)p(D_2|M_2)p(H_1)}{p(D_1 \cup D_2|M_{1\cup 2})p(H_0)} \tag{3.9}$$

La prise de décision est ensuite obtenue par le critère de maximum de vraisemblance a posteriori. Dans ce formalisme, la mesure entre l'information disponible et les différentes hypothèses est représentée par une distribution de probabilité conditionnelle. Celle-ci peut être représentée comme fréquence relative d'apparition des différentes classes. Ici, la mise à jour des informations (modélisées par des distributions de probabilité) se fait à l'aide du théorème de Bayes.

L'évaluation de rapport de ces deux probabilités conditionnelles en utilisant la loi de bayes, revient à calculer trois densité de probabilités des deux blocs ainsi que leurs union. Fait que le changement ne peut être détectée que si l'hypothèse d'avoir le même locuteur dans les deux blocs est fausse, et que la distribution de l'union des deux blocs ne suit pas une distribution gaussienne, (voir la formule. 3.9).

Afin d'optimiser la performance de la détection de changement de locuteur à l'aide de test d'hypothèses Bayésien, nous optons pour l'utilisation du critère *Bayésian Inférence Critéria* (BIC). Ce critère offre une estimation du meilleur (vrai) modèle dont les caractéristiques représentent les données d'observation D_i. Pour plus de précision nous allons présenter dans la prochaine section la technique de segmentation basée sur le critère de sélection Bayésien BIC.

3.3.4 BIC Segmentation

La sélection des modèles est un problème bien connu en statistique. Lorsque le modèle est fixé, la théorie de l'information fournit un cadre rigoureux pour l'élaboration d'estimateurs performants. Mais dans un grand nombre de situations, les connaissances *a priori* sur les des données ne permettent pas de déterminer un modèle unique pour réaliser l'inférence. C'est pourquoi depuis la fin des années 70 les méthodes pour la sélection de modèles à partir des données ont été développées.

Nous nous intéressons ici au critère BIC qui se place dans le contexte bayésien de sélection de modèles. Certains points de sa construction et de son interprétation sont régulièrement omis dans les démonstrations proposées dans la littérature [47]. Il est bien connu que le critère BIC est une approximation du calcul de la vraisemblance des données conditionnellement au modèle fixé. Cependant les résultats théoriques utilisés sont souvent peu explicités, tout comme les hypothèses nécessaires à leurs applications. Par ailleurs, l'interprétation de BIC et la notion "consistance pour la dimension" ne sont pas toujours très claires pour les utilisateurs.

3.3.4.1 Construction du critère BIC

Dans cette partie, nous présentons la construction du critère BIC. Pour cela, nous nous appuyons sur la présentation proposée par *Raftery A. E* [62].

Soit n échantillons $X = (X_1, ..., X_n)$ de variables indépendantes de densité inconnue f. L'objectif est d'estimer f, pour cela, nous définissons d'abord une collection finie de modèles $\{M_1, ..., M_m\}$. Un modèle M_i correspond à une densité g_{M_i} de paramètres θ_i. Il s'agit de choisir un modèle parmi cette collection de modèles.

Le critère BIC se place dans un contexte de Bayes : θ_i et M_i sont vues comme des variables aléatoires et sont munies d'une distribution *a priori*. La distribution *a priori* sur M_i est notée $P(M_i)$.

Pour un modèle Mi donné, la distribution a priori du paramètre μ_i est notée $P(\mu_{ij}, M_i)$. L'avantage d'une telle approche est qu'elle permet de prendre en compte des informations que peuvent détenir l'utilisateur, en donnant à certains modèles un poids plus important. Cependant, la distribution a priori posée sur les modèles M_i est souvent non informative (uniforme) et nous verrons par des considérations assymptotiques que la distribution a priori des μ_i n'intervient pas dans la forme du critère BIC. Ce dernier permet de sélectionner le modèle M_i qui maximise la

probabilité a posteriori $P(M_i|X)$:

$$M_{BIC} = \arg\max_{M_i} P(M_i|X_i) \tag{3.10}$$

En ce sens BIC cherche à électionner le modèle le plus vraisemblable au vu des données. D'après la formule de Bayes, $P(M_i, X)$:

$$P(M_i|X) = \frac{(X|M_i)P(M)}{P(X)} \tag{3.11}$$

Nous supposons dans toute la suite que la loi *a priori* des modèles M_i est non informative :

$P(M_1) = P(M_2) = ... = P(M_m)$

Ainsi, aucun modèle n'est privilégié. Sous cette hypothèse et d'après (formule 3.10 et 3.11), la recherche du meilleur modèle ne nécessite que le calcul de la distribution $P(X|M_i)$. Ce calcul s'obtient par l'intégration de la distribution jointe du vecteur des paramètres θ_i et les données X, $P(X, \theta_i|M_i)$, sur toutes les valeurs de θ_i :

$$P(X|M_i) = \int_{\theta_i} P(X, |\theta_i|M_i)d\theta_i = \int_{\theta_i} g_{M_i}(X, \theta_i)P(X, |\theta_i|M_i)d\theta_i \tag{3.12}$$

où $g_{M_i}(X, \theta_i)$ est la vraisemblance correspondant au modèle M_i de paramètres θ_i :

$$g_{M_i}(X, \theta_i) = P(X, \theta_i, M_i) \tag{3.13}$$

Cet intégral est redéfini sous la forme :

$$P(X|M_i) = \int_{\theta_i} \exp^{g(\theta_i)} d\theta_i; g(\theta_i) = log(g_{M_i}(X, |\theta_i)P(\theta_i|M_i))) \tag{3.14}$$

La probabilité $P(X|M_i)$ est appelée *vraisemblance intégrée* pour le modèle M_i. Le calcul exact de cette probabilité est compliqué. En utilisant la méthode d'approximation de *Laplace* où le modèle BIC_i est donné par :

$$BIC_i \approx -2\log(g_{M_i}(X|\widehat{\theta_i})) - \frac{K_i}{2}\log(n) \tag{3.15}$$

C'est de cette approximation que le critère BIC est issu. Plus précisément, pour le modèle M_i correspond à l'approximation de $-2\log P(X|M_i)$ et donc défini par :

$$BIC_i = -2\log(g_{M_i}(X|\widehat{\theta_i})) + K_i\log(n) \tag{3.16}$$

Le modèle sélectionné par ce critère est :

$$M_{BIC} = \arg\min_{M_i} BIC_i \tag{3.17}$$

Dans la suite, la phase de détection de changement de locuteur est réalisée à l'aide de test d'hypothèses Bayésien. Cependant, la sélection du meilleur modèle à l'aide du critère BIC permet d'améliorer l'expression (3.9) en utilisant l'approximation donnée par l'expression (3.16).

3.3.5 Interprétation du critère BIC

L'une des difficultés du critère BIC est son interprétation. La question est la suivante : quel est le modèle que nous cherchons à sélectionner par le critère BIC ? À ce niveau, les notions de probabilité *a priori* ou *a posteriori* d'un modèle sont peu explicites et ne donnent pas une idée intuitive de ce que BIC considère le "meilleur" modèle. Les considérations asymptotiques présentées ici vont nous permettre d'interpréter cette notion de meilleur modèle et de déterminer la probabilité a posteriori d'un modèle. Cette interprétation nous permettra aussi de discuter la nécessité de l'hypothèse d'appartenance du vrai modèle à la liste des modèles considérés.

3.3.5.1 Le "quasi-vrai" modèle

Nous reprenons ici la remarquable présentation de cette notion proposée par *Burnham K.* [14]. Nous supposons les connaissances du critère AIC et sa démonstration acquise.

Rappelons que la densité à estimer est f. On suppose que les m modèles $M_1, ..., M_m$ sont emboîtés. La pseudo distance de Kullback-Leibler (appelée par la suite distance) entre deux densités f et g est définie par :

$$d_{KL}(f, g) = \int_{\Omega} \log(\frac{f(x)}{g(x)}) f(x) dx \tag{3.18}$$

Par abus de notation, on définit la distance de KL de f au modèle M_i par :

$$d_{KL}(f, M_i) = inf_{\theta_i} d_{KL}(f, g_{M_i}(., \theta_i)) \tag{3.19}$$

Puisque les modèles sont emboîtés, la distance KL est une fonction décroissante de la dimension K_i. On note M_t le modèle à partir duquel cette distance ne diminue plus. Du point de vue de la distance KL, M_t doit être favorisé parmi les sous-modèles $M_i, i = 1, ..., t-1$ puisqu'il est plus proche de f. Par ailleurs, M_t doit aussi

être préféré à tous les modèles d'ordre supérieurs $M_i, i = t+1, ..., m$ puisqu'ils sont plus compliqués que M_t sans pour autant être plus proches de f. Nous allons montrer que le critère BIC est consistant pour ce modèle particulier, désigné comme le meilleur modèle appelé 'le quasi-vrai' [14]. Pour n supposé grand, on s'intéresse à la différence :

$$BIC_i - BIC_t, i \neq t$$

Premier cas : $i \prec t$

d'après la formule (3.16), on a :

$$BIC_i - BIC_t = -2\log(g_{M_i}(X, \widehat{\theta}_i)) + 2\log(g_{M_i}(X, \widehat{\theta}_t)) + (K_i - K_t)log(n) \quad (3.20)$$

$$= 2n\left[-\frac{1}{n}\sum_{k=1}^{n}\log(g_{M_i}(x_k, \widehat{\theta}_i)) - \frac{1}{n}\sum_{k=1}^{n}\log(g_{M_t}(x_k, \widehat{\theta}_t))\right] + (K_i - K_t)log(n) \quad (3.21)$$

$$= 2n\left[-\frac{1}{n}\sum_{k=1}^{n}\log(\frac{f(x_k)}{g_{M_i}(x_k, \widehat{\theta}_i)}) - \frac{1}{n}\sum_{k=1}^{n}\log(\frac{f(x_k)}{g_{M_t}(x_k, \widehat{\theta}_t)})\right] + (K_i - K_t)log(n) \quad (3.22)$$

Les deux dernières sommes sont des estimateurs consistants des quantités $d_{KL}(f, M_i)$ et $d_{KL}(f, M_t)$, voir [66] respectivement, pour n grand, on a donc :

$$BIC_i - BIC_t \approx 2n\left[d_{KL}(f, M_i) - d_{KL}(f, M_t)\right] + (K_i - K_t)log(n) \quad (3.23)$$

Deuxième cas : $i \succ t$

Dans ce cas, on reconnaît dans le terme $2\log(g_{M_i}(X, \widehat{\theta}_i)) - 2\log(g_{M_i}(X, \widehat{\theta}_t))$ la statistique du test du rapport de vraisemblance pour deux modèles emboîtés, qui sont sous l'hypothèse H_0 suit asymptotiquement une loi du Chi-2 à $(K_i - K_t)$ degrés de liberté. On a donc :

$$BIC_i - BIC_t \approx -\chi^2_{(K_i - K_t)} + (K_i - K_t)log(n) \quad (3.24)$$

Dans cette équation le second terme qui domine et tend vers l'infini avec n, les modèles $M_i, i = t+1, ..., m$ sont eux aussi disqualifiés. Le terme en $log(n)$ joue donc un rôle fondamental : il assure que le critère BIC permet de converger vers le quasi-vrai modèle. C'est cette convergence vers le modèle quasi-vrai, même s'il est emboîté dans un modèle plus général, que l'on appelle la consistance pour la dimension.

Il nous est maintenant possible d'interpréter clairement ce que l'on entend par probabilité *a posteriori* du modèle M_i, elle s'estime à partir des différences $\Delta BIC_i = BIC_i - BIC_{min}$, où BIC_{min} désigne la plus petite valeur observée de BIC sur les m modèles. On a :

$$P(M_i|X) = \frac{\exp(-\frac{1}{2}\Delta BIC_i)}{\sum_{t=1}^{m} \exp(-\frac{1}{2}\Delta BIC_t)} \tag{3.25}$$

Cette probabilité tend vers 1 pour le modèle quasi-vrai lorsque n tend vers l'infini, et vers 0 pour tous les autres. Selon des considérations précédentes, nous pouvons définir cette probabilité comme la probabilité que M_i soit le modèle quasi-vrai de la liste considérée.

3.3.5.2 La performance du critère de sélection de "vrai" modèle

La question de savoir si le vrai modèle ayant engendré les données doit apparaître ou non dans la liste des modèles considérés. Cette hypothèse demeure longtemps en suspens dans la littérature consacrée au critère BIC, bien que cette hypothèse n'apparaisse nulle part comme nécessaire dans la construction du critère BIC. C'est pourquoi certains auteurs ont choisi de poser cette hypothèse sans justifier son utilité [73, 62]. La partie précédente résout de manière simple ce dilemme : le critère BIC assure la convergence en probabilitée vers le "quasi-vrai" modèle lorsqu'il est unique (ce qui est vrai dans la grande majorité des cas).

Néanmoins, le "quasi-vrai" modèle peut être très éloigné (au sens de la distance KL) du vrai modèle. Ainsi, une probabilité *a posteriori* élevée, aussi proche de 1 soit elle, ne justifie pas le choix du modèle retenu comme le vrai modèle. Pour pouvoir garantir que le modèle choisi est le vrai modèle, il faut pouvoir garantir que ce dernier fait partie de la liste $M_1; ..., M_m$. C'est dans ce cas l'hypothèse d'appartenance à la liste $M_1, ..., M_m$ du vrai modèle est nécessaire.

3.3.6 Conclusion

La détection de changement de locuteur permet de segmenter le document audio en deux classes (locuteur/locuteur-différent). Pour chaque fenêtre de $(20ms)$ un vecteur de MFCC est extrait de dimension $d = 13$, plus le dérivé temporelle entre les deux vecteurs de son voisinage. Le test d'hypothèses Bayésiennes utilise deux blocs de données D_1 et D_2 (en pratique chaque bloc représente $4sec$) et leurs union D_{1U2}. Les données sont représentées par une simple gaussienne, le processus est incrémental, c'est-à-dire que la détection de changement est réalisée directement lors de la lecture de fichier audio avec seulement un décalage de la taille d'un bloc de données. Il est important de noter que grâce à la souplesse offerte par la représentation gaussienne simple en utilisant un rapport

de vraisemblance de données le coût de calcul est très satisfaisant.

Comme conclusion ce chapitre a donné un aperçu général des problèmes de détection et d'estimation de ruptures sur un signal aléatoire à temps discret, et ce, afin de situer le cadre formel couvert par cette étude. Généralement on distingue deux types de méthodes : les algorithmes séquentiels et ceux non-séquentiels. Les applications visées par ce travail ne sont pas contraintes à un traitement séquentiel. Ainsi, l'accent a été mis sur les méthodes non-séquentielles.

De plus, différentes techniques ont été proposées dans la littérature, leurs objectifs est d'améliorer la performance de regroupement des segments après une première phase de détection de rupture. Le critère BIC est présenté comme une technique de validation, améliorant à la fois la performance de détection de rupture ainsi la sélection de modèle correspond au segment de données de test.

Le but du prochain travail est de proposer une structure de modèles de locuteurs fiable et efficace permettant la gestion d'un grand volume de données ou un flux en continu audio.
Nous présenterons dans la suite les techniques de mesure de similarité entre modèles de mélange de gaussiennes. La similarité entre modèles locuteurs permet ainsi de répartir les modèles de locuteurs. Cependant, la mesure de similarité entre modèles de mélange de gaussiennes est utilisée comme outil d'organisation arborescente.

PARTIE II

Vers une structuration hiérarchique des documents audio au sens du locuteur adaptée au problème incrémental

CHAPITRE 4

Techniques d'indexation et d'organisation des modèles de locuteurs

Le but de ce chapitre est de présenter un état de l'art des techniques de mesure de similarité entre les modèles de mélange de gaussiennes, en particulier, nous focaliserons notre étude sur les techniques de mesure de similarité sans avoir recours aux données d'apprentissage. La divergence de Kullback-Leibler (KL) est au centre de nos intérêts car elle représente un outil de calcul de divergence entre MMG efficaces et peu coûteux. Dans ce travail la mesure de similarité à l'aide de la divergence de KL est la clé du système d'identification du locuteur. En outre, le développement d'autres approximations de KL permettant de réduire le coût d'identification entre MMG de locuteur, avec l'objective de garder une meilleure performance d'identification.

Par la suite, nous allons présenter des méthodes de création des structures binaire ou n'aire des modèles de locuteur. Les éléments constituants l'arbre de recherche sont des MMG de locuteurs ou un ensemble de MMG groupés selon un critère de similarité. La complexité linéaire devient logarithmique en choisissant une structure arborescente et un nombre réduit des descripteurs utilisés.

4.1 Etat de l'art des techniques de mesure de similarité entre MMG

Souvent l'estimation imparfaite du modèle de locuteur, dû au volume restreint des données d'apprentissage implique que le cadre de la théorie bayésienne ne soit pas strictement respecté. Un seuil unique de décision mène généralement à des performances non optimales. Des biais provenant des disparités entre les contextes d'apprentissage et de test apparaissent dans les valeurs du rapport de vraisemblance.

Le contexte d'un ensemble de données, qu'il soit destiné à l'apprentissage ou pour une requête, regroupe les conditions d'acquisition du signal (bruits ambiants, prise de son, transmission), le contenu linguistique du message et l'état du locuteur. Ces contextes varient fortement avec les applications visées, dans les applications de RAL par téléphonie (terrestre et/ou mobile) et en mode dépendant du texte.

Les données d'apprentissage sont souvent réutilisés pour le test d'identification. En revanche, le volume restreint des données d'apprentissage ou d'identification affecte généralement la performance de l'identification dans le mode indépendant de texte. Il existe de nouvelles techniques de comparaison basées sur une mesure de similarité ou de dissimilarité selon leurs utilisation. Cette technique permet d'extraire une information entre deux modèles en utilisant que les paramètres de modèle (i.e le vecteur des moyennes, la matrice des covariance). Cette mesure est moins coûteuse que la mesure de score de vraisemblance entre les données et le modèle requête. Dans le cas des MMG, la divergence de Kullback-Leibler modifiée KL_m(voir équation. 4.5) donne des résultats satisfaisants à moindre coût.

Plusieurs techniques d'estimation de modèle permettent de réduire l'impact de la variabilité et la taille des segments des descripteurs utilisés dans le module d'apprentissage et la reconnaissance. Le MMG Universel Background Model *UBM-MMG* consiste une des solutions permettant d'offrir des informations *a priori* sur l'espace des locuteurs afin de générer un modèle unique pour chaque entraînement utilisant EM avec les données du même locuteur [64].

4.2 Divergence de Kullback-Leibler modifiée

La divergence de KL est généralement utilisée comme une mesure de similarité ou une mesure de distance entre deux densités de probabilité. En particulier, cette mesure aide à interpréter certaines caractéristiques d'une distribution par rapport à l'autre. En outre, ces divergences sont directement liées au rapport de vraisemblance, de part leur définition mathématique.

Dans la suite de ce travail, nous utilisons l'expression symétrique de la divergence de Kullback-Leibler comme une mesure de distance entre deux MMG de locuteurs.

Soit p_1 et p_2 deux distributions, la divergence de Kullback-Leibler est donnée par :

$$KL(p_1, p_2) = \int_{-\infty}^{+\infty} p_1(x) log(\frac{p_1(x)}{p_2(x)}) dx \tag{4.1}$$

Il est facile de remarquer que l'expression de la divergence de Kullback-Leibler ci-dessus est non symétrique. Afin de l'utiliser comme distance, la forme symétrique s'impose ainsi, la formule devient :

$$d_{KL}(p_1||p_2) = \frac{[KL(p_1, p_2) + KL(p_2, p_1)]}{2} \tag{4.2}$$

Dans le cas d'une distribution gaussienne simple, la divergence de KL a pour expression :

$$KL(p_1||p_2) = \frac{(\Sigma_1^2 - \Sigma_2^2)^2 + (\mu_1^2 - \mu_2^2)(\Sigma_1^2 - \Sigma_2^2)^2}{4\Sigma_1^2\Sigma_2^2} \tag{4.3}$$

Tel que, Σ_i et μ_i représentent successivement la covariance et la moyenne de la distribution p_i. Cette expression est définie pour une simple gaussienne, dans la suite, nous allons présenter la formule de la divergence de Kullback-Leiblier dans le cas d'un mélange de gaussiennes décrite dans [32] :

Soit f un mélange de gaussiennes de k composantes de gaussiennes de dimension d :

$$f(y) = \sum_{i=1}^{k} \alpha_i N(y_i; \mu_i; \Sigma_i) = \sum_{i=1}^{k} \alpha_i f_i(y_i) \tag{4.4}$$

La mesure de divergence entre les deux mélanges de gaussiennes $f = \sum_{i=1}^{k} \alpha_i f_i$ et $g = \sum_{i=1}^{k} \beta_i g_i$ est donnée par :

$$d(f, g) = \sum_{i=1}^{k} \alpha_i \min_{j=1}^{k} KL(f_i||g_j) \tag{4.5}$$

avec i et j les index des mélanges de gaussiennes, la valeur de similarité du couple (i, j) est calculée comme suit :

$$D_{KL}(i, j) = \sum_{l=1}^{k} \alpha_l \min_{m=1}^{k} KL(i_l, j_m) \tag{4.6}$$

D'après l'equation 4.5 la divergence de KL_m entre deux MMG (*i.e :* f et g) est n'est d'autre que la somme entre chaque composante i et j pondéré par le poid des composantes de MMG de f.

En pratique, chaque terme est considéré comme une divergence de KL entre deux distributions gaussiennes N_1 et N_2 tel que : $N_1(\mu_1, \Sigma_1)$ and $N_2(\mu_2, \Sigma_2)$

$$KL(N_1, N_2) = \frac{1}{2}(log\frac{|\Sigma_2|}{|\Sigma_1|} + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T\Sigma_2^{-1}(\mu_1 - \mu_2) - d) \tag{4.7}$$

4.3 Structure arborescente basée sur un regroupement des MMG

Dans cette section, nous allons examiner quelques problèmes liés à la construction d'une structure arborescente à l'aide du calcul de similarité entre les modèles de locuteurs.

Deux méthodes vont être présentées dont le but est de créer une structure arborescente à l'aide d'une mesure de similarité précédemment définie entre modèles de mélange de gaussiennes. La première méthode est basée sur la mesure des scores du maximum de vraisemblance comme facteur d'organisation des modèles par similarité. La deuxième méthode proposée permet de calculer les distances entre les modèles de locuteurs sans avoir recours aux données. Il s'agit de calculer la divergence de Kullback-Leibler entre modèles de mélanges de gaussiennes des locuteurs inscrits dans la base de données.

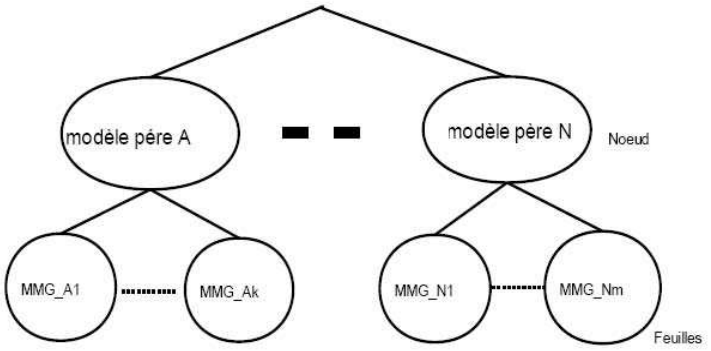


Figure 4.1 – Exemple de structure arborescente des modèles MMG.

4.3.1 Première expérience avec un regroupement du modèle

La construction d'une structure de données adéquate aux grands volumes de données exige un travail rigoureux et de réflexion sur sa conception, sur sa mise à jours sur les outils d'exploration et de navigation. L'idée est de pouvoir regrouper les locuteurs par un critère de similarité afin de faire en premier lieu, des recherches par "vue" ou groupe et en deuxième une recherche par modèle. Les modèles dont les caractéristiques partagent certaines propriétés seront groupés soit par le critère de similarité, soit par leur score de vraisemblance sur une même données d'observation (de test).

Une première expérience réalisée sur 20 modèles MMG de locuteurs regroupés d'une manière aléatoire a permis d'avoir une sous classification (voir le tableau 4.1). L'exploration de l'arbre est effectuée à l'aide de la mesure

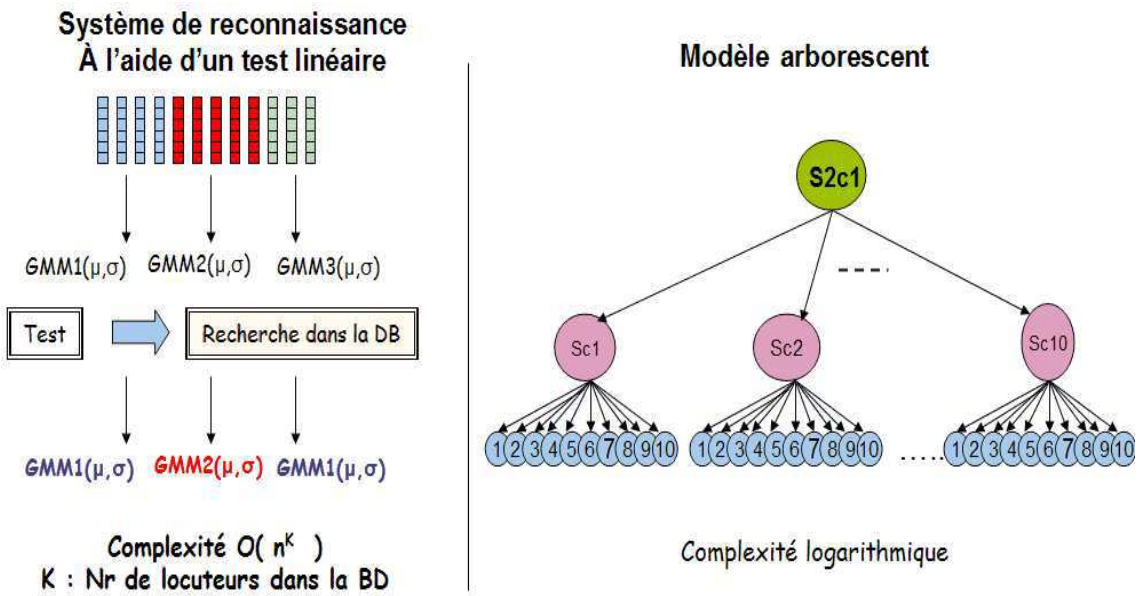


Figure 4.2 – Première expérience de sous classification des MMG locuteurs.

de score donné par la log-vraisemblance pour chaque test de noeud. Deux techniques sont alors utilisées pour l'exploration. La première est une exploration classique qui consiste à calculer le score de vraisemblance du modèle requête de la racine aux feuilles. La deuxième est une recherche itérative qui consiste à parcourir l'arbre de la racine au feuilles au premier temps, puis recommencer à partir d'un niveau supérieur "classe/groupe" dans le cas d'une erreur d'identification donné par un score supérieur au seuil maximum d'identification au premier test des modèles feuilles.

Dans les prochaines sections, nous allons présenter deux méthodes de classification des modèles MMG utilisant deux critères de similarité.

4.3.2 Méthode 1 : Calcul de la matrice de distance à l'aide du critère de similarité par calcul des scores de vraisemblance

La densité de probabilité d'un vecteur x de dimension d suivant une loi de mélange de K gaussiennes est calculée comme suit :

$$g(x) = \sum_{k=1}^{K} p_k g_k(x) \tag{4.8}$$

Recherche itérative			
Ordre de MMG	taille de données	Performance	Id-locuteur non reconnu
16	5 sec	100%	Néant
16	15 sec	100%	Néant
5	5 sec	90%	{1,15}
Recherche des modèles MMG classique			
16	15 sec	85%	{1,6,18}
20	30 sec	90%	{1,18}
60	15 sec	95%	{1}
60	30 sec	100%	Néant

Table 4.1 – Performance de l'exportation dans un arbre de recherche crée par regroupement de 20 MMG locuteurs.

tel que

$$g_k(x) = \frac{1}{2\pi^{d/2}|\sigma_X|^d} exp(-\frac{1}{2}(x-m_k)^T \sigma_X^{-1}(x-m_k)) \tag{4.9}$$

$$\text{Avec } \sum_{k=1}^{K} p_k = 1.$$

Les paramètres du modèle de mélange de gaussiennes sont $\theta = \{(p_k, m_k, \sigma_k)\}_{1 \leq k \leq K}$ où p_k, m_k et σ_k respectivement le poids, la moyenne et la matrice de covariance de la gaussienne d'indice k.

La probabilité de X sachant le modèle θ s'écrit :

$$p(X|\theta) = \prod_{t=1}^{T} p(x_t|\theta) \tag{4.10}$$

Un groupe de locuteurs $S = \{S_1, S_2, S_3, ..., S_N\}$ est représenté par les paramètres statistiques $\theta_1, \theta_2, \theta_3, ..., \theta_S$.

L'identification de locuteur requête est réalisée à l'aide du calcul du maximum de vraisemblance *Maximum-Likelihood* (ML) entre les segmenys de descripteurs associés aux locuteurs déjà inscrits dans la bases de données des locuteurs :

$$S_{loc} = argmax_k P(X|\theta_k) \text{ d'où}$$

$$S_{loc} = argmax_k (\prod_{t}^{T} \sum_{k=1}^{K} p_k . g_k(x_t)) \tag{4.11}$$

La formule de la matrice de distance D_{ML} des scores ML ainsi obtenue comme suit :

$$D_{ML}(i,j) = -2log(p(X_{loci}|\theta_j)) = -2\sum_{t=1}^{T} log(p(x_{ti}|\theta_j)) \tag{4.12}$$

Pour des applications liées à la reconnaissance automatique de locuteur nous utilisons 5 sec des données avec des modèles MMG de 16 composantes gaussiennes de dimension $d = 26$ (vecteur MFCC de 13 arguments plus 13 arguments issues d'une dérivation temporelle).

4.3.3 Construction de l'arbre à l'aide du critère de maximum de vraisemblance

La recherche et la nivigation par contenu lié à l'identité de locuteur dans le cas d'un processus de traitement incrémental consiste à effectuer des calculs d'estimation et de reconnaissance à l'aide du maximum de vraisemblance à l'issue de chaque changement détecté produisant un segment de locuteur inconnu (requête). Les modèles bien obtenus permettent de créer une matrice de distance à l'aide des scores de vraisemblance entre les modèles de locuteurs et leurs segments deux par deux. Le calcul de la matrice de distance est effectué à partir de locuteurs déjà inscris dans la base de modèles. En effet, La matrice obtenue permet alors de répartir les locuteurs dans un espace de distance afine de dimension $d = 2$, ce qui permet de représenter les MMG par un dendrogramme (voir la figure 4.3).

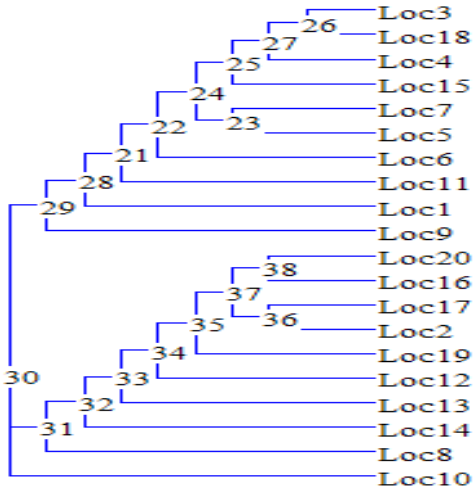


Figure 4.3 – Dendrogramme généré à partir de la matrice de distance des score de vraisemblance de 20 modèles MMG de locuteurs.

4.3.3.1 Méthode 2 : utilisation de la divergence de Kullback-Leibler pour la génération de l'arbre de décision

La divergence de Kullback-Leibler est une approximation d'une distribution P_1 par P_2. Une première évaluation entre deux distributions gaussiennes peut servir de distance entre deux distributions gaussiennes (voir l'equation 4.1). En outre, l'expression donnée par l'expression (4.2) est employée pour évaluer la distance entre les ensembles d'apprentissage et de validation croisée.

La section (4.2) décris d'une manière détaillée la technique de transformer la divergence KL simple à une divergence $KL_{\text{modifiée}}$ pour deux MMG. La version symétrique employée comme une distance est données par (l'équation 4.7). Le dendrogramme est généré à partir de la matrice de distance des modèles déjà inscrits, notons ici que le coût de création de la matrice de distance est largement moins coûteux par rapport à la matrice de distance générée à partir du score de vraisemblance. Le dendrogramme 1 (voir figure 4.3) présente une répartition en deux

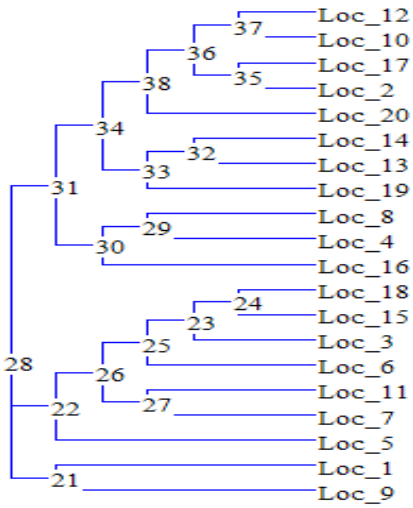


Figure 4.4 – Dendrogramme construit à l'aide de la matrice de distance à l'aide de LM_m.

grandes classes de locuteurs. L'arbre est de type binaire représente une forme non équilibrée. Par conséquence, le coût d'exploration est similaire à celui d'un traitement linéaire réparti en deux classes. Dans cette exemple voir figure. fig :dendrogramme2

4.3.4 Regroupement par critère de similarité

Cette méthode à pour but de découper l'arbre en choisissant un niveau permettant de segmenter l'ensemble des locuteurs en différentes classes jugées similaires. Cette méthode permet d'améliorer le regroupement des MMG de locuteurs basé sur les dendrogrammes de similarité, Les matrices de distance utilisent soit :

- le score de vraisemblance entre les modèles deux par deux ;
- la divergence de Kullback-Leibler.

Un découpage du dendrogramme de distance permet alors de classifier les locuteurs, selon le critère de similarité. La recherche et l'identification se résume alors à un test de N parmi les M locuteurs présents. Cette technique nous a permis de valider l'approche d'une sous classification appliquée sur des MMG, qui demeure une des différentes techniques de base présentées dans l'état de l'art [51].

L'utilisation du dendrogramme est présentée par cet exemple (voir la figure 4.3) nous permettra d'organiser les modèles de locuteurs en trois sous classes (voir tableau 4.2) puis en deux sous groupes bien équilibrés, comme la répartition donnée par (tableau 4.3) :

Dans cet exemple deux sous-classes "noeud du dendrogramme" sont créées grâce à l'ensemble des données de

Classe	Sous classe 1	Sous classe 2	Sous classe 3
Locuteurs label	{3, 18, 4, 15, 7, 5, 6, 11, 1, 9}	{20, 16, 17, 2, 19, 12, 13, 14, 8 }	10

Table 4.2 – Regroupement des modèles de locuteurs à l'aide de critère de maximum de vraisemblance comme mesure de similarité.

Classe	Sous classe 1	Sous classe 2
Locuteurs label	{3, 18, 4, 15, 7, 5, 6, 11, 1, 9}	{20, 16, 17, 2, 19, 12, 13, 14, 8 } + 10

Table 4.3 – Regroupement des MMG à l'aide d'un découpage de la structure arborescente binaire basée sur KL comme mesure de similarité.

modèles concernés. Le coût d'identification et de recherche est réduit en $card(Sc) + card(Sc_i)$ tests inférieur à N tel que Sc est l'ensemble. Par conséquent, la méthode d'organisation des modèles en une structure arborescente équilibrée dépend directement du dendrogramme construit à partir de la matrice de distance. De plusCependant, la

mesure de distance basée sur le critère de similarité entre modèles statistiques permet d'obtenir une structure arborescente légèrement équilibrée. Cette difficulté est assez commune dans le domaine de la recherche par similarité dans les grands volumes de données.

Ordre de MMG sous-classe(5 MMG)	Taille de données d'entraînement	Taux de reconnaissance
16	15 sec	85%
20	30 sec	90%
60	15 sec	95%
60	30 sec	100%

Table 4.4 – Performance de reconnaissance en fonction de taille de données d'apprentissage des modèles regroupés par une sous-classe

Cette première expérience d'organisation de locuteurs a permis d'abord de valider l'approche hiérarchique d'organisation des mélanges de gaussiennes. Cette technique est simplement basée sur la représentation d'un groupe de modèle par un seul MMG partageant les mêmes propriétés. Le modèle ainsi créer est un modèle de mélange de gaussiennes nommé "sous-classe" par rapport à son niveau dans la structure. Pour une meilleure performance nous constatons Ce dernier est composé d'un nombre de composantes gaussiennes égale à la somme des composantes gaussiennes de tout les modèles regroupés (voir tableau 4.4). Cependant, l'organisation de modèle de locuteur sous forme de structure arborescente est une technique très intéressante pour la phase d'exploration, mais elle demande un coût considérable lors de sa création. D'où la nécessité d'utiliser des techniques de réduction de MMG [33].

4.4 Conclusion

Ce chapitre a permis de dresser les outils de base permettant de construire une structure arborescente afin de valider l'approche de classification hiérarchique. Cette approche consiste à effectuer une sous classification à travers un groupement des données de chaque modèle de mélanges et estimer un modèle représentant l'union de ces données. Les critères de similarité utilisés ont pour but d'améliorer le regroupement d'une manière supervisée. Les classes ou les modèles de locuteurs similaires sont regroupés en découpant un dendrogramme à un niveau donné. Ce dendrogramme est crée à l'aide de deux critères : le score de vraisemblance de données et la mesure de

KL_m. La matrice de distance est alors générée, offrant plusieurs possibilité de générer un arbre binaire suite à un découpage linéaire ou adapté.

Nous avons présenté une expression de la divergence de KL adaptée pour le calcul rapide et moins coûteux de la distance entre modèles de mélanges de gaussiennes de locuteurs. Le dendrogramme résultant par les deux mesures est légèrement différent. Cepedant le résultat de regroupement supervisé par un groupement par similarité permet d'améliorer la représentation des MMG par desclasses, augmentant ainsi sa performance d'exploration.

Dans la suite, nous tirons profit de cette étude afin de réduire le nombre de composantes de MMG noeud qui représentent un ensemble de MMG par un critère de similarité. Dans ce but, une technique de fusion de MMG sera développée dans la capitre suivant. Le MMG généré par cette fusion permet de représenter deux MMG jugés similaire en ayant le même nombre de composantes de modèle.

CHAPITRE 5

Regroupement ascendant des MMG de locuteurs

L'objectif principal de ce chapitre est de proposer une nouvelle approche d'organisation arborescente ascendante des MMG. La création de la structure arborescente est effectuée sur un lot de MMG déjà inscrit dans la base de données. Cette organisation ne peut être réalisée au fur et à mesure que des nouveaux modèles de locuteurs sont détectés. La structure générée est binaire tel que les feuilles sont représentées par les MMG des locuteurs, les noeuds sont des MMG générés par un algorithme de fusion de GMM détaillé dans la suite de ce chapitre .

Grâce à l'efficacité de la représentation d'un mélange de gaussiennes au locuteur via ses descripteurs MFCC, l'approche proposée permet d'organiser les modèles de locuteurs par lot sous une structure arborescente. Après avoir effectué un prétraitement de tri par similarité, le regroupement par deux MMG donne forme à une structure arborescente crée d'une manière ascendante.

5.1 Introduction des techniques de regroupement des modèles des locuteurs

Le contexte de ce travail est défini de tel manière que, le processus d'indexation des documents audio ne contenant que de la parole, dont lequel un flux est introduit dans un système d'indexation. Des documents comme : des journaux d'informations, des débats télévisés et un flux radio sont au coeur du processus d'indexation au sens de locuteur et de recherche par contenu. Le but de notre contribution est de produire une technologie de recherche par locuteur des documents audio ne contenant que de la parole où plusieurs locuteurs échangent des tours de parole. L'application envisagée permet aux utilisateurs de chercher par identité de locuteur (i.e l'intervenant(s) ou tous les interventions) par durée de segment ou contexte temporel. Souvent dans le cadre d'applications les courtes périodes de silence et de publicité peuvent être négligées, ceci grâce aux propriétés des descripteurs acoustiques MFCC qui

ne garde que les informations pertinentes vis-à-vis du locuteur est calculé sur des fenêtres d'obseravation assez large de l'ordre de (4 sec).

Les techniques de modélisation statistique ainsi que les critères d'authentification sont en générale des techniques communes à un système d'identification de locuteur. Une solution classique à la tâche mentionnée ci-dessus consisterait à segmenter le flux audio en segments de locuteur homogènes. Dans l'implémentation, le coût de la tâche qui consiste à comparer la représentation d'un segment de données acoustiques appartenant à un locuteur (récemment détecté et inconnu) avec l'ensemble des modèles des locuteurs déjà enregistrés augmente linéairement en fonction du nombre de locuteurs testés. Le but principal de ce chapitre est de proposer une nouvelle technique d'organisation de l'ensemble des représentations des locuteurs afin de réduire le temps de recherche dans la cas d'une bases de données volumineuse. Le processus ainsi conçu permet de gagner un temps d'exécution considérable en faveur d'une légère perte de performances d'identification.

La tâche d'indexation au sens de locuteur est liée étroitement à des questions classiques de structure de données et d'indexation des données complexe. La communauté de bases de données propose plusieurs contributions basées sur une variété de structures arborescentes. La particularité du problème courant résulte de la nature des entités à classer, à savoir les modèles statistique, pour lesquels les structures classiques d'indexation sont encore inadéquates.

La modélisation des descripteurs acoustiques est garantie par l'utilisation des MMG, une technique qui reste toujours la plus dominante dans le domaine de RAL grâce à ses performances de représentation et de reconnaissance. À l'aide d'un segment de données acoustiques correspond au $k^{\text{ème}}$ locuteur, le modèle de mélange de gaussiennes est donné par la formule (voir eq. 5.1) ci-dessous, tel que $N_k^i(x)$ est la composante gaussienne, de moyenne μ_k^i, de covariance Σ_k^i et de w_k^i les poids tel que $\sum_{i=1} w_k^i = 1$.

$$M_k(x) = \sum_i w_k^i N_k^i(x) \tag{5.1}$$

Voici les raisons pour lesquelles le choix de MMG est justifié pour la représentation des données acoustiques de locuteur :

1. Nous optons pour des modèles génératifs plutôt qu'une approche discriminante (comme par exemple des machines de vecteurs), sachant que le nombre de classes croît en continu au fur et à mesure que des nouveaux locuteurs apparaissent dans le système.
2. Le MMG permet d'offrir une bonne performance de représenter l'information issue du timbre vocal pour un

grand nombre de locuteurs en mode dépendant de texte dans un espace multidimensionnel. En particulier, les mélanges de gaussiennes sont commodément favorables à l'identification à partir de la manipulation des paramètres sans avoir recours directement aux données.

3. La mise à jour des modèles MMG de locuteurs est assurée à l'aide d'un algorithme qui permet la fusion des paramètres des modèles de mélange de gaussiennes.

5.2 Classification hiérarchique des MMG de locuteur

5.2.1 Principe et état de l'art

Rappelons dans le cas d'un schéma incrémental d'acquisition le processus d'identification d'un modèle de locuteur récemment détecté par rapport aux locuteurs candidats déjà enregistrés dans la base de données des MMG est très coûteux. L'organisation des modèles de locuteur en structure qui permet de réduire la complexité de la recherche s'avère idispensable.

Les travaux proposés actuellement dans l'état de l'art se basent sur des techniques d'ancrage de modèles de locuteur [81, 49].

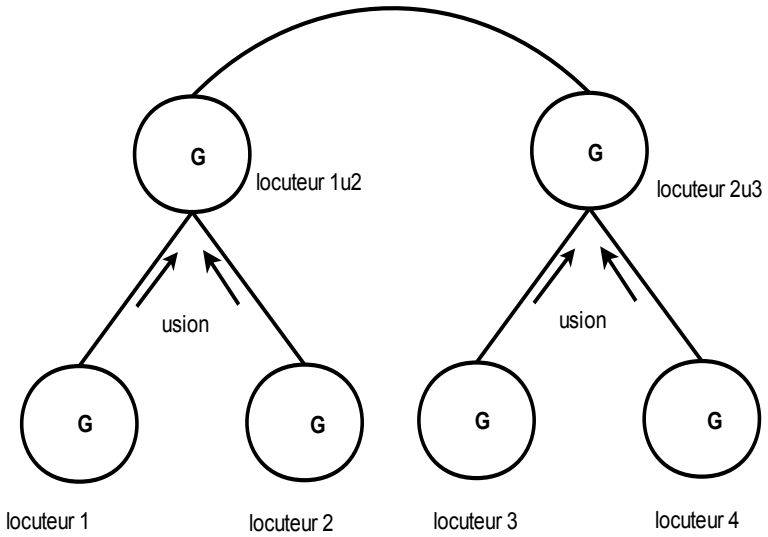


Figure 5.1 – Principe de regroupement des MMG sous forme d'une structure hiérarchique

5.2.2 Fusion des MMG

Dans cette section, des travaux récents portant sur la classification hiérarchique des modèles de mélanges basés sur la technique de réduction de MMG par le critère de similarité exprimé à l'aide d'une approximation de la divergence de KL [33]. En outre, dans ce travail nous proposons une extension de cet technique dans le but fusionner deux MMG en utilisant que leurs paramètres (μ_k^i, Σ_k^i et w_k^i). Le but de fusionner deux MMG est de créer une structure arborescente binaire de recherche à moindre coût. Le choix de structurer des modèles de locuteur d'une manière arborescente a été traité par [22, 65, 33] afin de proposer une approche hiérarchique de classification de locuteurs de large volume de données.

5.2.2.1 Arbre binaire de recherche des modèles statistiques de locuteur

Nous avons choisi de structurer les modèles de locuteurs sous forme d'un arbre binaire de recherche selon les arguments suivants :

1. La structure binaire est crée à moindre coût ce qui permet sa mise à jour pour chaque lot de MMG détecté ;
2. L'algorithme d'exploration utilise les même outils de recherche linéaire qui sont plus familiés au utilisateurs (ML, KL_m).

Le groupement de deux modèles de mélange de gaussiennes est basé sur un calcul de similarité à l'aide de la mesure de KL_m. A partir de la matrice de similarité nous choisissons d'abord de fusionner les deux modèles qui représentent une forte similarité à l'aide des deux critères *ML* ou KL_m

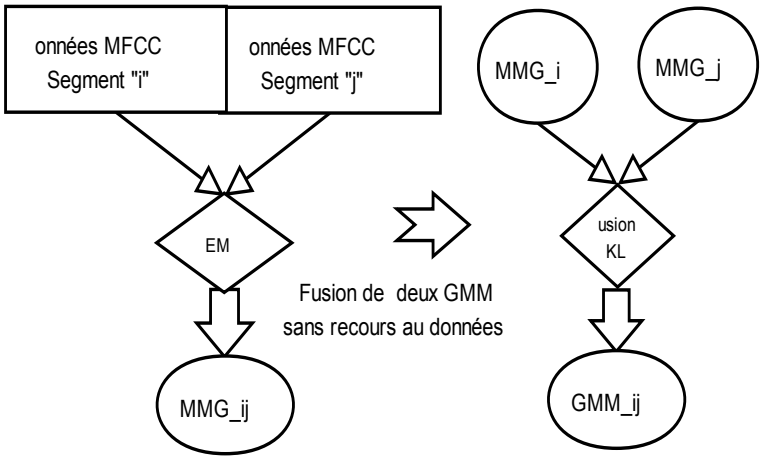


Figure 5.2 – Création d'un MMG de fusion de deux vrais locuteurs à moindre coût.

5.2.3 Algorithme de fusion de MMG

A l'aide des paramètres des deux mélanges de gaussiennes de locuteurs à fusionner, l'algorithme de fusion génère un modèle de mélange de gaussiennes "noeud" des travaux similaires sont décrits dans [32]. L'algorithme de fusion permet alors de générer un modèle composé du même nombre de composantes gaussiennes. Ainsi que, les modèles de fusion représentent les noeuds dans la structure à construire. L'algorithme de fusion est appliqué après une sélection de deux par deux des modèles mélanges de gaussiennes de locuteurs selon un critère de similarité (score *ML* ou KL_m) (voir figure 5.2). En outre, l'avantage principal de la fusion des mélanges de gaussiennes est de préserver à l'issue de cette fusion de MMG le même nombre de composantes gaussiennes et représente les deux modèles sources. Enfin, l'algorithme de fusion doit être moins coûteux et fournir de bonne performance d'identification lors du passage à l'échelle incrémentale.

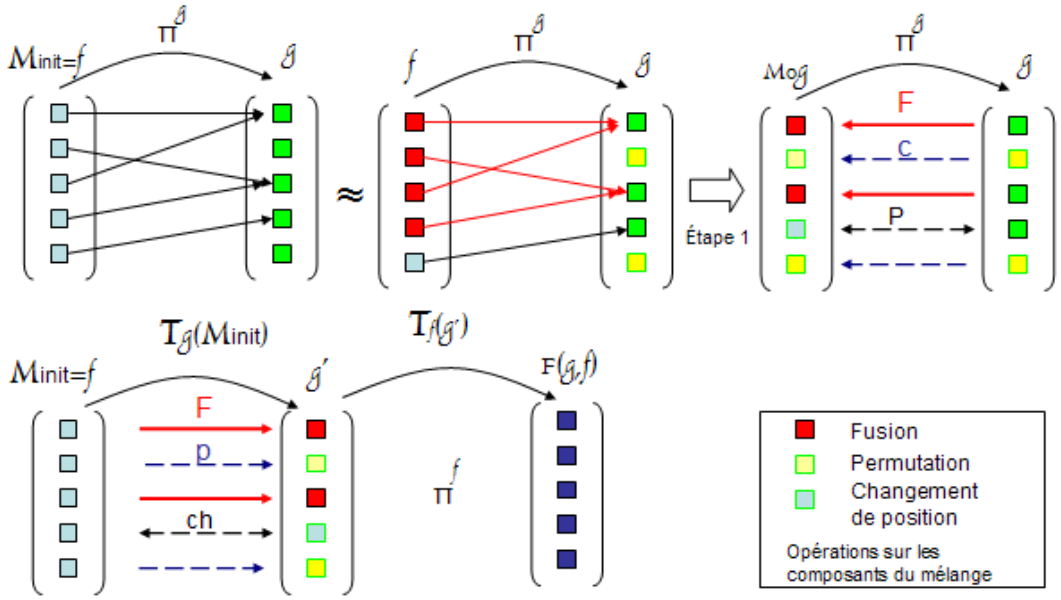


Figure 5.3 – Processus de fusion des deux mélanges de gaussiennes

L'algorithme de fusion de MMG est composé de trois opérations de base sont définies (*RÉDUIRE, AJOUTER et PERMUTER*), le but de ces opérations et de minimiser la distance KL_m entre les deux mélanges de gaussiennes source et le modèle résultant par fusion. L'évaluation de l'algorithme de fusion du modèle est effectué à l'aide de deux critère : BIC et la divergence de Kullback-Leibler. Ces deux critère permet de valider l'efficacité de cette technique à produire des modèle qui permet de représentés à la fois les deux modèles source ainsi les données pour les quelles ont était crées. Dans la pratique deux itérations de base sont nécessaire e.i. $g^{'} = T_g(f)$ puis

$M_{g \cup f} = T_f(g^{'})$.

Une illustration simplifiée du principe de fusion est basé sur une somme euclidienne des paramètres des gaussiennes de dimension 2 (voir la figure 5.3). Deux composantes gaussiennes seront fusionnées s'ils présentent une distance de similarité très proche.

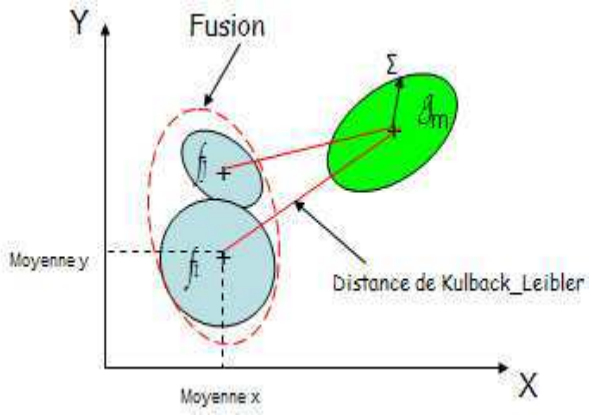


Figure 5.4 – Représentation 2D de fusion de deux composantes Gaussiennes

L'algorithme de fusion prend en entrée deux modèles mélanges de gaussiennes de même nombres de composantes. Trois opérations sont alors effectuées dans l'ordre suivant :

1. **"RÉDUIRE" : Réduire par fusion des composantes gaussiennes** :
   L'algorithme de fusion entre deux modèles de mélanges ($f = \sum_{i=1}^{m} \alpha_i N(\mu_i, \sigma i)$ et $g = \sum_{j=1}^{m} \alpha_j N(\mu_j, \sigma j)$) s'applique d'abord sur un des modèles à fusionner (par exemple g). La première étape consiste à calculer une fonction de transfert π en utilisant KL_m. Ensuite, une phase d'initialisation avec un MMG choisi aléatoirement (en pratique $M_{init} = f$ ou à l'aide d'une initialisation aléatoire). Notons par π^g la fonction de transfert entre g et M_{init}. La deuxième étape consiste à simplifier le modèle (exp. $M_{init} = f$) en réduisant le nombre de composantes à l'aide d'algorithmes donnés par *J. Goldberger et S. Roweis* dans [33] en utilisant la fonction de transfert π^g. La technique de réduction de modèles consiste à regrouper toutes les composantes du modèle M_{init} qui possèdent une image unique par rapport à la fonction de transfert π^g. soit *GoM(m)* l'ensemble de mélanges de gaussiennes avec *m* composantes gaussiennes et Ψ l'ensemble de fonctions défini de $\{1, ..., k\}$ à $\{1, ..., m\}$. Pour chaque $\pi \in \Psi$ et $M_{init} \in$ *GoM(m)* la réduction de MMG est définie :

$$d(f, g, \pi) = \sum_{i}^{k} \alpha_i KL(g_{\pi(i)} || M_{init_i}). \tag{5.2}$$

Pour un modèle donné $M_{init} \in GoM(m)$ la fonction de transfert permettant de donner une distance minimale de KL est définie comme suit :

$$\pi^g(i) = \arg\min_{j=1}^{m} KL(g_i||M_{init_j}) \text{tel que} i = 1, ..., k \tag{5.3}$$

Nous pourrons montrer facilement que :

$$d(M_{init}, g) = d(M_{init}, g, \pi^g) = \min_{\pi \in \Psi} d(M_{init}, g, \pi) \tag{5.4}$$

D'où, la fonction de transfert (5.4) optimale π^g est obtenue entre les composantes de M_{init} et g, en utilisant l'équation (5.3) le modèle réduit est alors considéré comme une solution d'une double minimisation :

$$M_c^{'} = \arg\min_{M} \min_{\pi \in \Psi} d(M_{init}, g, \pi) \tag{5.5}$$

Pour m$>$ 1, la double minimisation de l'équation (5.5) ne peut pas être résolue avec une méthode analytique. Nous pouvons employer la minimisation alternative pour obtenir un minimum local. En effet, pour une fonction de transfert donnée et pour $\pi^g \in \Psi$, on peut définir $g^{'}$ tel que :

$$M_{c_j}^{'\pi} = \frac{\sum_{i \in \pi^{-1}(j)} \alpha_{M_{init_i}} M_{init_j}}{\sum_{i \in \pi^{-1}(j)} \alpha_{M_{init_i}}} \tag{5.6}$$

L'expression de la moyenne et la covariance de M_{init_j} ($\mu_j^{'}$ et $\Sigma_j^{'}$) sont alors écrites de la mainère suivantes :

$$\mu_j^{'\pi} = \frac{1}{\beta_j} \sum_{i \in \pi^{-1}(j)} \mu_i \tag{5.7}$$

et

$$\sigma_j^{'} = \frac{1}{\beta_j} \sum_{i \in \pi(j)^{-1}} \alpha_i (\sigma_i + (\mu_i - \mu_i^{'})(\mu_i - \mu_i)^{'T}) \tag{5.8}$$

avec

$$\beta_i = \sum_{i \in \pi^{-1}(j)} \alpha_{M_{init_i}} \tag{5.9}$$

D'où $M_{ci}^{'\pi} = N(\mu_i^{'}, \Sigma_i^{'})$ est une composante gaussienne obtenue à l'aide de la fusion des composantes gaussiennes de $M_{init_j}^{'\pi}$ partageant la même image par la fonction de transfert π^g, en une simple gaussienne :

$$M_{c_i}^{'\pi} = N(\mu_i^{'}, \Sigma_i^{'}) \tag{5.10}$$

$$= \arg\min_{M} KL(M_c^{'\pi}||g) = \arg\min_{M} d(M_c^{'\pi}, g) \tag{5.11}$$

2. **"AJOUTER"**

   Pour tout *k* tel que $\pi^{g^{-1}}(k) = \phi$, l'opération "AJOUT :ADD" consiste à ajouter les composantes de g à $g^{'}$. Les composantes ajoutées sont alors celles qui n'ont pas d'image par rapport à la fonction de transfert. Cette opération permet de garder les composantes de g et celles ajoutées au modèle $g^{'}$ pré-estimé avec le même nombre de composantes en sortie.

3. **"PERMUTER"** :

   Pour tout *s* tel que $Card(\pi(s)) = 1$, Nous procédons par un changement de position de la composante gaussienne comme suite (cette opération n'affecte pas la représentation de modèle, mais elle consiste seulement à obtenir une bijection entre les modèles de fusion et le modèle estimé) : $M_p^{'}(\pi(s)) = M_{init}(s)$, afin de garder les composantes significatives, qui représentent un meilleur mélange de gaussiennes original.

Le MMG obtenu $g^{'}$, présente le résultat de la première étape de fusion $M_{init} = f$ avec *g* :

$$g^{'} = \sum_{(i,\pi^{g-1}(i)\neq\phi)} (\frac{\beta_{M_{c_i}^{'}}}{\Omega}) M_{c_i}^{'} + \sum_{(j,\pi^{g}(j)=\phi)} (\frac{\alpha_{Add_j}}{\Omega}) M_{Add_j}^{'}$$

$$+ \sum_{(k,Card(\pi(k))=1)} (\frac{\alpha_{p_k^{'}}}{\Omega}) M_{p_k}^{'} \tag{5.12}$$

avec :

$$\Omega = \textstyle\sum_n \beta_{M_{c_n}} + \sum_m \alpha_{Add_m} + \sum_l \alpha_{p_l}^{'}$$

Enfin le modèle de fusion $M_{\hat{Merge}}$ est obtenu en faisant appel aux étapes précédentes en remplaçant M_{init} par $g^{'}$ et g par f.

5.2.3.1 Résumé de l'algorithme de fusion des mélanges de gaussiennes de locuteurs

1. Trouver la fonction de transfert π^g entre M_{init_i} et g ;
2. Opération **"Réduire :Collapse"** l'ensemble des composantes Gaussiennes ayant la même image dans g par π^g, ainsi notée $\Rightarrow M_{c_i}^{'}$;

3. Opération **"Ajout :Add"** Ajouter les composantes Gaussiennes de g n'ayant pas d'image avec l'inverse de la fonction de transfert : $M_{Add'_k} = g_L$ tel que $\pi_k^g = L$;
4. Opération **"Permuter :Permute"** M_{init_i} les composantes Gaussiennes ayant une seule et unique image, i.e $M_{p'_k} = M_{init_n}$, tel que $Card(\pi(k)) = 1$ et $\pi(k) = n$;
5. Faire 1,2,3,4 pour f avec $g^{'}$ pour obtenir : $\hat{M_{Merge}}$;
6. Répéter jusqu'à ce que $d_{KL}(\hat{M_{Merge}}, f) \prec \epsilon$ et $d_{KL}(\hat{M_{Merge}}, g) \prec \epsilon$, tel que ($\epsilon$ =seuil), d'où $M_{Merge} = \hat{M_{Merge}}$.

Par la suite, nous discuterons de la performance de l'algorithme en utilisant une initialisation aléatoire de modèle M_{init}. Le critère BIC est employé pour la sélection du meilleur modèle qui répond au critère de similarité par rapport à la mesure de KL_m. Le but est d'avoir une convergence par rapport à la mesure de KL_m et une approximation du maximum de vraisemblance pénalisée par le critère BIC.

5.3 Étude de performance d'algorithme de fusion des mélanges de gaussiennes

L'étude de performance de la technique de fusion est basée sur le critère BIC de sélection du modèle mélanges de gaussiennes par rapport aux données $D_{BIC} = D_f \cup D_g$ des deux mélanges de gaussiennes fusionnés. Nous nous plaçons dans le contexte bayésien de sélection de modèle dans le but de vérifier la performance de l'algorithme par rapport au critère de maximum de vraisemblance. La mesure de KL_m entre les deux modèles d'entrée et le modèle crée est utilisée comme critère d'arrêt dans le cas de validation du premier critère.

Soit $M_A, M_B \in MoG(m)$ et $M_{AUB} \in MoG(m)$ tel que $M_{AUB} = Fusion(M_A, M_B)$, l'estimation du modèle de fusion en utilisant le critère BIC défini tel que :

$$BIC_{M_{AUB_i}} = -2log(P(X_A \cup X_B | \theta'_{M_{AUB_i}})) + d.log(n_A + n_B) \tag{5.13}$$

Ainsi, la sélection du meilleur modèle est donnée par :

$$M_{AUB_{BIC}} = \arg\min_{Mi} BIC_{M_{AUB_i}} \tag{5.14}$$

Rappelons que l'algorithme de fusion des MMG est itératif, l'initialisation est aléatoire, mais dans la pratique l'initialisation est réalisée en choisissant un des modèles d'entrée (f ou g). Le critère d'arrêt utilise la mesure

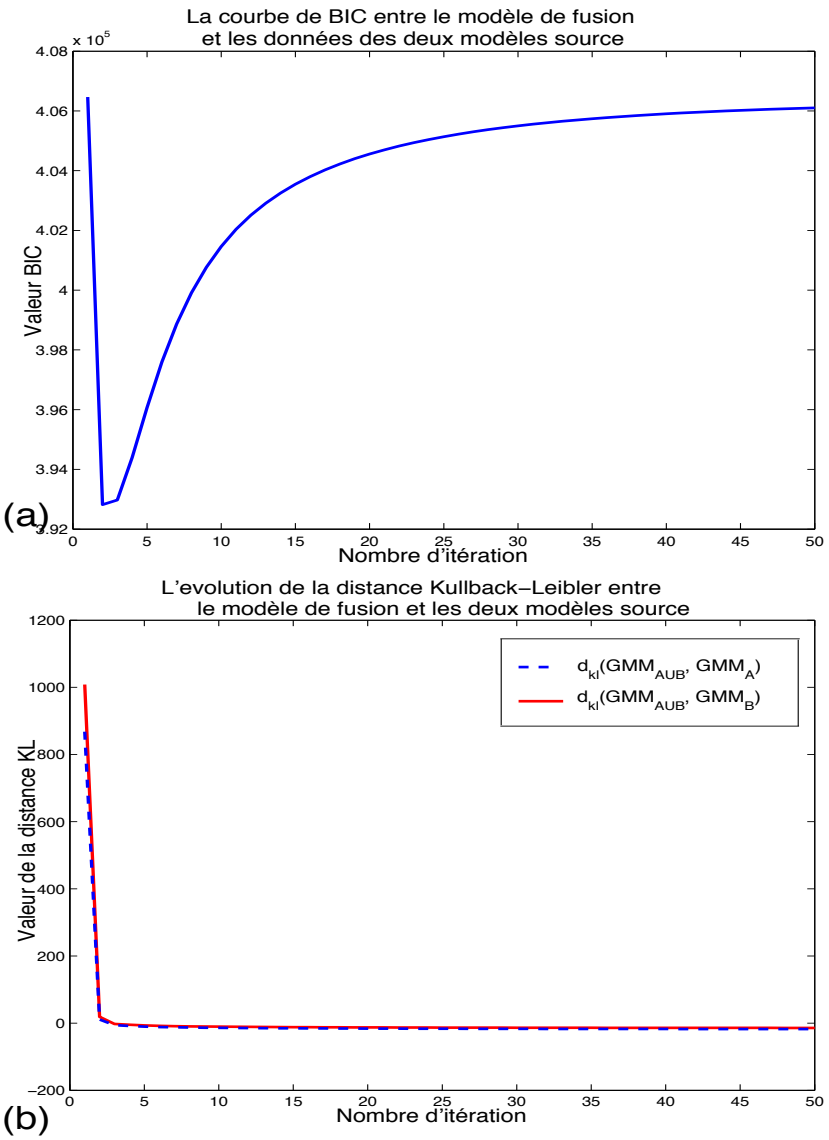


Figure 5.5 – (a) Distance de KL_m entre le modèle fusionné et les deux modèles à fusionner. (b) La valeur de BIC pour un modèle fusionné comparé à l'ensemble des segments de descripteurs des deux modèles

de KL_m entre les deux modèles d'entrée et le modèle de sortie. Cette mesure converge vers un minimum après quelques itérations (voir figure 5.5). La figure de gauche (5.5(a)) présente l'évolution de l'estimation du modèle BIC par rapport à l'ensemble des données des deux segments modèles d'entrées M_A et M_B. La courbe représente un minimum local après un nombre très faible d'itérations égal à trois, ce qui signifie que le modèle résultant représente bien les deux modèles d'entrées de la même manière que s'il a été généré directement en utilisant l'en-

semble des deux segments $D_A \cup D_B$. La deuxième partie de la courbe diverge rapidement après les trois première itérations (voir figure 5.5(a)), le modèle généré présente toujours une distance très faible par rapport à KL_m, en revanche, il ne représente pas l'union des deux segments des modèles d'entrées en terme de ML (voir figure 5.5(b)).

La courbe de droite (figure 5.5(b)) traduit le même résultat que celui de la courbe gauche sur la première partie, en revanche, la deuxième partie de la courbe de la distance KL_m entre le modèle de fusion et les deux modèles d'entrée continue de converger et reste pratiquement stable durant la deuxième partie supérieur à la $3^{\text{ème}}$ itérations de l'algorithme. Cependant, les modèles fusionnés sont indépendant aux données des modèles, de plus, le modèle résultant doit vérifier les critères pour lesquels il était généré c'est à dire avoir une représentation en un MMG des deux modèles jugés similaires sous la forme d'un modèle de dimension identique. En résumé, le critère de BIC peut être utiliser comme critère d'arrêt ce qui permis de valider l'efficacité de la fusion en seulement quelques itérations même avec une initialisation aléatoire. Rappelons qu'il suffit d'utiliser une initialisation avec un des deux modèles.

5.3.1 Arbre binaire de recherche basé sur l'algorithme de MMG-Fusion

Actuellement, la recherche et l'identification des locuteurs se basent sur des algorithmes de complexité linéaire non adaptée à l'indexation audio dans le cas de grands volumes de données. L'algorithme de fusion permet d'organiser les modèles des bases de données des modèles locuteurs sous forme d'une structure arborescente binaire. La structure hiérarchique des modèles de locuteurs réduit considérablement la complexité d'exploration ainsi l'identification. Pour cela, l'efficacité de la mesure de similarité joue un rôle important dans le processus de la construction de la structure et aussi l'exploration rapide. La structure hiérarchique doit être flexible, dynamique et moins coûteuse. En revanche, une telle structure pose un risque de perte d'information lors de sous classification qui peut engendrer des erreurs d'identification. Cette erreur n'est pas encore traitée dans cette partie de travail.

La structure hiérarchique est générée premièrement à l'aide de l'algorithme *'Neighbor Joining'* [72] (voir figure 5.8). Cet algorithme utilise la matrice de distance KL_m entre modèles mélanges de gaussiennes locuteur. L'arbre binaire de recherche est créee d'une manière ascendante par regroupement deux par deux des modèles par similarité à l'aide de KL_m. Le choix d'utiliser KL_m est largement moins coûteux car il utilise seulement les paramètres des MMG de locuteurs générés suite à une étape de détection de changement de locuteur qui génére les locuteurs requêtes.

Le premier exemple illustré dans la figure (5.6) permet de valider l'approche de regroupement afin de construire une structure arborescente. Soit deux modèles MMG_A et MMG_B dont la mesure de similarité est très importante. La distance KL_m calculée entre les deux modèles MMGs est très faible. En outre, soit $MMG_{F=A\cup B}$ le MMG résultant de la fusion des deux MMG_A et MMG_B, l'interprétation de ce graphe est résumé dans les points suivantes :

1. La distance du KL_m de MMG_A avec MMG_B décroît en fonction du nombre de composantes des deux modèles (exp. :$2, 4, 8, 16, 26...$),
2. La distance du KL_m entre MMG_A ou MMG_B avec MMG_F représentent un écart considérable par rapport à $KL(MMG_A, MMG_B)$, le but est d'avoir une meilleure discrimination entre MMG_A et MMG_B par rapport à MMG_F,
3. Pour un nombre de composantes égal à 16, le modèle de fusion MMG_F présente une distance minimale avec les deux modèles MMG_A et MMG_B par rapport à la distance de MMG_A avec MMG_B.

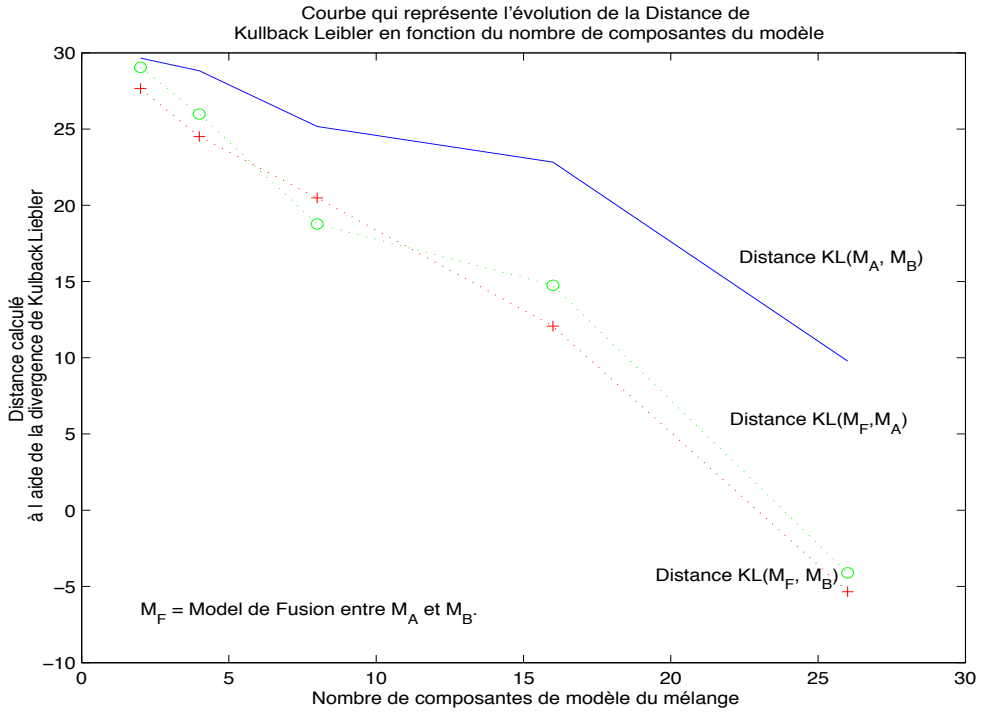


Figure 5.6 – L'évolution de la distance du modèle de fusion avec le modèle original en fonction du nombre de composantes du mélange de gaussiennes

Dans un deuxième exemple, l'évaluation de l'approche de construire une structure ascendante binaire utilise

trois mélanges de gaussiennes de locuteurs sont regroupés à l'aide d'algorithme de fusion. Le but est de tester la fiabilité de l'exploration à l'aide de la mesure de KL_m. Rappelons que l'algorithme de fusion permet la construction d'un modèle par couple de MMG. La première étape consiste à prendre les deux modèles de locuteurs qui représentent une forte similarité (exp. :A et B) noté $MMG_{A\cup B}$. Ce dernier est fusionné avec un troisième modèle nommé C noté $MMG_{A\cup B\cup C}$. La figure (5.7) illustre les courbes de distance de KL_m par rapport au nombre de composantes gaussiennes de chaque MMG locuteur. L'objectif d'une telle structure est de vérifier que la distance de KL_m respectivement des deux modèles premièrement fusionnés A et B avec le modèle noeud $A \cup B$ est toujours inférieure à la distance de A et B avec le modèle C. Remarquons que l'écart de distance entre MMG_C et $MMG_{A\cup B}$ augmente considérablement en fonction du nombre de composantes de MMG.

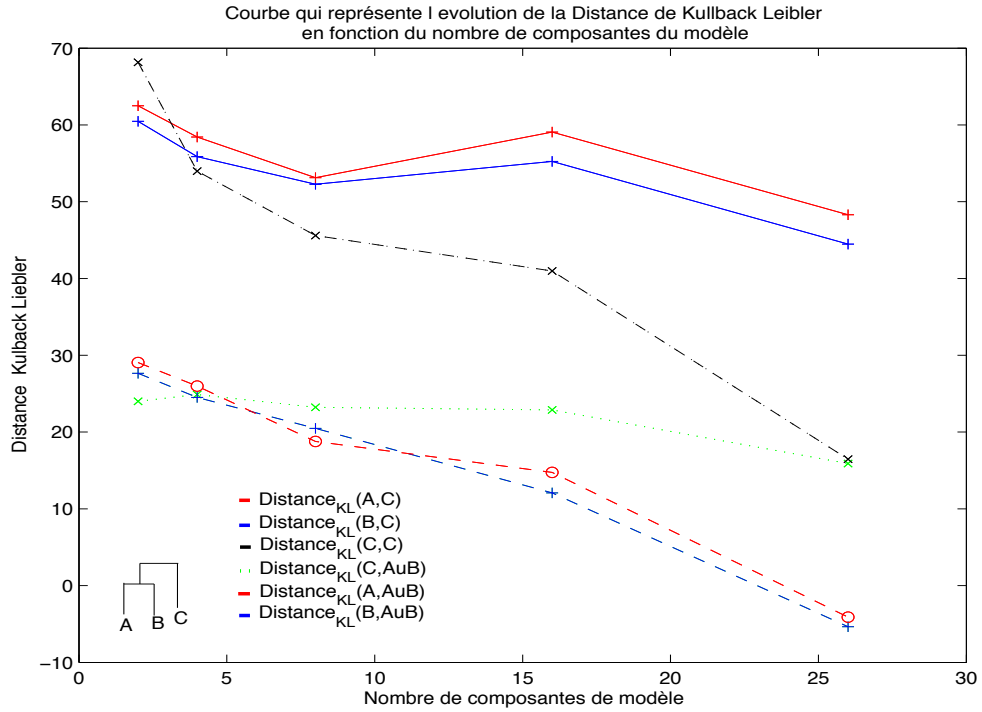


Figure 5.7 – L'évolution de la distance du modèle de fusion avec les modèles originaux en fonction du nombre de composantes du mélange

5.3.1.1 Algorithme (1) : Le plus proche voisin

Le premier algorithme proposé permet de créer la structure arborescente binaire directement à partir de l'ensemble des modèles MMG. Après avoir calculé la matrice triangulaire de distance des modèles existants en utilisant

la mesure de divergence de KL_m, la complexité de la construction de l'arbre binaire dépend du nombre N des MMG des locuteurs. L'algorithme de fusion crée de nouveaux MMG-noeuds pour chaque couple de MMG puis des MMG noeuds, la complexité de cette algorithme dépend du nombre des locuteurs :

$$Card(MMG - noeuds)_N = \sum_{k=1}^{N/2} 2 * k, \text{ si N est paire} \quad (5.15)$$

$$= (\sum_{k=1}^{\frac{N+1}{2}} 2 * k) - 1, \text{ si N est impaire} \quad (5.16)$$

Cet algorithme nous permet de grouper des modèles à l'aide du critère de similarité. Cependant, ce choix de regroupement produit. Souvent, un arbre binaire non équilibré liée au manque d'information sur la représentation des MMG dans un espace euclidien (voir le figure 5.8).

Algorithme (1) :Le plus proche voisin

1. Matrice$D = (D_{ij})$: matrice des similarités entre modèles (classes)
2. Trouver les classes r et s tels que $D_{rs} = min_{ij}(D_{ij})$
3. Recherche plus proche tel que $D(k, \{k, s\}) = min\{D(k, r), D(t, s)\}$.
4. Répéter le processus jusqu à trouver une seule classe

5.3.1.2 Algorithme (2) : "Neighbor Joining"

Le second algorithme utilisé nommé *"Neighbor Joining"* permet de mettre à jour la matrice de distance après chaque fusion de modèles en tenant compte les modèles résultants par la fusion. Cet algorithme permet de réorganiser les modèles de locuteurs et leur fusion à chaque niveau de l'arbre selon un critère le similarité. La représentation est améliorée grâce à l'information prise en compte des modèles de fusion après chaque affectation de noeud à l'arbre binaire. Le seul inconvénient est présenté dans le calcul de la matrice de distance après chaque fusion. Cette stratégie offre une fidèle représentation des modèles locuteurs (feuilles) et les modèles de fusion (noeuds) selon le critère de similarité. Cependant, la structure générée est bien équilibrée à l'aide d'utilisation de cet algorithme 2, en revanche le coût de la création de la structure est plus coûteuse que celle générée à l'aide du premier algorithme (voir figure 5.9).

Algorithme (2) : "Neighbor Joining"

1. Calculer la matrice de similarité à l'aide de KL_m entre MMG locuteurs $D = (D_{ij})$,

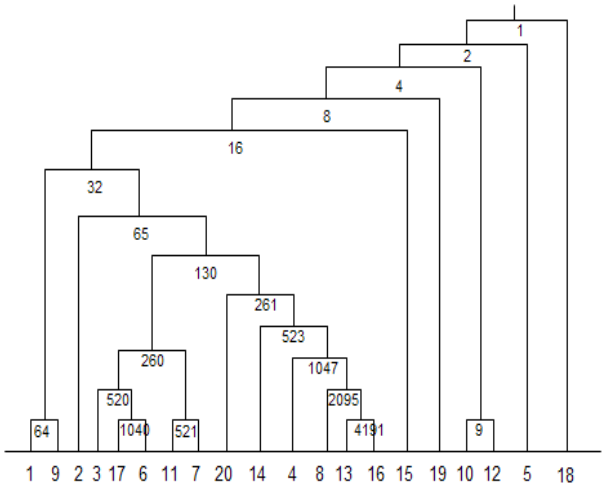


Figure 5.8 – L'algorithme (1) utilise la matrice de distance calculée à l'aide de KL_m sans tenir en compte les modèles fusionnés comme des nouveaux condidats. La création de l'arbre binaire est basée sur une vue globale produit souvent une structure non équilibrée.

2. Trouver les classes r et s tel que $D_{rs} = min_{ij}(D_{ij})$,
3. Fusionner les modèles MMG r et s en modèle MMG t,
4. Mettre à jour la matrice de similarité D, après avoir remplacé r et s par t tel que *t= MMG_Merge(r,s)*.
5. Répéter les étapes jusqu'à regroupement l'ensemble des modèle locuteurs et la création de modèle racine de l'arbre binaire.

5.4 Conclusion

Ce chapitre a cité d'abord les motivations et les objectives de la proposition d'une organisation hiérarchique des modèles de locuteurs. Ce travail est basé sur un concept de la modélisation statistique en MMG des segments de descripteurs acoustiques. Les justifications liés à l'intérêt de créer une structure arborescente ont été évoquées dans le but de réaliser un système l'indexation par locuteur dans le cas de passage à l'échelle incrémentale. Notre travail concerne plus précisément l'étude et l'organisation des modèles MMG de locuteurs basée sur une mesure de similarité. En effet, les techniques basées sur les paramètres propres au modèle MMG permet l'élaboration d'un nouvel algorithme de fusion qui cherche à mieux représenter les modèles MMG sous forme d'une nouvelle structure permettant une recherche non-linéaire.

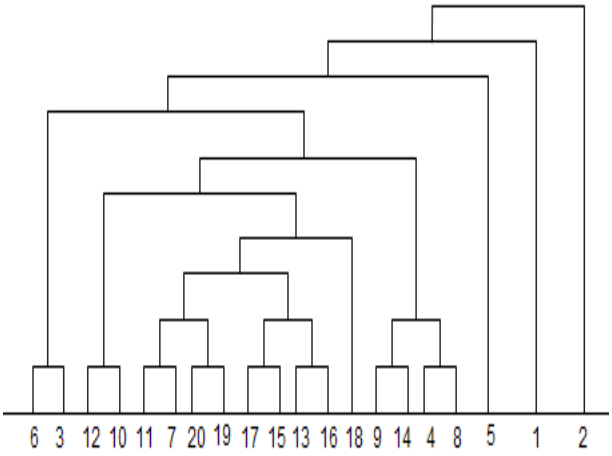


Figure 5.9 – L'algorithme (2) nommé 'Neighbor Joining' utilise la matrice de similarité ré-estimée après chaque fusion.

L'organisation de modèles MMG locuteurs en structure arborescente est un choix permettant de réduire considérablement la complexité du traitement lors du passage à l'échelle incrémental. L'algorithme proposé dans ce chapitre, permet de grouper deux modèles MMG en un modèle MMG ayant un nombre de composante gaussienne éguale à celui d'un des MMG source. Le modèle MMG résultant de la fusion permet de représenter à la fois par le critère de similarité et aussi par la vrasemblance de données des MMG fusionnés. L'évaluation de l'algorithme de fusion du modèle est effectué à l'aide de deux critère : BIC et la divergence de Kullback-Leibler. Ces deux critère permet de valider l'efficacité de cette technique à produire des modèle qui permet de représentés à la fois les deux modèles source ainsi les données pour les quelles ont était crées. La mesure de divergence de Kullback-Leibler est réutilisé dans la phase de l'exploration, les mesures de KL entre modèles fils et le modèle père sont quasiment identiques et inférieur à la distance présenté entre les deux modèle fils déjà jugé similaire.

Par la suite, nous avons effectué des comparaisons statistiques des différentes méthodes de construction de l'arbre de recherche binaire qui ont révélées les bonnes propriétés de la technique de regroupement par fusion. Les premières comparaisons ont permis de vérifier la topologie de la structure par rapport à la complexité de sa construction, les deux méthodes appliquées (avec ou sans mise à jour de la matrice des distances des MMG locuteurs) produitent un arbre de recherche moyennement équilibré. Néanmoins, la performance d'exploration reste très faible à cause du manque d'information lié à l'évolution ascendante de la structure. Nous allons présenter dans le chapitre suivant une nouvelle approche d'organisation des MMG locuteurs. Premièrement, la structure arborescente ascendante nécessite la création des noeuds MMG suplémentaires pour chaque niveau de l'arbre

minimum $(N-1)$. En plus, la racine de l'arbre est un modèle MMG qui représente l'ensemble des MMG existants produit des erreur biaisés de calcul de similitude avec le MMG requête.

CHAPITRE 6

Structure arborescente binaire adaptée au problème incrémental

Ce chapitre a pour objectif de présenter une nouvelle proposition d'organisation des MMG sous forme d'une structure arborescente binaire descendante. L'arbre binaire de recherche est construit directement à partir des MMG détectés. Les noeuds et les feuilles sont représentés par des vrais MMG locuteurs. L'approche proposée dans ce chapitre permet de résoudre les inconvénients liés à la construction d'un arbre binaire non équilibré. Pour cela, une nouvelle proposition de structure de MMG offre des fonctionnalités pratique dans le cas du traitement d'un flux audio tel que la mise à jour des MMG.

La mise à jour des modèles locuteurs est effectuée grâce à l'algorithme de fusion des MMG discuté précédemment (voir section. 5.2.3). Nous citerons dans un premier temps les motivations principales des techniques de classification hiérarchique des MMG. Dans un deuxième temps, nous présenterons un algorithme d'indexation adapté au problème incrémental. Pour cela, l'organisation des modèles MMG peut être actualisée dans le but de minimiser sa profondeur dans le cas d'un arbre binaire par exemple. Ceci permet de minimiser la profondeur de l'arbre en rendant la structure plus équilibrée.

6.1 Contexte de travail

Comme nous l'avons déjà mentionné au chapitre précédent, l'organisation hiérarchique des MMG locuteurs permet en premier temps de résoudre partiellement le coût d'exploration et de recherche dans les bases de données au sens de locuteurs. L'organisation des MMG sous un arbre binaire ascendant en utilisant le dendrogramme basé sur le critère de similarité ne permet pas de maîtriser la topologie de la structure. De plus, au cours du processus

de création de l'arbre, des noeuds qui représentent les sous-classes sont ajoutés au nombre des vrais modèles de locuteurs de bases à explorer. Cela peut conduire à des structures peu représentatives et coûteuses.

Etant donné un ensemble d'objets décrits par un nombre fixe d'attributs, l'objectif d'une tâche de classification non supervisée [42, 34] consiste à proposer une partition des objets en k sous-ensembles où le paramètre k est le nombre de regroupements attendus par l'utilisateur. Une variante de cette tâche est de ne pas utiliser le nombre attendu de regroupements comme une donnée du problème. Plusieurs stratégies permettent la recherche des regroupements dans l'espace de toutes les partitions. Deux méthodes peuvent être distinguées, les méthodes procédant par partitionnement, les méthodes hiérarchiques (ascendantes ou procédant par divisions successives) et les méthodes basées sur les densités et les méthodes de quantification. Dans certaines applications, les données regroupées ne sont pas représentées par des vecteurs d'attributs, elles peuvent par exemple être représentées par des graphes, des listes d'attributs de longueur variable, des mots ou collection de mots, des images etc. La seule information accessible est alors une mesure de similarité entre les objets. Plusieurs algorithmes de classification non supervisée [42] peuvent être appliqués à partir de la matrice contenant les mesures de similarité entre les objets pris deux à deux. Les données peuvent être assignées à des regroupements, qui peuvent être associés jusqu'à obtenir une hiérarchie de partitions. La présentation hiérarchique des données se révèle souvent utile car la relation de catégorie à sous-catégorie existe dans la plupart des applications.

Cependant, de nombreux travaux d'organisation hiérarchique des MMG consistent à représenter un ensemble de MMG par une sous-classe à l'aide d'un seul modèle GMM. Certaines techniques utilisent par exemple la réduction des paramètres de tous les modèles à grouper après leur concaténation [32] ou encore le travail présenté dans (section 4.3). L'organisation des MMG locuteurs proposée dans cette section repose sur l'utilisation de la distance de KL_m présentée dans (section. 4.2). Cependant, cette mesure peut être considérée comme une distance uniforme permettant de représenter tous les MMG de locuteurs dans un espace affine euclidien de dimension 2. En effet, ceci suppose que la représentation des locuteurs à l'aide de modèles MMG est performante en terme de reconnaissance de locuteur. Rappelons que dans notre contexte de travail, la reconnaissance de locuteur est réalisée en mode indépendant du texte. La durée des échanges à tour de rôle de parole lors d'une conversation au minimum de 3 à 4 seconds, sachant que les locuteurs ne parlent pas simultanément. En outre, la performance de la représentation de l'identité du locuteur via un modèle de MMG dépend de plusieurs conditions d'application. La contrainte de la taille des données utilisées dans la phase d'entraînements nous oblige à utiliser un seuil pour fixer une taille minimale de données avant de passer à la création de modèle. En effet, l'étude est réalisée sur

un nombre $(N = 20)$ limité de locuteurs afin de tester la performance de la représentation. La performance de la représentation de données acoustiques au sens du locuteur varie proportionnellement avec le nombre de composantes de son MMG.

6.2 Classification hiérarchique des modèles de mélange de gaussiennes

Cette section présente les techniques de classification hiérarchique des modèles de mélange de gaussiennes afin de réduire le coût de recherche linéaire. Au début, nous rappelons les étapes du processus d'indexation par locuteur ainsi que la génération de modèle de locuteur. Puis, nous allons proposer une méthode d'organisation des MMG à l'aide d'un arbre binaire descendant (crée à partir de la racine). Le schéma d'organisation est incrémental, en outre, la mise à jour des modèles est effectuée à l'aide d'algorithme de fusion des MMG utilisé auparavant dans la création d'un arbre binaire ascendant (voir section 5.3.1).

6.2.1 Processus d'indexation et classification hierarchique

L'indexation automatique du locuteur est définie comme la reconnaissance de celui qui a parlé et quand est-ce qu'il a parlé, sur une zone de parole prononcée par un ou plusieurs locuteurs, dans un dialogue multi-locuteurs. Toutefois, une des tâches les plus importantes en indexation de locuteur est la segmentation de la parole en zones délimitant l'identité des différents locuteurs parlant et en zones coïncidant avec le silence, avec une grande résolution temporelle et un grand taux d'identification. Les tests en laboratoire ont montré que la meilleure performance est obtenue pour une durée segmentale de 3 secondes en donnant un taux d'erreur moyen de 7.65%. Ce qui implique que le taux d'indexation vaut 92.35% (pourcentage de segments correctement indexés) [59].

Dans le cas de flux en continu, après chaque détection de changement de locuteur, un nouveau MMG de locuteur est généré à partir d'un du segments de descripteurs correspondants. Le coût d'un test d'identification d'un modèle requête augmente considérablement en fonction du nombre des modèles existants (notamment dans le cas d'un nouveau modèle détecté).

Le système d'indexation s'appuit sur des techniques utilisées dans le domaine de reconnaissance automatique de locuteur. Un prétraitement composé de trois étapes : extraction des descripteurs pertinents vis-à-vis des locuteurs, segmentation non supervisée basée sur le test d'hypothèses Bayésien et la représentation en modèle de MMG. Notre travail se focalise sur la représentation des modèles de locuteurs en structure arborescente afin de

réduire le coût d'exploration lors du passage à l'échelle incrémentale. Le schéma thématique du système d'indexation proposé dans ce chapitre est illustré par (figure 6.1).

Notre proposition est basée sur une structure arborescente binaire. Dans un espace de mesure de similarité basé sur la distance KL_m, l'idée consiste à répartir les MMG des locuteurs sous forme d'une structure binaire équilibrée. Pour cela, nous ferons appel à l'histogramme de distance L_m d'un ensemble de MMG locuteurs de référence comme un paramètre de création de l'arbre de recherche garantissant une meilleur répartition.

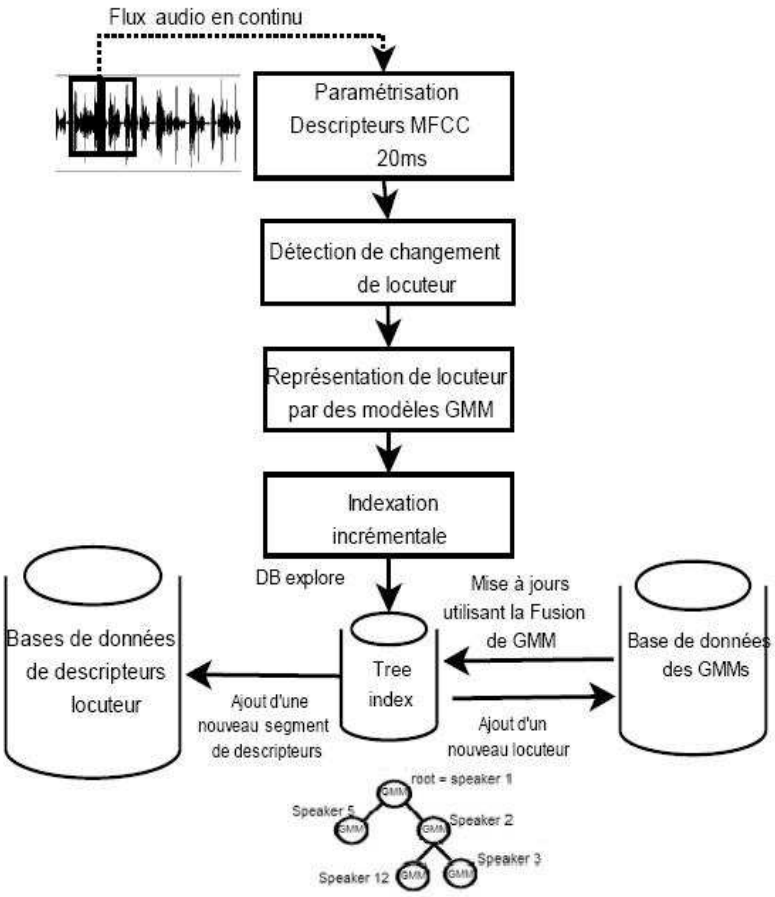


Figure 6.1 – Synoptique du système d'indexation au sens de locuteur adaptée au flux incrémental

De nombreux travaux proposent une organisation arborescente des modèles de locuteurs grâce à une classification hierarchique. L'utilisation du modèle d'ancrage dans [49] [81] se base sur la mesure de distance des vecteurs avec le critère de maximum de vraisemblance, ce qui permet leur répartition suivant une structure arborescente

d'ancrage. Le présent travail propose une technique de classification hiérarchique en utilisant une représentation par (arbre binaire ou encore n-aire). ce principe est partagé avec plusieurs travaux utilisant les MMG pour la RAL dans le cas de grands volumes de données, en particulier [82] qui proposent une représentation en *Multi-Grained Modeling*.

Par la suite, nous allons proposer une approche incrémentale et décrire le processus de création de la structure arborescente. L'algorithme présenté dans (figure 6.6) repose sur la souplesse et la flexibilité de la représentation statistique des données acoustiques vis-à-vis du locuteur à l'aide des modèles de mélange de gaussiennes. Nous allons citer les avantages justifiant le choix de cette technique :

Avantages :

- L'arbre binaire est construit entièrement à partir des MMG détectés, c'est-à-dire que les noeuds et les feuilles seront représentés par des vrais MMG locuteurs,
- Le processus d'indexation garantissant la mise à jour des modèles, est réalisé à l'aide de l'algorithme de Fusion des MMG,
- Le traitement est incrémental adapté au flux audio continu (cas d'indexation des données provenant d'une station radio).

Inconvénients :

- L'organisation de l'arbre dépend de l'ordre d'arriver des modèles,
- Aucune notion de similarité ni de vues.

6.2.2 Arbre de recherche binaire pour l'indexation de locuteur

L'organisation des MMG locuteurs repose principalement sur la performance de la mesure de distance à l'aide la divergence de KL_m. En effet, la structure est incrémentale au fur et à mesure que les modèles MMG sont générés suite à une détection de changement de locuteurs. La détection automatique de changement de locuteur est une étape de segmentation en locuteur qui produit l'événement requête déclanchant le processus d'indexation. Le principe de système d'indexation par locuteur proposé pour un flux en continu doit assurer trois fonctions principales :

1. initialisation des paramètres système tels que η seuil utilisé pour l'exploration de l'arbre et ϵ utilisé pour l'identification ;

2. segmentation en continu du flux audio permet de déclencher l'événement "nouveau segment/nouveau modèle" ;

3. l'exploration de l'arbre de recherche permet de décider l'ajout, la mise à jours ou le rejet du modèle requête crée à l'étape 2 ;

L'initialisation du seuil η est réalisée après un calcul d'histogramme de distance KL sur des modèles de référence. Cette initialisation est très recommandée pour une organisation meilleure de MMG à l'aide d'un arbre binaire bien équilibrée. En revanche, dans le cas du manque d'informations *a priori* sur le signal, le seuil η est choisi aléatoirement au début, puis estimé à partir d'un nombre des MMG inscrits dans la base de données. La deuxième étape, consiste à segmenter le signal audio entrant par locuteur à l'aide des techniques de détection de changement de locuteur (voir section 3.3.3). La troisième étape, consiste à interroger le système à l'aide du modèle MMG récemment détecté. Cependant, le modèle requête est testé en premier lieu avec la racine de l'arbre binaire s'il existe, sinon le modèle requête représentera la racine de l'arbre soit le premier modèle MMG détecté. En revanche, le modèle requête est redirigé à un nouveau test avec le modèle MMG noeud fils (droite ou gauche) selon la valeur (supérieur ou inférieur au seuil η) de la distance mesurée avec le modèle MMG noeud père (voir schéma illustrée par la figure. 6.6).

Algorithme et processus de création de l'arbre :

Initialisation :

- Le seuil est choisi à partir d'un ensemble des modèles de référence ou à l'aide d'un algorithme adaptatif.
- La racine de l'arbre (root) : (premier MMG de locuteur détecté).

Première étape :

L'algorithme se met en attente d'un nouveau modèle de locuteur généré par la phase de segmentation progressive.

Deuxième étape :

1. Affectation de nouveau modèle (à gauche ou à droite) selon la distance de ce dernier avec le noeud père (à commencer de la racine),
2. Si le nouveau modèle MMG existe déjà, on le fusionne avec le modèle MMG existant afin de le mettre à jour si nécessaire,
3. Retour à la première étape.

6.2.3 Choix du seuil

La phase de l'initialisation est très importante dans le processus de création de l'arbre binaire. Une première étude consiste à trouver le seuil initial pour établir une répartition équilibrée des MMG locuteurs. Le calcul du seuil noté η consiste à chercher une mesure moyenne de la distance la plus fréquente entre les MMG de locuteurs dont nous disposons à un instant t. Le seuil de répartition des MMG est défini comme suit :

$$\eta = \max\left[hist\left(\text{de mesure de distance entre tout MMG}\right)\right] \tag{6.1}$$

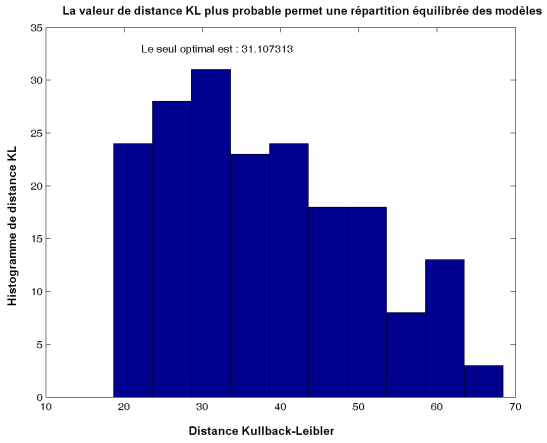
Le seuil d'identification noté ϵ tel que ($\epsilon \approx 0$) est donc choisie manuellement sa valeur peut changer selon les variantes de la formule de KL. Ensuite ψ représente la distance de KL_m entre le modèle requête et le modèle noeud, tel que $\psi = KL_m(Noeud, \text{MMG-reqûete})$.

A partir de la matrice de distance entre les MMG référence le calcul de l'histogramme des distances KL_m, nous permet de définir une valeur ou un intervalle, dont la plupart des MMG locuteurs ont une distance qui apparaît fréquemment. Un exemple d'histogramme calculé à partir d'une base de référence de 20 locuteurs est présenté dans (figure. 6.2).

Deux seuils sont alors utilisés pour la création de la structure binaire de recherche, le premier noté η précédemment cité, permet d'orienter l'exploration dans la structure. Le deuxième noté ϵ, représente le seuil utilisé dans la phase de l'identification entre modèle requête et le modèle noeud de l'arbre. Le cas échéant, le modèle sera ajouté comme fils (gauche ou droit) du dernier noeud qui ne possède pas un ou les deux fils concernés par l'addition. Dans l'exemple illustré dans (figure. 6.7), l'organisation hiérarchique incrémentale des MMG de locuteur, est basée principalement sur la mesure de divergence de KL. En outre, le processus de la construction de la structure est basé sur deux opérations : **Ajouter** ou *mettre à jour* le modèle requête. L'exploration de la structure permet soit : d'identifier le modèle requête avec ceux qui existent, dans ce cas l'opération de la mise à jour est effectuée à l'aide de l'algorithme de fusion décrit dans (section. 5.2.3), soit : ajouter directement le modèle requête à droite ou à gauche à l'endroit ou l'exploration est aboutie sans trouver un modèle similaire.

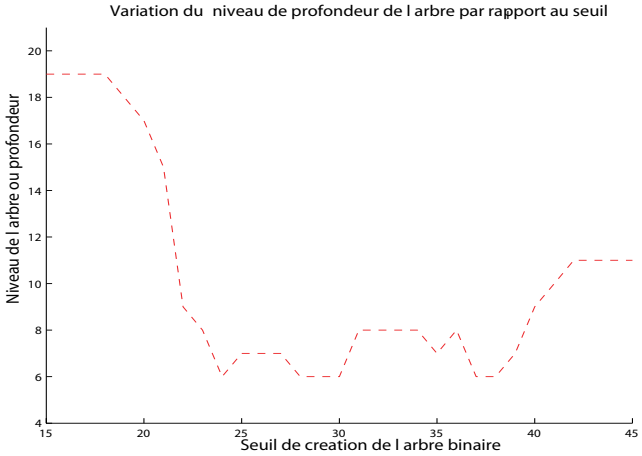
.

Comme présenté dans (figure. 6.4), le seuil η est choisi à partir de l'histogramme des mesures de KL_m de tout les MMG locuteurs inscrits à l'instant t donné. Le but est de répartir les modèles de locuteurs sous un arbre bi-



a)

Figure 6.2 – Longueur de plage des mesures de distance entre un ensemble des MMG locuteurs



b)

Figure 6.3 – Le seuil optimal est donné par le minimum des niveaux de profondeur soit un intervalle donnant lieu à une profondeur minimale

naire équilibré, par conséquence, le niveau de profondeur de l'arbre diminue considérablement équivalent à une recherche avec un coût optimisé. Le nombre de tests maximum sera donc $P + 1$, tel que P est la profondeur de l'arbre binaire de recherche nécessaire pour une identification rapide.

Dans le cas d'indexation de flux audio, le système reçoit des MMG locuteurs en continu, au fur et à mesure que le changement de tour de parole est détecté. De plus, l'organisation des modèles selon une structure arborescente est conçue pour un processus sans connaissance *a priori*. Cependant, le choix du seuil η peut être adaptatif, c'est-à-dire que l'initialisation du seuil sera aléatoire au début du processus de création de l'arbre, et au bout d'un certain

nombre de modèles inscrits dans la base de données de MMG, le seuil peut être estimé, selon les informations données par l'ensemble des modèles au fur et à mesure que le nombre de modèles afin d'assurer d'une manière continue une structure binaire équilibrée.

L'étude illustrée par l'histogramme donné par (figure. 6.3), montre que le seuil peut appartenir à un intervalle large, tout autant que l'architecture de l'arbre de recherche respecte une organisation qui réduit la complexité de l'exploration.

La variation du seuil η implique un changement de structure ainsi le niveau d'équilibre de l'arbre, de plus, un choix adéquat du seuil η permet de réduire au minimum la profondeur de l'arbre. Cependant, le seuil est choisi à partir d'un histogramme de mesure de KL_m de toutes les combinaisons possibles des locuteurs déjà identifiés. Le seuil est déduit à partir du maximum donné par l'histogramme.

L'approche proposée peut être testée directement sur un flux radio (exp. *Europe1*[1]). Le processus d'indexation est réalisé en trois modules : d'abord, le module de segmentation par locuteur produit des segments descripteurs MFCC. En suite, à l'aide d'algorithme EM permet de générer le MMG_requête (composé de 16 gaussiennes de dimension 26) suite à chaque changement de locuteur (voir figure. 6.5). L'étape de détection de changement de locuteurs ne prend en considération que les changements de rôle de parole supérieur à (5 secs). En effet, la fiabilité du modèle MMG généré dépend directement de la taille des données utilisées pour l'apprentissage. Pour cela, La mise à jour des modèles déjà inscrits dans la structure arborescente consiste en une solution alternative à l'insuffisance de données d'apprentissages. Cependant, la mise à jour des modèles de la structure continue au fur et à mesure que les données correspondantes sont détectées. La mise à jour est réalisée grâce à l'algorithme de fusion entre GMMs.

6.2.4 Arbre de recherche binaire ascendant des MMG

L'arbre binaire proposé se base sur l'utilisation d'une mesure de distance entre MMG. La répartition des MMG suppose que l'espace de la distance de KL_m est un espace euclidien uniforme. Cependant, Les MMG estimés avec suffisamment de données d'apprentissage (supérieur à *5 secs*) permettent une meilleure organisation de répartition des modèles de locuteurs représentés dans le cas du passage à l'échelle. Le segment de données acoustiques (≥ 5

[1] exemple de flux mms : "mms ://vip8.yacast.fr/encodereurope"

secs) est composé de 100×5 vecteurs MFCC extrait d'une bande audio échantillonnée entre 11000Hz à 44100Hz, selon le type de codage et de transmission. Dans un processus d'indexation par locuteur qui permet de gérer un grand nombre de MMG, il est recommandé d'utiliser le modèle MMG-UBM (Universel Background Model) afin d'entraîner les données des segments avec une initialisation basée sur un modèle référence universel [63].

Résumé du processus de création de la structure hiérarchique incrémentale

1. **initialisation**
   - initialisation des paramètres η et ϵ ;
   - racine : premier modèle arrivé.
2. **Attendre un événement : détection d'un nouveau modèle**
3. exploration de l'arbre :
   - 1^{er} cas : modèle existe déjà dans la base donc aucune modification.
   - 2^{me} cas : modèle n'existe pas dans la base, ce qui implique mise à jour de l'arbre et ajout de ce dernier.
4. répéter à partir de l'étape 2.

Dans un cas réel, le processus est appliqué sur des flux en continu d'une station radio de durée de 4 heures. L'étude de performance est difficile à faire manuellement à cause de la taille du document source. Cependant le test de validation des approches proposées sera premièrement effectué sur des bases de données limitées (voir figure. 6.7).

L'organisation hiérarchique incrémentale d'un flux de durée 4 heures est illustrée dans (figure. 6.7). A partir de 72 changements de locuteur détecté, 46 modèles ont été crées, cependant, 16 locuteurs parmi les 46 qui ont été identifés. Le niveu de profondeur de la structure arborescente obtenue est de 5, ce qui réduit d'une manière significative le nombre de tests dans la recherche parmi les 16 en mode de recherche linéaire.

6.2.5 La performance de la structure proposée face au problème incrémental

La méthode proposée dans ce chapitre, une nouvelle technique dont le but est de réduire de plus le coût du système d'identification dans le cas d'un traitement incrémental. En outre, l'avantage de cette approche est que la structure est créer sans avoirne nécessite pas de prétraitement ou des information *a priori* sur les modèles MMG, de plus la création est réalisée dans la phase l'identification via une exploration menée à l'aide de KL_m. En effet, la répartition des MMG à structurer selon une approche incrémentale est réalisée en mode non supervisée et sans

aucune connaissance *a priori*. Notre approche répond fortement au problème incrémental d'arrivée et d'indexation des MMG entrants (voir figure. 6.8). Néanmoins, la performance d'identification souffre encore du problème de la variabilité interne des données ainsi que l'insuffisance des données d'apprentissage.

Le test mené sur le choix du seuil est validé par la courbe (voir figure 6.3), nous constatons que le seuil qui permet une répartition équilibrée des MMG dans un espace de distance KL peut être initialisé aléatoirement, puis un seuil adaptatif peut être utilisé afin de réorganiser les modèles pour obtenir une profondeur d'ordre minimum. Le coût de la réorganisation de l'arbre de recherche est très coûteux se qui permet de réorganiser les MMG sous forme d'un arbre binaire équilibrée à l'aide d'une adaptation du seuil de répartition.

En effet, La variation du seuil implique un changement de la structure et le niveau de l'arbre le plus optimal est celui qui minimise la valeur de la profondeur de l'arbre, sachant que le seuil, peut être choisi dans la plage de distance donnée par la matrice de distance par un histogramme. Sachant que ces mesure de distance KL_m sont données par l'ensemble ou une partie des modèles déjà inscrit (voir figure. 6.4).

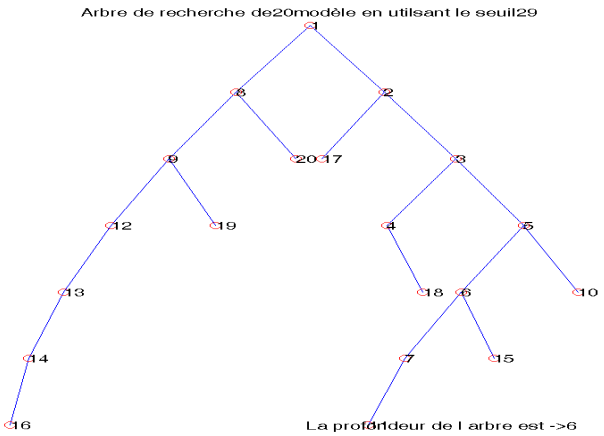
6.3 Conclusion

Dans ce chapitre nous avons présenté une réponse au problème la représentation des MMG par un graphe. La structure hiérarchique decscendante est mieux adapté au schéma incrémental et ne nécessite pas des connaissances *a priori*, alors que le seuil peut être choisi avec des méthodes classiques, arbitraires ou adaptatives. En outre, afin d'assurer une fiabilité de la représentation face au problème de l'insuffisance des données d'apprentissage, dans ce cas, l'algorithme de fusion est fortement sollicité. Le modèle crée à partir d'un segment de descripteur de taille importante et ayant une distance de KL_m faible avec le modèle correspondant déjà inscrit dans l'arbre de recherche seront fusionnés, avec ce moyen, le modèle résultant de la fusion est ré estimé à moindre coût et bénéficiera d'une meilleure représentation.

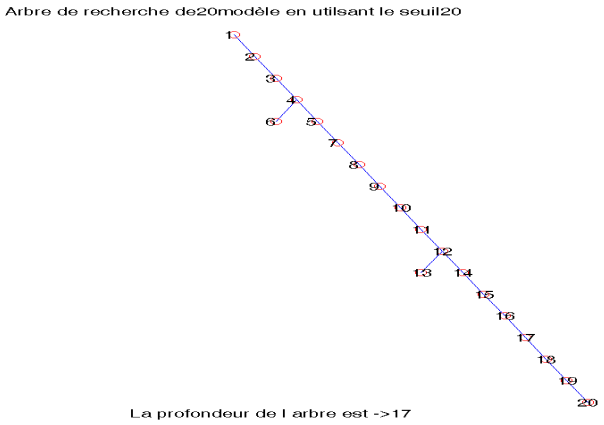
La technique proposée dans ce chapitre est adaptée aux conditions d'un schéma incrémental pour l'indexation par locuteur. La création de la structure commence par une exploration de l'arbre suivi les opérations (ajout, mise à jour ou rejet) sont appliquées au fur et à mesure que les modèles de locuteurs arrivent. L'ajout des modèles et la mise à jour dépendent de la mesure de KL_m entre le modèle requête et le modèle noeud de l'arbre corespondent.

L'utilisation des deux seuils afin d'organiser l'arbre de recherche met cette technique face à un risque majeur de problème de paramétrisation et d'adaptation liée aux données source à traiter.

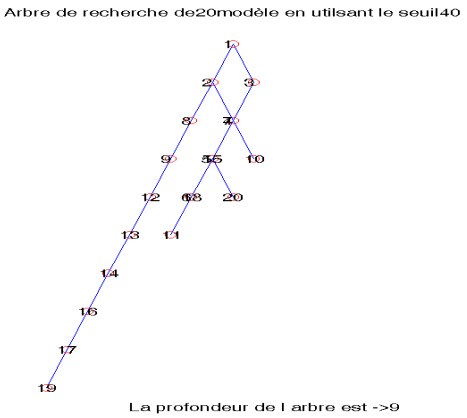
Dans le chapitre suivant, nous allons aborder ce problème d'une autre facette, le problème d'utilisation de seuil sera au centre de notre intérêt. Nous allons développer une technique de calcul de probabilité liée à la perte d'information dans le cas d'une organisation hiérarchique des MMG. Le but est de mettre en oeuvre un système de classification hiérarchique par regroupement des modèles statistiques en proposant une approximation de recherche linéaire optimisée.



a).



b).



c).

Figure 6.4 – a). Arbre binaire de recherche bien équilibré pour $\eta = 31.107313$; b). L'Arbre est balancé à droite $\eta = 20$; c). L'arbre est complètement blancé à gauche pour $\eta = 20$

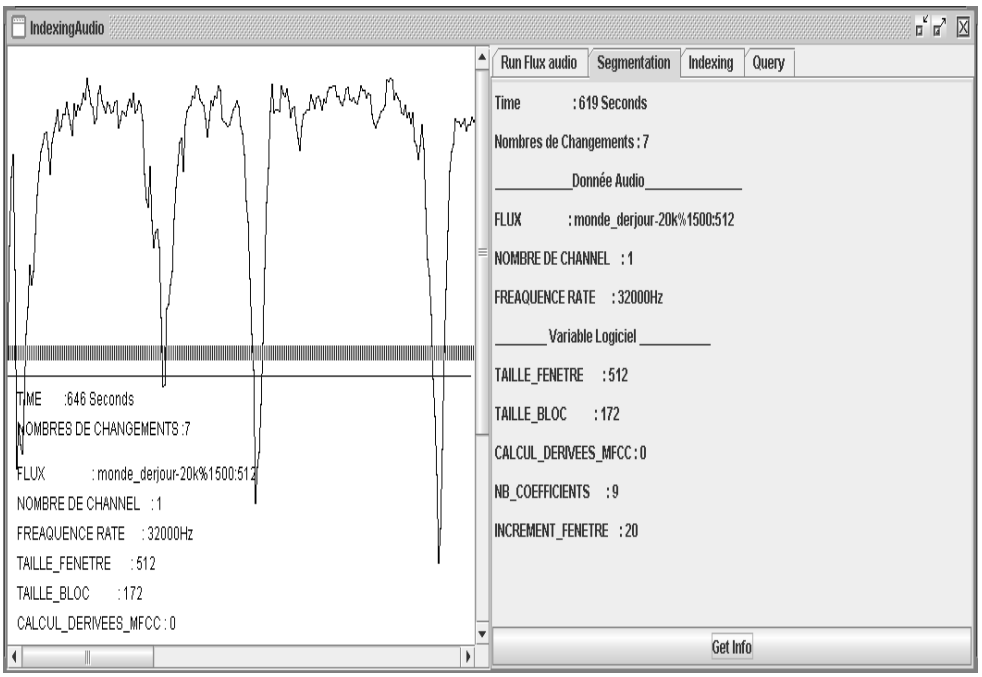


Figure 6.5 – Segmentation incrémentale au sens de locuteur

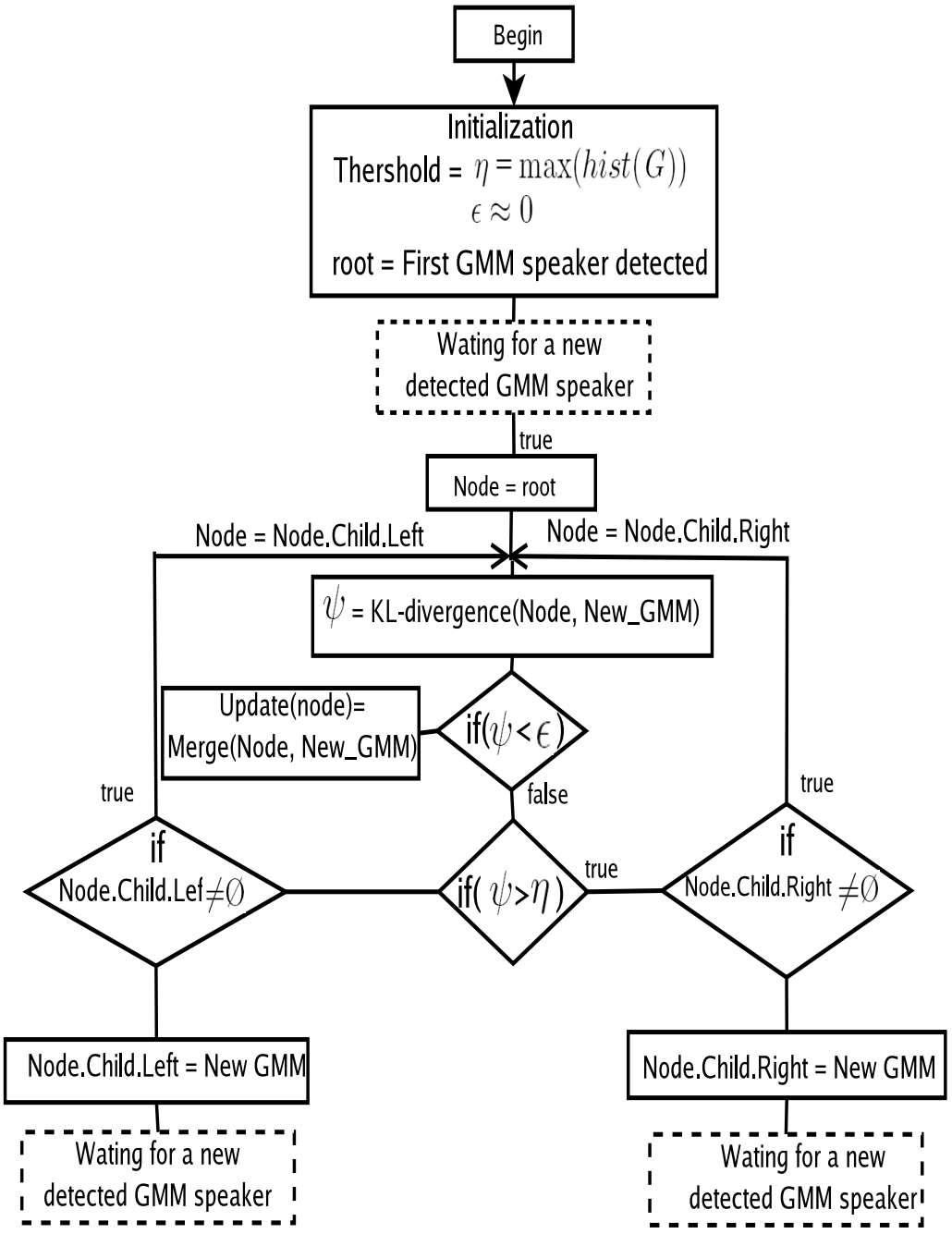


Figure 6.6 – Processus de création d'une structure binaire descendante des MMG

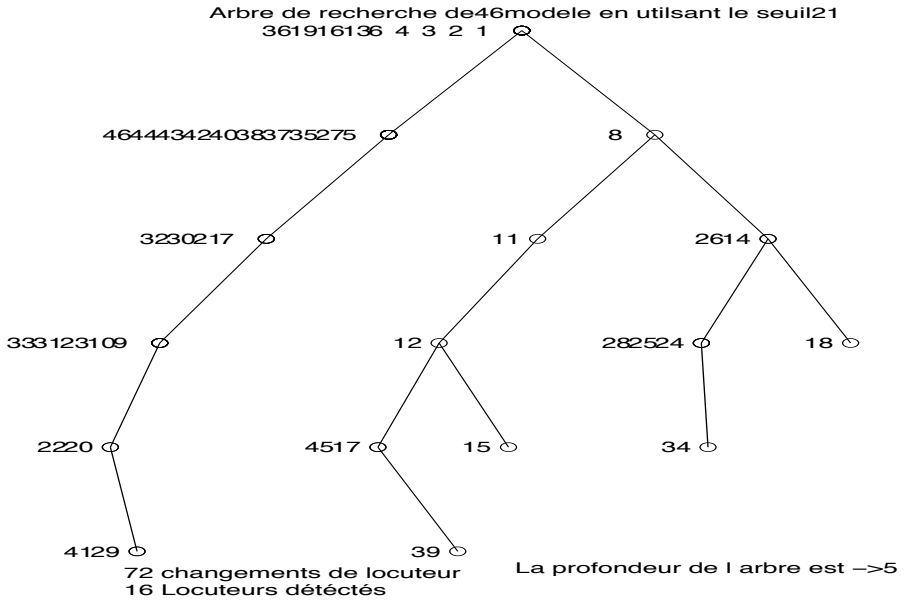


Figure 6.7 – L'organisation hiérarchique incrémentale d'un flux de durée 4 heures

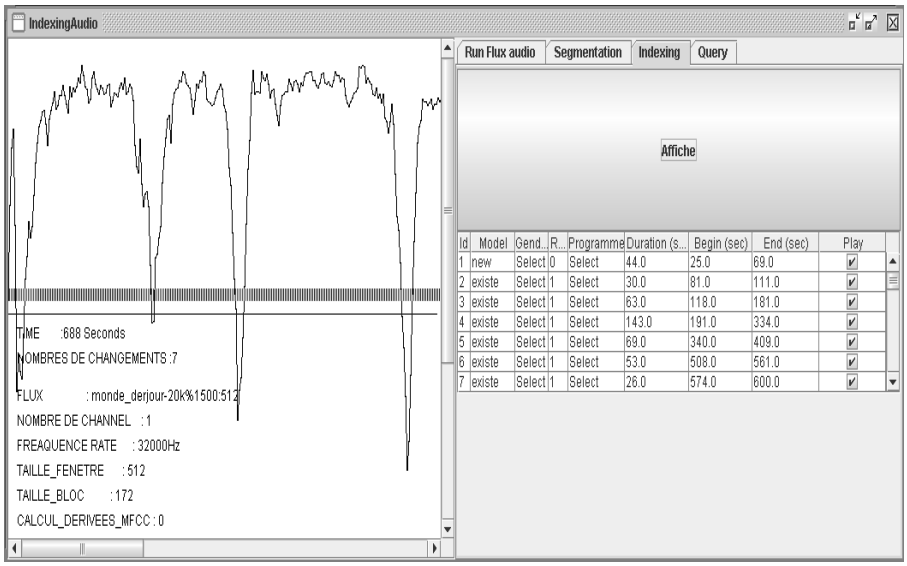


Figure 6.8 – Table d'index des MMG locuteurs et méta données

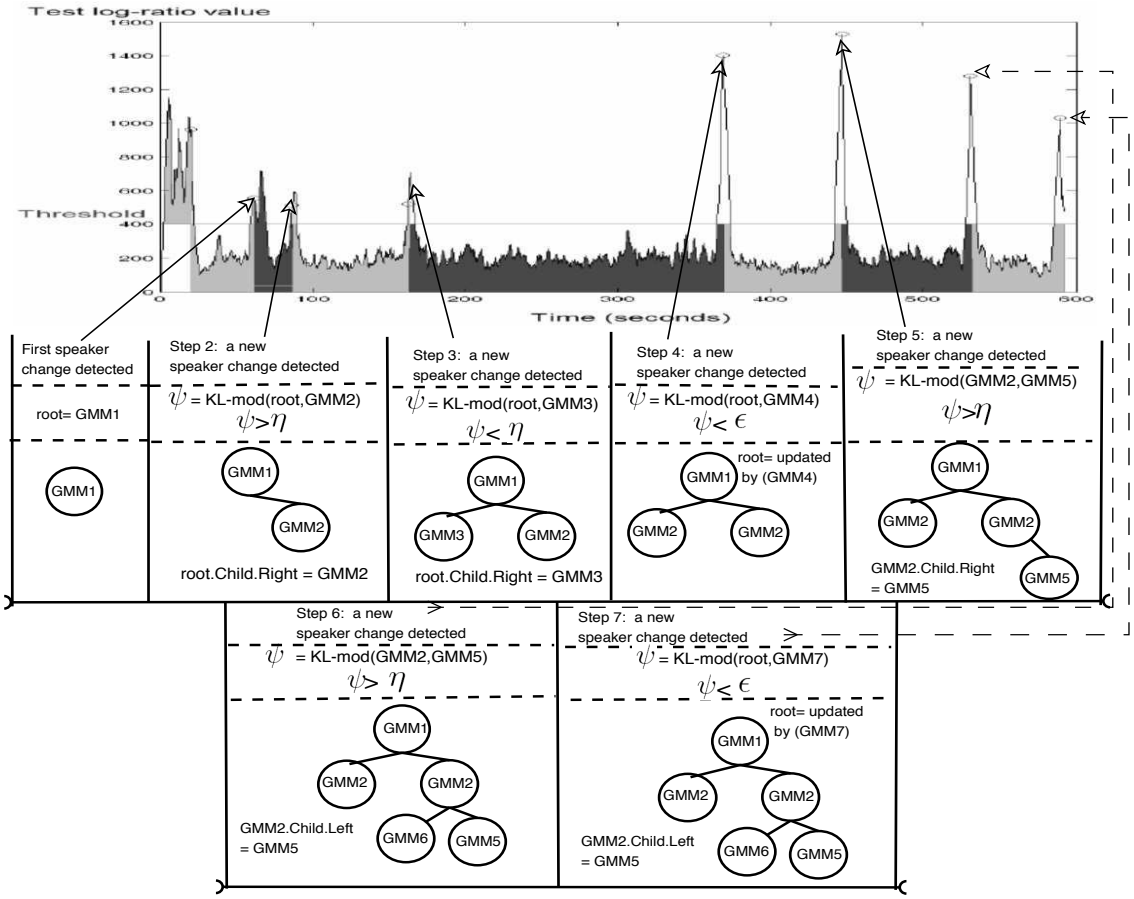


Figure 6.9 – Processus de construction de l'arbre binaire en utilisant des données réelles

CHAPITRE 7

Organisation hiérarchique des MMG locuteurs

Nombreuses sont les applications de reconnaissance automatique qui font face à une tâche coûteuse d'évaluation de données en utilisant le critère de maximum de vraisemblance, notamment, dans le cas d'un grand nombre de candidats à tester [81, 43]. Dans le cas général, le problème de la reconnaissance automatique de locuteurs agit sur les performance d'un tel système d'indexation et de recherche par contenu dans des archives audio. Plus précisément, nous proposons de réduire la complexité de temps de requête, par une organisation hiérarchique *a priori* des modèles des locuteurs. L'organisation hiérarchique est à l'origine d'une technique classique de sous classification en utilisant une représentation vectorielles dans un espace multidimensionnel. A cet effet, ce chapitre introduit une étude comparative entre deux propositions via une approximation de la recherche linéaire optimisée sur un dendrogramme de MMG généré à l'aide du critère de similarité.

A l'aide de l'expression modifiée de la divergence de Kullback-Leibler entre le noeud père et le neoud fils, des techniques d'optimisation de recherche optimisée sont dérivés. En outre, l'expression statistique de la relation noeud père et noeud fils permet alors de développer une nouvelle proposition efficace de création à la construction d'un arbre de modèles, utilisant des technique de classification (*dendogram-based ou K-mean like*).

7.1 Objectifs et motivations

L'importante tâche d'indexation par contenu, permet la navigation et la recherche dans de grands volumes de données audio, ce qui nécessite une technique *a priori* de structuration temporelle, via un étiquetage des entités extraites tel que l'assignement de l'identité du locuteur a un segment temporel (cette tâche est connue aussi comme

suivi de locuteur : speaker diarization). Cet axe de recherche est fortement sollicité par un grand nombre de travaux, notamment, le domaine d'indexation par contenu des données multimédia [55, 71, 48]. Notre travail est focalisé sur le traitement de reconnaissance de locuteur en mode indépendant de texte, appliqué sur des archives audio contenant que de la parole. Un prétraitement classique de segmentation en locuteur est réalisé en premier lieu par notre système d'indexation qui est décrit dans le chapitre 3.

Dans le cas idéal, l'indexation d'un flux audio est réalisée en mode incrémental. Deux phases d'indexation impliquent l'utilisation de la technique d'identification de locuteur. Premier cas, lors de la phase d'étiquetage, quand deux différents segments temporels représentent le même locuteur. Deuxième cas, lors d'une requête formulée par un utilisateur. Le traitement basé sur un algorithme incrémental est alors nécessaire lors de la manipulation de la base de données des modèles de locuteurs soit par un ajout d'un nouveau modèle ou d'une mise à jour avec plus d'informations. Le choix du schéma est alors justifié par une utilisation de modèles génériques au lieu des techniques discriminantes entre modèles de locuteurs.

Une solution classique de la reconnaissance automatique de locuteur consiste à explorer linéairement l'ensemble de $S_1, ..., S_M$ de modèles de locuteurs inscrits. L'évaluation peut être réalisée à l'aide de différentes techniques, une des fameuses techniques utilisées est le calcul de maximum de vraisemblance entre les données du locuteur requête et tout les modèles candidats. L'objectif de ce travail est de pouvoir créer un système de recherche par locuteur dans le cas de grands volumes de données. Nous proposons une organisation d'un ensemble de modèles de locuteurs candidats sous forme d'un arbre binaire de recherche, dans le but d'obtenir complexité de recherche logarithmique et non linéaire (*i.e :* $\prec O(M)$) qui permet d'optimiser le coût de calcul en terme de temps d'évaluation.

Il existe des travaux alternatifs qui permettent de réduire le coût de recherche, basés sur l'optimisation de la complexité des algorithmes. A cet effet, deux voies sont envisagées : adapter des stratégies de recherche ou de construire une structure de modèles adéquats sous espace cepstral [58, 87, 82]. En outre, l'organisation à l'aide d'un modèle d'encrage permet d'attribuer pour chaque modèle un vecteur de scores par rapport à un ensemble de modèles de références [49, 81]. Toujours dans le but de réduire le coût de la recherche des modèles statistiques de locuteurs, une technique alternative a pour objectif de réduire les paramètres du MMG. Ces derniers peuvent représenter une classe de MMG de locuteurs (un ensemble de MMG de locuteurs partageant une propriété) tout en préservant les caractéristiques pertinentes. Ces approches proposent une optimisation du coût et de la fiabilité

de la recherche. Notre travail de recherche est qualifiée comme complémentaire par des contributions dans le sens horizontal et vertical de l'existant.

Les différentes techniques de recherche d'indexation sont basées sur une conception classique des structures d'index de données multidimensionnelles. Les techniques d'organisation, de gestion et de stockage des données multimédia non demeure moins. Actuellement, la communauté des bases de données met en avant un nombre considérable des contributions basées sur des structures arborescente afin de traiter et gérer des données de type variées (audio, image, vidéo) [7, 86]. La particularité du problème réside alors, dans la nature des entités des index à traités. Notamment le cas de gérer des distributions de probabilité, pour les quelles une structure classique n'est pas appropriée. La recherche par contenu est un domaine en plein expansion qui nécessite l'extension et l'adaptation des techniques de gestion des données multimédia, notamment le cas d'indexation par locuteur dans un flux continu audio, des représentations statistiques des données acoustiques sont utilisées ce qui implique que la complexité des applications de recherche par contenu augmente exponentiellement notamment dans le cas de grands volumes de données.

Ce chapitre est organisé comme suit, la section (7.2) présente des définitions de base permettant de formuler une expression statistique de la relation entre modèle père et un ensemble de modèle fils du même père. Ainsi, la méthode de création du modèle père qui offre une meilleure représentation des modèles fils d'une manière à optimiser la recherche des fils via le modèle père. La section (7.3) détaille les différentes techniques de création d'une structure arborescente en utilisant les outils définis dans la section précédente. Des résultats expérimentaux sont présentés dans la section (7.4), une analyse de performance ainsi que des commentaires seront développés lors d'une conclusion présentée en section (7.5).

7.2 Définition de relation statistique entre les noeuds fils-père

Nous supposons que la technique de reconnaissance est basée sur l'évaluation à l'aide du maximum de vraisemblance des données requête D par rapport à l'ensemble M des modèles candidats. La forme exhaustive de la recherche de locuteur utilisant une forme basique du maximum de vraisemblance, sera comparée avec la technique proposée.

L'objectif est de créer à partir des modèles des locuteurs une forme hiérarchique par regroupement ascendant de M modèles. Afin, de justifier le critère de regroupement proposé ci-dessus basé sur la similarité père-fils, considérons un arbre simplifié, dont deux locuteurs S_1 et S_2 sont présentés par un seul modèle MMG père S_{12}. Cependant, la technique permet une extension directe et arbitraire d'un nombre supérieur de modèle fils.

La réduction du coût en terme de temps de requête durant l'exploration de l'arbre de la racine aux feuilles, est obtenue par le calcul de simple valeur $P(D|S_{12})$ au lieu de $P(D|S_1)$ et $P(D|S_2)$. Par conséquent, S_{12} doit être construit tel que $P(D|S_{12})$ et le plus proche possible de $P(D|S_1)$ et $P(D|S_2)$, dont le but de générer une erreur de classification la plus faible possible que celle d'un test linéaire exhaustif. Le nombre de m_{12} doit aussi être nettement plus petit que $m_1 + m_2$ afin d'assurer une réduction du coût d'évaluation.

Les sous-sections suivantes définies respectivement (section 7.2.1) un critère de création d'un modèle père optimal par rapport au modèle fils donnés et la section (7.2.2) présente une technique d'optimisation du critère de déterminer un modèle père avec des modèles fils donnés.

7.2.1 Évaluation d'organisation hiérarchique par optimisation de la mesure KL entre le noeud père et fils

Dans le cas d'un regroupement de modèles MMG, il est important de mesurer l'information perdue due à ce choix d'organisation. Le système d'exploration d'une structure arborescente fiable doit tenir en compte une mesure d'erreur dans l'estimation du meilleur chemin d'exploration. Cette erreur est considérée comme une information de perte liée au choix de regroupement d'un sous ensemble des modèles d'une structure arborescente. En particulier, l'estimation de perte d'information liée au regroupement de modèle est réalisée sur une optimisation de mesure KL_m entre le noeud père et le nud fils.

La perte prévue dans l'expression du log de vraisemblance donnée par approximation de S_1et S_2 par S_{12} est exprimée comme suit :

$$E_{S_k}[\ln p(D|S_k)] - E_{S_k}[\ln p(D|S_{12})], \text{avec } k = 1, 2 \tag{7.1}$$

Supposant que tous les candidats sont équiprobables, le meilleur modèle de mélange de gaussiennes qui représente $\hat{S}_{12}$ réduisant au minimum cette erreur exprimée par la perte d'information d'une telle sous classification est

ainsi défini :

$$\hat{S}_{12} = \arg\min_{\Im} \left[- \int S_1(x) S_{12} dx - \int S_2(x) S_{12} dx \right] \tag{7.2}$$

Tel que les intégrales engendrent l'espace de vecteurs et $\Im$ l'espace de recherche, discuté ci-dessous. Ceci correspond en fait à réduire au minimum la divergence de Kullback-Leibler $KL(S_{1+2}||S_{12})$ [10], tel que S_{1+2} désigne $\frac{1}{2}(S_1 + S_2)$ d'où :

$$\hat{S}_{12} = \arg\min_{\Im} \left[- \int S_{1+2}(x) \ln \frac{S_{12}}{S_{1+2}} dx \right] \tag{7.3}$$

Le calcul pratique de l'expression (voir équation 7.3) présente un intérêt majeur dans le cas du manque d'une approximation rigoureuse de la divergence dans le cas de mélange de gaussiennes. Afin d'éviter l'évaluation exhaustive en utilisant l'estimation Monte-Carlo [21], nous proposons une expression d'approximation ci-dessous. La linéarité de l'intégrale est appliquée à l'équation (7.2) donnent :

$$\hat{S}_{12} = \arg\min_{\Im} \left[- \sum_{i}^{m_1+m_2} w_{1+2}^i \int N_{1+2}^i(x) S_{12}(x) dx \right] \tag{7.4}$$

Dans chaque terme de la sommation dans (voir équation 7.4), une approximation de mélange de gaussiennes S_{12} est exprimée à l'aide d'une des composantes gaussiennes choisie pour une meilleur approximation de N_{1+2}^i au sens de KL. Par conséquent, l'utilisation de la mesure de similarité notée KL_m, entre le modèle de l'union S_{1+2}. Le modèle S_{1+2} est constitué de toutes composantes des modèles fils. Le but est de réduire le nombre de composantes gaussiennes et ainsi le coût de test de vraisemblance, d'où le besoin d'une approximation de S_{1+2} en S_{12}, tel que :

$$\hat{S}_{12} = \arg\min_{\Im} \left[KL_m(S_{1+2}||S_{12}) \right] = \arg\min_{\Im} \left[\sum_{i}^{m_1+m_2} w_{1+2}^i \min_{j}^{m_{12}} KL(N_{1+2}^i||N_{12}^j) \right] \tag{7.5}$$

L'expression suivante est utilisée pour la comparaison entre le modèle fils et le modèle père dans l'arbre de recherche :

$$KL_m(S_k||S_{12}) = \sum_{i}^{m_k} w_k^i \min_{j=1}^{m_{12}} KL(N_k^i||N_{12}^i), \text{k=1,2} \tag{7.6}$$

La mesure de similarité peut être mesurée facilement avec un coût très faible. Cependant, la divergence de KL entre deux composantes gaussiennes, dont les paramètres sont (μ_1, Σ_1) et (μ_2, Σ_2), bénéficie de l'expression

suivante :

$$\frac{1}{2}\left(\log\frac{|\Sigma_1|}{|\Sigma_2|} + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu 2)^T\Sigma_2^{-1}(\mu_1 - \mu 2)\right) - \delta \tag{7.7}$$

Avec δ est la dimension de l'espace des descripteurs (i.e descripteurs MFCC extraites à partir des échantillons audio). La démonstration de l'équation (7.7) est donnée dans [32] qui optimise l'équation (7.5) dans le but de trouver la fonction de transfert π optimale entre $m_1 + m_2$ composantes gaussiennes de S_1 et les composantes $m_{12} \prec m_1 + m_2$ de S_{12}. Ceci permet de réduire le nombre de composantes dans le mélange S_{1+2} pour construire S_{12}, tout en réduisant au minimum la distorsion entre les densités, dans le sens de KL_m. L'espace $\Im$ de recherche consiste ainsi en général de grouper les composantes de $m_1 + m_2$ en un seul groupe m_{12}.

7.2.2 Recherche du modèle optimal de mélange noeud-père

Dans la pratique, il est coûteux d'effectuer une recherche linéaire dans l'espace de recherche de modèles MMG de locuteurs. Par conséquent, l'optimisation locale du critère de (7.6) avec une organisation itérative est détaillée dans l'algorithme 1. L'approximation donnant lieu à un modèle de groupe est proposée par [33], plus particulièrement le cas d'un groupement hiérarchique des gaussiennes simples (plutôt que des mélanges de gaussiennes). Le processus d'organisation est réalisé suivant le même scénario que l'algorithme classique $k-means$, ce dernier est basé sur la même action d'optimisation locale pour une affectation alternative des éléments en groupes puis estimé le groupe représentant. Dans notre contexte, les éléments sont les composants de S_{1+2} et les représentants est de S_{12}.

Souvent avec *k-means*, l'affectation de π^0 pour la quelle le processus présente une optimisation locale est choisi aléatoirement. Dans ce chapitre, nous proposons des critères plus efficaces d'une initialisation adapté à notre contexte : généralement, les composantes d'un mélange de gaussiennes ne sont pas similaire (i.e : ne sont pas redondantes), ainsi notre proposition consiste à choisir π^0 d'une manière aléatoire et suivant une contrainte, en effet, les composantes appartenant au même mélange de gaussiennes ne doivent pas être groupées dés la phase initiale. En revanche à l'aide du schéma itératif du processus les composantes gaussiennes initialement définies aléatoirement peuvent être groupées ultérieurement dans le cas où les données s'impliquent à cette convergence.

7.3 Regroupement des modèles de locuteur

Cette section présente une application de l'approximation statistique de la relation père-fils et une optimisation des techniques précédemment citées. En outre, trois méthodes d'organisation des modèles de locuteurs sous forme d'une structure arborescente seront présentées. Notre proposition permet de créer et de manipuler une seule couche entre la racine et les feuilles, ceci peut être considérées comme une technique de sous-classification.

7.3.1 Dendrogramme par regroupement de modèles

Dans un premier lieu, nous allons présenter une projection des fameuses techniques de classification sur notre problème étudié, appelée classification ascendante hiérarchique, tels que tous modèles de mélange de gaussiennes locuteurs seront présentés par des feuilles. (voir figure. 7.1) :

1. La matrice carrée de similarité $M \times M$ est crée à partir de l'ensemble des modèles inscrits dans la base de données des modèles MMG. La valeur de similarité entre S_1 et S_2 est calculée comme suit :

$$KL_m(S_1||S_2) + KL_m(S_2||S_1) \tag{7.8}$$

2. La structure arborescente est obtenue à l'aide d'un découpage (voir figure "ligne en pointillés" 7.1) de l'arbre à un niveau donné. Le nombre des noeuds au dessus de la ligne de découpage est proche de $\log_2(M)$. Ces noeuds héritent de leur noeuds fils, et les modèles correspondants sont déterminés à l'aide d'optimisation du critère (voir équation 7.5). De même, une structure arborescente avec un nombre variable des modèles est réalisée et sera utilisée dans le cas de la recherche et la navigation.
3. Les modèles les plus similaires sont groupés en un seul modèle (une opération ne permet pas de réduire S_{1+2} en S_{12}, mais permet toujours de garder une représentation des deux modèles), et ainsi de suite, jusqu'a ce qui reste seulement deux modèles. La matrice de similarité est mise à jour à chaque opération de groupement.

Algorithme 1

Algorithme d'optimisation itératif utilisé pour l'estimation de modèle S_{12} réduit, donné par le critère (voir équation 7.6).

Début : initialisation aléatoire avec $\hat{\pi}^0$ (ou donnée si valable)

$it = 0$

répéter

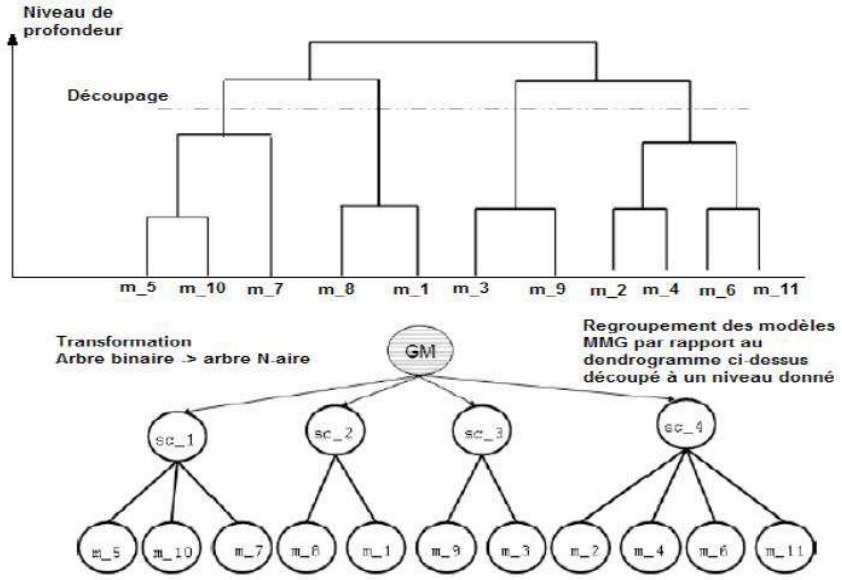


Figure 7.1 – Classification hiérarchique des MMG locuteurs. (Au dessus) dendrogramme est construit en premier lieu, puis (au dessous) le découpage permet de déterminer les noeuds MMG formant la couche intermédiaire

1. **Réduire le mélange S_{12} :**

   pour $\hat{\pi}^{it}$, initialisée aléatoirement ou calculée dans l'itération précédente, la mise à jour des paramètres du mélange de gaussiennes S_{12} est définie comme suit :

$$\hat{S}_{12}^{it} = \arg \min_{S_{12} \in \Im_{m12}} KL_m(S_{1+2}, S_{12}, \hat{\pi}^{it}) \tag{7.9}$$

   Tel que $\Im_{m12}$ est l'espace de mélange avec $m12$ composantes obtenues après un groupement des composantes de $M_{c'}$. En effet, l'estimation et la mise à jour de chaque composantes de S_{12}. Pour une composante j :

$$\hat{w}_{12}^{j} = \sum_{i \in \pi^{-1}(j)} w_{1+2}^{i} \tag{7.10}$$

$$\hat{\mu}_{12}^{j} = \frac{\sum_{i \in \pi^{-1}(j)} w_{1+2}^{i} \mu_{1+2}^{i}}{\hat{w}_{12}^{j}} \tag{7.11}$$

$$\hat{\Sigma}_{12}^{j} = \frac{\sum_{i \in \pi^{-1}(j)} w_{1+2}^{i} (\Sigma_{1+2}^{i} + (\mu_{1+2}^{i} - \hat{\mu}_{12}^{j})(\mu_{1+2}^{i} - \hat{\mu}_{12}^{j})^{T})}{\hat{w}_{12}^{j}} \tag{7.12}$$

   Avec $\pi^{-1}(j)$ une légère notation de $\hat{\pi}^{-1,it}(j)$, l'ensemble des projections de la composante j de S_{1+2} en S_{12}.

2. **Groupement des composantes :**

   pour le mélange de gaussiennes $\hat{S}_{12}^{it}$ obtenu dans l'étape 1, nous cherchons la fonction de transfert π^{it+1}, définie de $1, ..., m_1 + m_2$ jusqu'au $1, ..., m_{12}$, pour lequel le meilleur groupe de composantes de $\S_{12}^{it}$, dans le sens suivant :

$$\hat{\pi}^{it+1} = \arg\min_{\pi} KL_m(S_{1+2}, \hat{S}_{12}, \pi) \tag{7.13}$$

   Autrement dit, chaque composante i de S_{1+2} est la projetion de j par $\hat{S}_{12}^{it}$ au sens de la mesure conjointe de KL (voir équation 7.14) ci-dessous. Dans cette phase, nous avons recours à une recherche linéaire parmi les composantes, ayant un coût réduit grâce à l'approximation donnée par l'expression (7.7).

$$\pi^{it+1}(i) = \arg\min_{j} KL_m(N_{1+2}^{i} || N_{12}^{j}) \tag{7.14}$$

3. $it = it + 1$

   jusqu'à convergence (i.e $\pi^{it+1} = \pi^{it}$)

   calcul

7.3.2 Regroupement itératif

Nous proposons un schéma similaire à la procédure de k-means, comme technique alternative d'une classification hiérarchique, tel que les éléments à organiser sont représentés par des modèles de MMG de locuteurs. Cette technique est décrite dans l'algorithme 2. En outre, le critère d'optimisation de la relation père-fils est défini dans la section (7.2), sur l'ensemble des modèles pères :

$$\sum_{S_p \in \text{pères}} \sum_{S_c \in \text{fils de } S_p} KL_m(S_p || S_f) \tag{7.15}$$

Algorithme 2 :

Optimisation itérative des paramètres du modèle père.

Début : choix aléatoire des modèles de locuteurs ;

répéter

1. Réduire le mélange de chaque noeuds père à l'aide d'algorithme 1,

   Cette étape implique l'algorithme k-means agit sur les composantes des mélanges de gaussiennes. (Contrairement appliqué sur des mélanges de gaussiennes réalisés par l'algorithme 1),

2. Affectation de chaque noeud-fils à l'ensemble des noeuds-père les plus similaires (dans le sens de $KL_m(S_p||S_f)$).

jusqu'à convergence effectuée.

A l'aide de cette approche, l'affectation des modèles de locuteurs à un groupe en utilisant le même procédé de K-means peut être facilement exploré, contrairement à une sous-classification par groupement des MMG basé sur un dendrogramme de similarité. Par conséquente, l'approche itérative peut être aussi appliquée à un processus incrémental, c'est-à-dire une mise à jour des couches intermédiaires par une optimisation locale de l'equation (7.15) et, si nécessaire, il est possible d'étendre le présent schéma permettant de créer plus de couches intermédiaires selon l'évolution au cours de l'arrivée des nouveaux modèles.

7.3.3 Utilisation de l'approximation d'erreur dans la structure arborescente

Etant donné l'arbre obtenu à l'aide des approches présentées dans les deux sections précédentes. Notons par S_p le noeud père de $S_1, S_2, ...$

Les principaux points présentés dans la suite de ce travail se résument en : une proposition d'une technique pour créer le modèle père S_p à moindre coût, plus explicitement, nous essayons de donner une expression approximative pour l'évaluation de tous modèles fils à l'aide du critère log de vraisemblance des données du modèle requête à identifier. Pour cela, l'expression doit vérifier pour tout noeud fils $\log p(D|S_p) \approx \log p(D|S_{fils})$. D'où, la valeur de $\log p(D|S_p)$ peut être sauvegardée et réutilisée à chaque évaluation des noeuds fils de S_p. Il est alors nécessaire de définir une nouvelle approche pour un traitement plus rapide, notamment dans le cas d'un nombre important de modèles de locuteurs, et tant que la discrimination entre modèle est réalisée avec des vecteurs de vraisemblance.

Un point supplémentaire résulte de l'estimation d'erreur, sachant que le but est non seulement de minimiser cette erreur lors de la création de la couche intermédiaire, mais encore, l'estimation d'erreur doit être prise en compte pour raffiner l'exploration dans la phase de classification avec le même coût de calcul. Plutôt qu'un remplacement, pour tous noeud-fils k, $\log p(D|S_k)$ par $\log p(D|S_p)$, l'association de la vraisemblance à chaque noeuds fils peut être estimée comme suit :

$$\underbrace{\log \hat{p}(D|S_k)}_{\text{Log-vraisemblance fils}} \approx \underbrace{\log \hat{p}(D|S_p)}_{\text{Log-vraisemblance père}} + \underbrace{KL(S_p||S_k)}_{\text{indépendamment aux données à classifiées}} \quad \text{k=1,2 ...} \tag{7.16}$$

Un point important dans le calcul cette approximation de $KL(S_p||S_k)$ ne dépends aux données a classifier. En pratique, nous appuyons l'idée d'utiliser une nouvelle expression de KL basée sur une transformation non linéaire *Unscented Transformation* notée par KL_{UT} utilisée dans le calcul de $KL(S_p||S_k)$ [41], grâce à sa performance

par rapport à KL_m. Cette approximation entre mélanges de gaussiennes n'agit pas sur la forme globale d'une gaussienne, mais la résume en quelques points d'intérêts statistiques, ainsi le calcul de KL_{UT} à l'aide d'une représentation résumé sous forme des points sigma (les points de sigma données par la matrice de covariance du MMG) rend dans l'ensemble une technique facile et efficace en terme de calcul. En effet, l'approximation de la vraisemblance désormais possible à moindre coût pour chaque noeuds fils et offre plusieurs possibilités pour l'exploration de l'arbre de modèles, par exemple :

1. par une recherche linéaire dans un ensemble de noeuds fils, en utilisant l'approximation $\log \tilde{p}(D|S_k$, ou,
2. par un calcul *a priori* du maximum et du minimum d'erreur entre le noeud père est ces noeuds fils.

$$Min_{ERR} = \min_k KL(S_p||S_k), \tag{7.17}$$

la classe correspondante lors de la recherche est caractérisée et donnée par le maximum de probabilité, la valeur de log-vraisemblance doit être comprise entre les valeurs $[\log p(D|S_p) + Min_{ERR}, p(D|S_p) + Max_{ERR}]$ donnant lieu à plusieurs schéma et possibilités de recherche.

7.4 Résultats

Le test expérimental présenté dans cette section est effectué sur des données réelles extraites de flux audio en langue française. Les 13 coefficients de MFCC et leurs dérivés temporels sont alors utilisés. Une segmentation temporelle du flux audio en segments générés à l'aide du critère BIC, (une approximation classique [73] en test d'hypothèse Bayésien) appliqué sur des fenêtre de $(4sec)$. La représentation du modèle pour chaque locuteur est basé sur une adaptation Bayésien [9]. Le flux audio ne contient généralement que de la parole (par exp. : programme du journal d'information radio), ce dernier inclus de courte plage de publicité ce qui peut être facilement éliminé grâce aux propriétés acoustiques des descripteurs *MFCC*.

La première expérience est appliquée sur 20 locuteurs. La performance de la phase d'exploration est évaluée comme suit : 40 échantillons à partir de 20 locuteurs sont utilisés pour la classification par deux locuteur.
Premièrement, nous rapportons l'expérimentation à une recherche linéaire, tel que le modèle requête est testé avec tout les modèles de locuteurs déjà inscrits en utilisant KL_m ou la log-vraisemblance (voir tableau 7.1). Le maximum de vraisemblance est performant en terme de reconnaissance, tandis que, KL_m est moins efficace.

Recheche linéaire	Performance de la reconnaissance		
Durée de requête	5	10	15
ML	100%		
KL_m	75%	82.5%	85%

Table 7.1 – Comparaison de la performance de la recherche de modèle de locuteur dans le cas linéaire, en se basant sur le maximum de vraisemblance (score de ML) ou KL_m entre le modèle requête et l'ensemble de modèles déjà inscrits dans la base de données.

7.4.1 Résultats de groupement hiérarchique en dendrogramme

Deux critères alternatifs de similarité sont utilisé :

- la vraisemblance des données, nécessite l'utilisation des vecteurs descripteurs correspondants à un locuteur donné, au point de vu du coût de calcul ceci reste non favorable pour un tel traitement de données,
- la version symétrique de KL_m.

La structure arborescente est obtenue à l'aide d'un découpage de dendrogramme à un niveau de profondeur , ceci permet d'obtenir une partition de modèles MMG sous forme d'un arbre N-aire, (voir figure. 7.1). Par conséquent, l'arbre généré à l'aide de la mesure de similarité KL_m est nettement bien équilibré avec un minimum de coût de calcul que celui crée en utilisant le critère de la mesure de vraisemblance à partir des données.

L'exploration de l'arbre à partir de la racine aux feuilles. La comparaison du modèle requête est réalisée à l'aide d'un des deux méthodes proposées : (i) la vraisemblance des données requête avec un modèle donné, ou (ii) à l'aide de la mesure de similarité KL_m. Dans les deux cas, il faut mentionner que KL_m est toujours appelée dans la phase de la construction de l'arbre.

Dendogramme hiérarchique	ML		KL_m	
durée de la requête	5	10	5	10
26 gaussiennes	92.5%	95%	47.5%	40%
16 gaussiennes	95%	95%	50%	45%

Table 7.2 – La performance de reconnaissance dans le cas d'une organisation des modèles de locuteurs suivant un découpage du dendrogramme donnée par (voir figure. 7.1).

Les résultats sont présentés dans le tableau (7.2). Comme le montre les expériences précédentes, l'utilisation de KL_m de même pour ML, introduit une dégradation. Cependant, l'utilisation de ML, permet d'avoir une performance satisfaisante. Ainsi l'approche proposée est toujours favorable dans le sens : (i) l'exploration est effectuée à travers un arbre contrairement a une exploration linéaire, (ii) l'utilisation de KL_m dans la construction de l'arbre demeure plus rapide et efficace.

7.4.2 Résultats d'une implémentation hiérarchique en utilisant un groupement hiérarchique des MMG

Le tableau (7.3) illustre la performance de la reconnaissance obtenues sous les même conditions précédemment citées, en revanche, l'arbre de recherche est générée à l'aide d'un schéma itératif (Algorithme 2) contrairement à la technique basée sur le dendrogramme de mesure de similarité. La qualité donnée par le résultat est similaire à celle donnée avec l'approche précédente.

Regroupement hiérarchique itératif	ML		KL	
durée de la requête	5	10	5	10
26 gaussiennes	92%	92.5%	45%	55%
16 gaussiennes	92.5%	90%	57.5%	60%

Table 7.3 – La performance de reconnaissance dans le cas d'une organisation des modèles de locuteurs à l'aide de l'algorithme 2

7.5 Conclusion

Dans ce chapitre, nous sommes focalisés sur les technique de regroupement de modèles MMG. De plus, nous avons étudier des techniques de mesure l'information perdue due à ce choix d'organisation. Le système d'exploration d'une structure arborescente fiable doit tenir en compte une mesure d'erreur dans l'estimation du meilleur chemin d'exploration. Cette erreur est considérée comme une information de perte liée au choix de regroupement d'un sous ensemble des modèles d'une structure arborescente. En particulier, l'estimation de perte d'information liée au regroupement de modèle est réalisée sur une optimisation de mesure KL_m entre le noeud père et le nud fils.

Nous avons défini deux méthodes de pré analyse permettant d'améliorer l'organisation des modèles MMG (dendrogramme et un algorithme itératif de regroupement) pour lesquelles la notion de similarité joue un rôle principal. En terme de performance d'identification et de reconnaissance nous avons constaté une légère perte par rapport à une recherche classique linéaire. Notons que, la mesure de similarité donnée par KL_m offre des perspectives prometteuses en terme d'optimisation de temps de traitement. La procédure itérative de regroupement des modèles est très intéressante en pratique, la technique est mieux adaptée au processus incrémental de traitement de données.

En perspective, la technique de regroupement itérative peut être généralisée sur plusieurs niveaux. En plus, l'estimation de la mesure de KL_m entre le modèle père et le modèle fils, peut être enrichie avec plus d'informations en rapport avec la vraisemblance des données d'un modèle fils donné, à présent que l'information donnée par le maximum de vraisemblance par rapport à un modèle père donné, est pris en considération.

Dans le chapitre suivant nous allons continuer notre investigation des techniques d'organisation des MMG locuteurs. Notre approche est basée sur une stratégie qui permet de mener chaque modèles MMG de locuteur récemment détecté par une information *a priori* obtenue à partir des modèles qui partagent une forte similarité et selon l'ordre historique de leurs détection. Pour cela, une structure de Treillis sur des régions temporelles de traitement.

CHAPITRE 8

Structure de Treillis définie sur des régions temporelles des segments des locuteurs

Deux contributions majeures seront présentées dans ce chapitre, d'un coté ce travail consiste à améliorer et adapter la divergence de Kullback-Leibler. L'approximation de la divergence de KL permet de fournir une expression robuste et discriminante entre modèle MMG de locuteur. D'un autre coté, nous proposons une structure arborescente adaptée au problème incrémental à l'aide de treillis. Le test expérimental est effectué sur le flux de modèles MMG bien défini selon un scénario aléatoire. Des stratégies d'amélioration ont été menées sur ces algorithmes afin de diminuer la complexité de manipulation et d'exploration permettant d'augmenter la performance de la reconnaissance et le coût de recherche en termes de temps de requête.

8.1 Contexte et motivations

La technique d'identification de locuteur utilise un critère de décision basé sur des mesures de distance entre modèles de chaque unité de test. Les descripteurs de chaque locuteur sont représentés par un MMG. Ces modèles statistiques sont plus adaptés à l'invariabilité des données source. Le but alors est de créer une structure hiérarchique de modèles GMM de chaque locuteur. Selon l'état de l'art, plusieurs techniques de segmentation des données acoustiques utilisent aussi une distance de mesure pour détecter le changement de locuteur [45]. Notre approche d'indexation audio par locuteur consiste à créer et améliorer une structure hiérarchique de modèles MMG de chaque locuteur détecté. L'étude des structures comme les treillis consistent à l'adapter à l'indexation au sens

de locuteur pour des flux audio, pour cela une forme modifiée de transformation non linéaire de Kullback-Leibler appelée *Unscented Transformation* sera implémentée (KL_m). En effet, le regroupement des modèles MMG suivant une structure arborescente exploite le fait de pouvoir grouper tous les GMM's selon un critère de similarité entre les modèles de locuteurs [81, 50]. Ainsi, la structure arborescente de treillis permet de représenter les liens de similarité entre les GMM locuteurs sans avoir besoin de regrouper les données liées à leurs modèles. Dans ce travail, plusieurs facteurs justifient la motivation de cette approche proposée pour concevoir de nouvelles stratégies des méthodes d'organisation de locuteur MMG. Nous allons citer les arguments pour lesquels nous avons optés pour le regroupement hiérarchique à l'aide de treillis et puis l'adapter au problème incrémental :

- l'utilisation d'un arbre binaire pour le regroupement des MMG engendre des pertes d'information par rapport à l'ensemble global de la structure (contrairement à un test linéaire complet), qui a pour conséquence des erreurs de reconnaissance sur chaque niveau de regroupement,
- la construction d'un arbre binaire ascendant est valide si le nombre de modèles des locuteurs est fini. Cependant, la version incrémentale est difficile à implémenter sachant qu'il faudra réorganiser l'ensemble de la structure,
- la classification par regroupement de modèles dans une seule classe rend la mise à jour selon la variabilité de locuteurs très difficile.

Les trois raisons précédemment citées, justifient notre motivation dans le but de créer une nouvelle technique d'organisation des modèles MMG des locuteurs basée premièrement sur une nouvelle approximation de KL_m [31], deuxièmement une organisation des MMG à l'aide d'une structure qui permet une exploration dans le sens ascendant que descendant. La structure choisie permet une connectivité libre selon l'évolution des modèles entrants et leur variabilité au cours du temps.

La prochaine section donne un aperçu des techniques de mesure de distance entre les modèles de mélange de gaussiennes. Dans la section 7.3 un algorithme d'organisation des modèles de locuteurs sera présenté. Ce dernier utilise la fréquence d'apparition des modèles similaires et les représente à l'aide d'un treillis. Dans la section 7.4 nous évaluons les méthodes suggérées par des données réelles de 20 locuteurs faisant plusieurs apparitions selon une loi aléatoire.

8.2 Identification de locuteur Indépendamment du texte

Le traitement efficace des systèmes d'indexation au sens de locuteurs est basé sur des stratégies de structure de données. Le but est de faciliter leurs accès, leur manipulation et leur exploration dans des grandes bases de données

multimédia. Cette section donne un aperçu sur le processus d'indexation illustré par (figure. 8.1). Après une phase d'acquisition du signal audio ne contenant que de la parole, le module de segmentation non supervisé génère des segments de descripteurs MFCC. Pour chaque détection de changement de locuteur un segment de descripteur d'un minimum $(5sec)$ est nécessaire pour la phase d'apprentissage. L'étape suivante consiste à représenter chaque bloc de données en un modèle statistique. Ce module exploite pleinement la performance de modélisation des descripteur pertinents vis-à-vis de locuteur à l'aide d'un modèle de mélange de gaussiennes fortement utilisé dans le domaine de la RAL. Ensuite, l'identification automatique du locuteur consiste à déterminer, parmi une population de N locuteurs qui existent dans la base de données des MMG, celui ou ceux qui ont la mesure de similarité très faible. Le résultat d'identification peut alors être exprimé par une liste de modèles les plus probables.

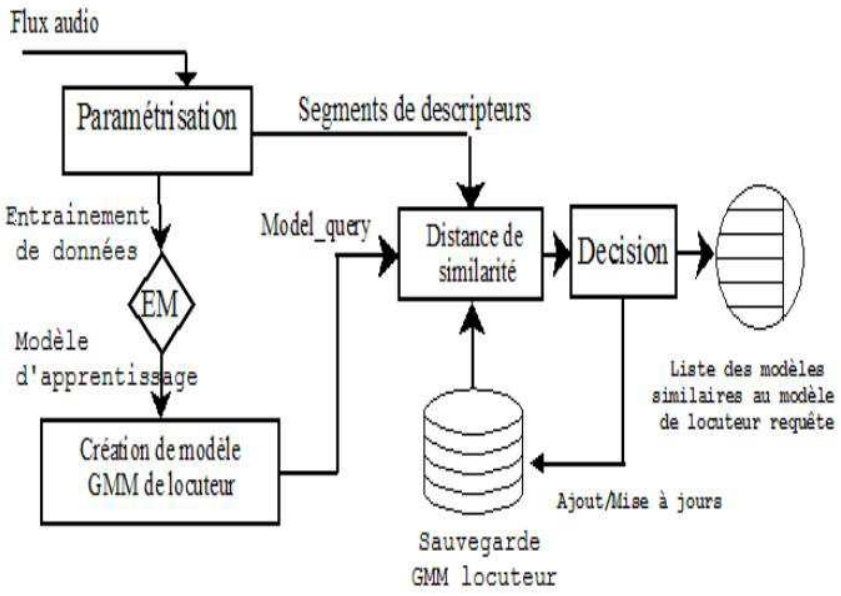


Figure 8.1 – Diagramme d'un système d'identification de locuteur

8.2.1 Mesure de similarité entre modèles MMG

8.2.1.1 Approximation robuste de la divergence de Kullback-Leibler

La mesure de similitude entre deux MMG générés après une phase d'apprentissage des segments de descripteurs MFCC, représente l'outil principal utilisé dans notre module d'identification. En outre, l'expression de KL_m n'utilise que les paramètres du modèle sans avoir besoin des données d'entraînement. Cette mesure permet non

seulement de déterminer l'identité d'un tel modèle, mais aussi elle joue un rôle très important dans l'organisation des modèles de locuteurs suivant le critère de similarité. Notons que la divergence de KL entre deux modèles du mélange de gaussiennes $f = \sum_{i=1}^{N} \alpha_i f_i$, $g = \sum_{j=1}^{M} \beta_i g_j$ est définie comme suit :

$$KL_m(f||g) = \int f \log(\frac{f}{g}) \approx \frac{1}{n} \sum_{t=1}^{n} log(\frac{f(x_t)}{g(x_t)}) \tag{8.1}$$

Notre contribution consiste à améliorer la mesure de similitude en proposant une nouvelle approximation de la divergence de KL à l'aide d'une transformation non linéaire *Unscented Transforamtion* modifiée notée (KL_{ut-mod}) à partir de l'expression (KL_{ut}) décrite dans les travaux de [32] (voir équation 8.2). En outre, l'approximation de *Monte-Carlo* est utilisée pour exprimer une nouvelle expression de la divergence de KL en KL_{ut}, cette nouvelle formule consomme moins de points sigma ce qui la rend moins coûteuse et rapide dans le cas de traitement d'un grand nombre de modèle de mélange de gaussiennes. Cette expression modifiée de KL_{UT-mod} est présentée par la formule (voir l'équation 8.3), permettant d'améliorer la performance en terme d'identification ainsi que le coût d'exécution de la moitié par rapport à la version KL_{ut}. Le nombre des "points de sigma" (les points sigma sont donnés par les coordonnées de la matrice de variance d'une distribution gaussienne) ce qui réduit considérablement la complexité de la mesure (voir la figure. 8.2). En outre, pour les composantes gaussiennes simples de chaque MMG locuteur, l'expression ci-dessus sera écrite comme suit :

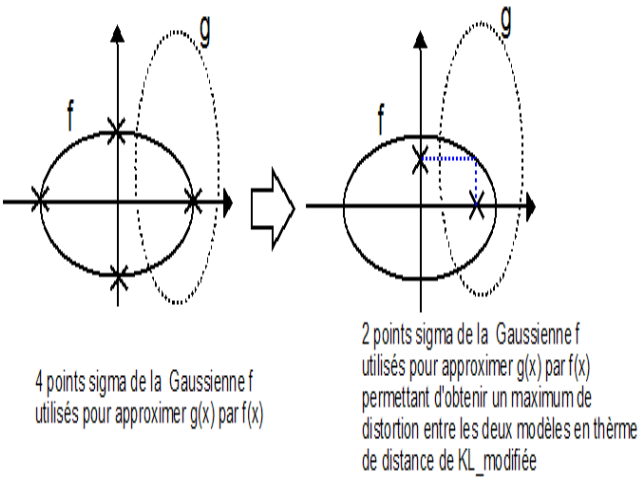


Figure 8.2 – Approximation par transformation non linéaire de la divergence de KL utilisée pour la mesure de similarité entre MMG et pour l'identification de locuteur

Rappelons que, la transformation non linéaire *Unscented Transformation* de KL est définie comme suit : supposez que x est une variable aléatoire $x \sim f_i$ puis $E_{f_i}(x) = \mu_i$ et $E_{f_i}(log(g(x))$ est l'espérance de la fonction

non linéaire $\log g(x)$ qui peut être employée et approximée par une transformation non-linéaire, par conséquent :

$$\int f \log g = \frac{1}{2d} \sum_{i=1}^{n} \alpha_i \sum_{k=1}^{2d} log(g(x_k)) \tag{8.2}$$

tel que :

$$x_{i,k} = \mu_{i,k} + (\sqrt{d\Sigma_i})_k \text{ avec } k = 1, ..., d$$

$$x_{i,k+d} = \mu_{i,k} - (\sqrt{d\Sigma_i})_k \text{ avec } k = 1, ..., d$$

Une nouvelle approximation est proposée. Celle-ci est basée sur la même technique de la transformation de variable et permet ainsi d'améliorer la performance de discrimination à l'aide de la mesure de similarité et de réduire considérablement le coût d'exécution. La nouvelle expression est définie comme suit :

$$KL(f||g) = \int f \log f - \int f \log g$$

avec :

$$\int f_i \log g_j = \frac{1}{d} \sum_{i=1}^{N} \alpha_i \sum_{k=1}^{d} log(\sum_{j=1}^{M} \beta_j g_{j,k}(X_{i,k})) \tag{8.3}$$

tel que :

$$X_{i,k} = \mu_{f_{i,k}} + \rho_{i,k}(\sqrt{d\Sigma_{f_i}})_k \text{ avec } k = 1, ..., d$$

d'où :

$$\rho_{i,k} = \frac{(\mu_{g_{j,k}} - \mu_{f_{i,k}})\sigma_{f_{i,k}}}{\sigma^2_{f_{i,k}} + \sigma^2_{g_{j,k}}} \tag{8.4}$$

En outre, la démonstration de paramètre $\rho_{i,k}$ (voir 8.4) est présentée explicitement dans l'annexe A.

8.3 Structure incrémentale des modèles de locuteurs MMG

Dans cette section, nous adoptons un certain algorithme de structure de données dans le but d'organiser les MMG des locuteurs en utilisant le critère de similarité. Ainsi, l'approximation de KL présente la clef de cette structure incrémentale proposée.

8.3.1 Objectifs

Le développement croissant du volume de données multimédias tels que le *Streaming*, *Podcasting*, archives ou transmission radio rend leur gestion et leurs manipulation et navigation par contenu une tâche très complexe. Notre objectif est de réduire la complexité d'un système d'indexation audio au sens de locuteur lors du passage à l'échelle. La représentation et l'organisation hiérarchique des modèles statistiques pour un SGBD audio sont les importantes préoccupations de notre travail. Des techniques de représentation arborescente ont été expérimentées dans les chapitres précédents ont permis de réduire considérablement le coût de recherche. En revanche, la perte de performance d'identification est fortement liée au choix de regroupement en classe des modèles de locuteurs par rapport à un critère de similarité. Cette information de perte peut être contournée en utilisant une nouvelle structure qui permet une libre exploration soit par région temporelle (derniers modèles détectés) et/ou par critère de similarité. En 1984, Antoin Guttman a présenté une structure nommée R-Arbre (arbre par région), Guttman donne les algorithmes afin de naviguer, mettre à jour et manipuler, etc.[36]. On trouve aussi plusieurs travaux qui s'intéressent à l'étude des treillis et des raccordements rigides qui peuvent être employés dans le travail ci dessous [44].

Après une courte présentation des méthodes fortement employées comme algorithmes et structures de données, nous présenterons l'algorithme principal à l'aide des techniques existantes adaptées au problème incrémental.

8.3.2 Proposition : Régions temporelles définies sous une structure de treillis

La structure proposée est créée suivant une procédure différente de celle d'une structure binaire. Rappelons que notre système d'indexation doit aussi admettre des opérations comme la mise à jour au fur et à mesure que les données arrivent. Dans l'état de l'art aucune étude ne permet de comparer la performance d'un arbre binaire par rapport à un R-Arbre. Un avantage majeur des treillis et le fait que ce n'est pas nécessaire de regrouper les noeuds (modèle locuteur MMG). La nouvelle structure se contente seulement d'une simple opération d'insertion.

Les connexions par ordre de priorité de la structure arborescente proposée sont réalisées comme suit : Pour un flux audio entrant la détection de changement de locuteur génère des segments homogènes de vecteurs MFCC puis sauvegardés. Au fur et à mesure que les premiers segments de descripteurs arrivent Puis, les segments des descripteurs MFCC seront sauvegardés ainsi permettent la création d'un modèle MMG appelé *Universel Backgroud Model* (UBM). Ce dernier est utilisé pour atteindre une performance élevée en phase d'entraînement. La deuxième étape consiste à comparer chaque modèle MMG de locuteur qui vient être détecté avec l'ensemble des modèles existants précédemment détectés et identifiés.

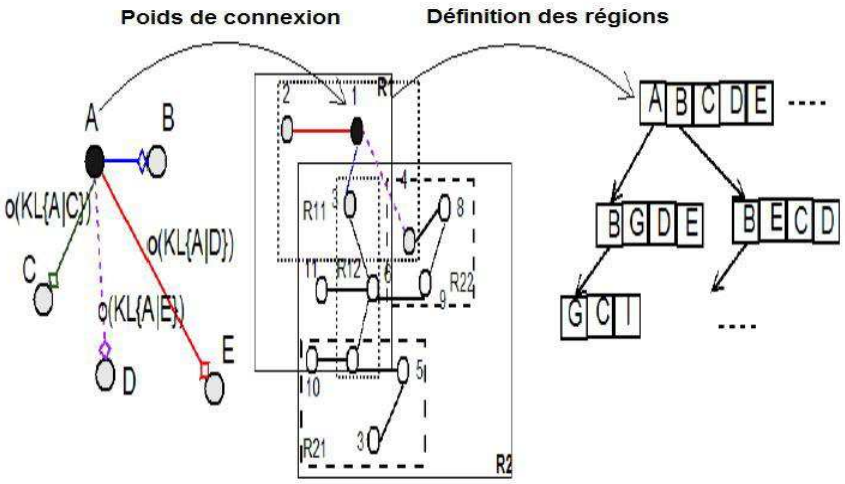


Figure 8.3 – Régions temporelles définies sous une structure de treillis utilisant des connections par ordre de priorité

L'exploration de la structure est réalisée selon un test exhaustif sur une liste des L derniers modèles de locuteurs précédemment ajoutés dans la base de données des MMG. La liste représente une région temporelle notée R contenant les derniers événements au cours du temps (par exemple pour un modèle récemment détecté loc la région temporelle R_{loc} comporte $L = 5$ locuteurs triés par ordre de similarité à l'aide des mesures conjointes de (KL_{ut}). L'identification de locuteurs par définition permet d'avoir une agrégation de modèles comme résultat du test. Chaque modèle détecté sera muni d'une liste de référence à des modèles précédemment détectés dont la distance par rapport KL_{ut} est très faible. Cependant, après N détections le modèle $(N + 1)$ aura la possibilité d'explorer la liste de ces derniers h modèles de locuteurs classés dans l'ordre de similarité. Cette technique permet cependant d'optimiser la recherche du locuteur correspondant et dans le cas échéant l'ajout d'une nouvelle entrée sera effectuée. (voir le Fig. 8.3).

Notons par $TR_W(S_k)$ la région temporelle du locuteur k, et S_{t_n} l'ensemble des locuteurs insérés dans la base de données à l'instant t_n, nous définissons comme suit :

$$TR_W(S_k) = \text{Sort}_{min}\{\text{KL}_{ut}(S_{query}|S_i), |i = card(S_t) - W, ..., card(S_t)\} \tag{8.5}$$

W correspond à l'intervalle W^{th} des derniers modèles MMG de locuteurs identifiés dans la base de données. La taille de W augmente par rapport au cardinal de S.

Notons maintenant par ϕ la fonction des connexions définie entre chaque modèle de locuteur appartenant à l'ensemble des locuteurs S :

$$\phi : S \rightarrow S^L$$

$$S_i \longmapsto \{S_1, ..., S_L\} = T_L(S_i)$$

Tel que , L est la taille de la liste des modèles similaires.

Une étude de performance de la technique proposée sera présentée dans la prochaine section. Le test est effectué sur un flux incrémental en utilisant un sénario aléatoire composé d'un ensemble de N locuteurs. Cependant, la simulation de l'arrivée des modèle des locuteurs est aléatoire et suit une distribution uniforme.

8.4 Résultats expérimentaux

La première étape du processus d'indexation, consiste à effectuer une segmentation du flux audio selon les apparitions des locuteurs à l'aide du test d'hypothèse bayésiennes. La deuxième étape consiste à créer et générer des modèles MMG à l'aide d'algorithme *EM*. Cependant, pour chaque bloc de descripteurs MFCC un modèle MMG composé de 16 densités gaussiennes de dimension 26 sera généré.

MMG	*KL modified*			KL_{ut}			KL_{ut-mod}		
TD	5	10	15	5	10	15	5	10	15
SG 1	2.00	**2.04**	1.39	0.43	0.44	0.44	**0.37**	0.38	0.38
SG 2	**1.99**	1.25	0.56	0.56	0.48	0.48	0.51	**0.39**	0.42
SG 3	0.35	**1.34**	0.85	0.43	0.36	0.41	0.35	**0.34**	0.37

Table 8.1 – Les mesures comparatives de distance des trois techniques effectuées de même modèle MMG de locuteur utilisant différents segments de données notant par TD = taille de données d'entraînement en *(sec)* et SG= MFCC segment de descripteurs MFCC

La mesure de similarité modifiée KL_{UT} est basée sur l'approximation non linéaire. Sa performance est prouvée non seulement par sa fiabilité en matière d'identification mais également par le coût de calcul satisfaisant ce qui est facile à démonter par une simple comparaison de complexité entre les deux formules de KL_{UT} et KL_{UT-mod}.

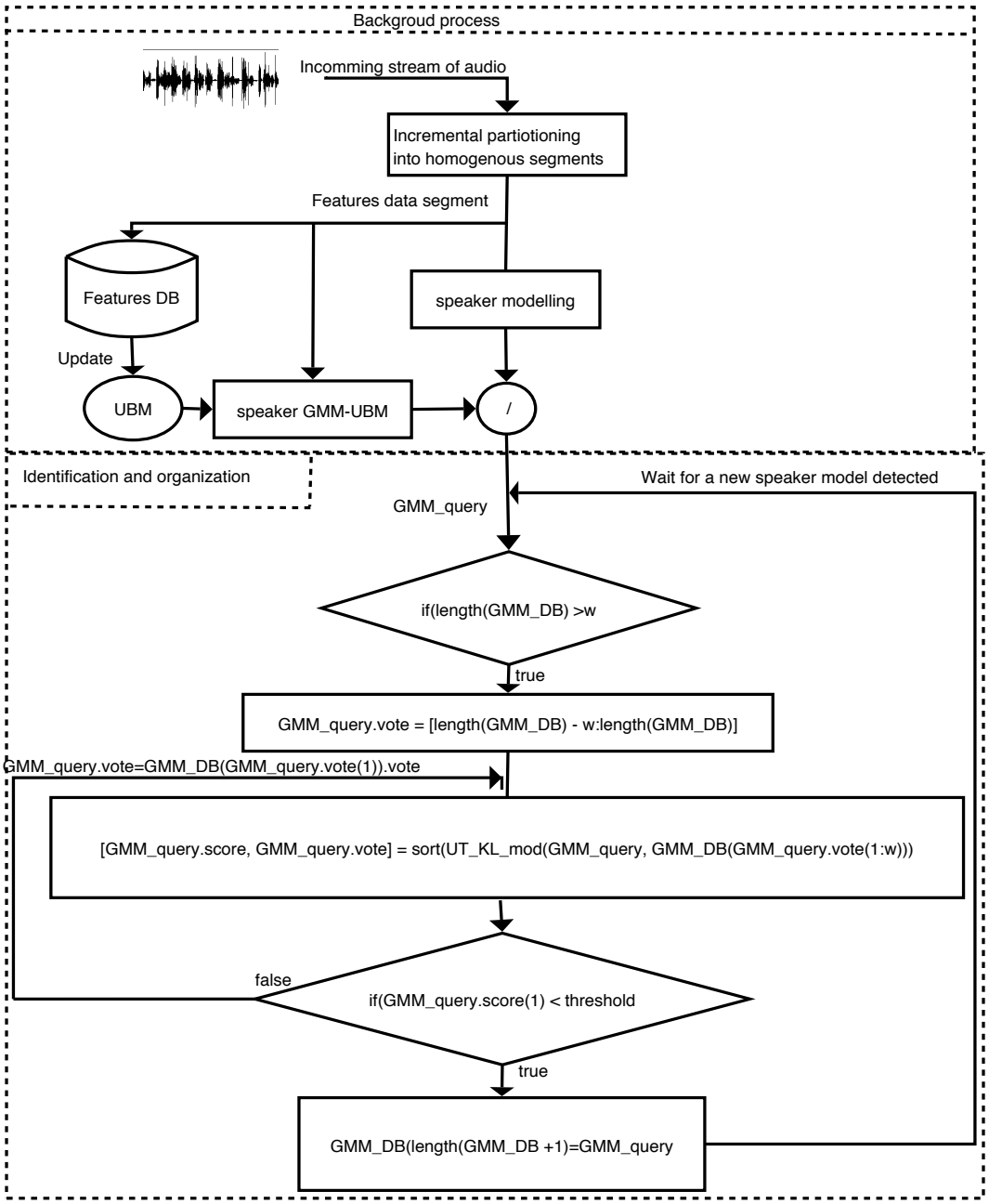


Figure 8.4 – Algorithme de création de la structure arborescente permettant l'indexation incrémentale au sens de locuteur

Cependant, la mesure de similarité entre deux MMG "différents" générée par différents segments de données correspondant au même locuteur à l'aide de KL_{UT-mod} est moins sensible aux effets de la taille de données utilisées dans la phase d'entraînement des modèles MMG de locuteur (voir Tab. 8.1).

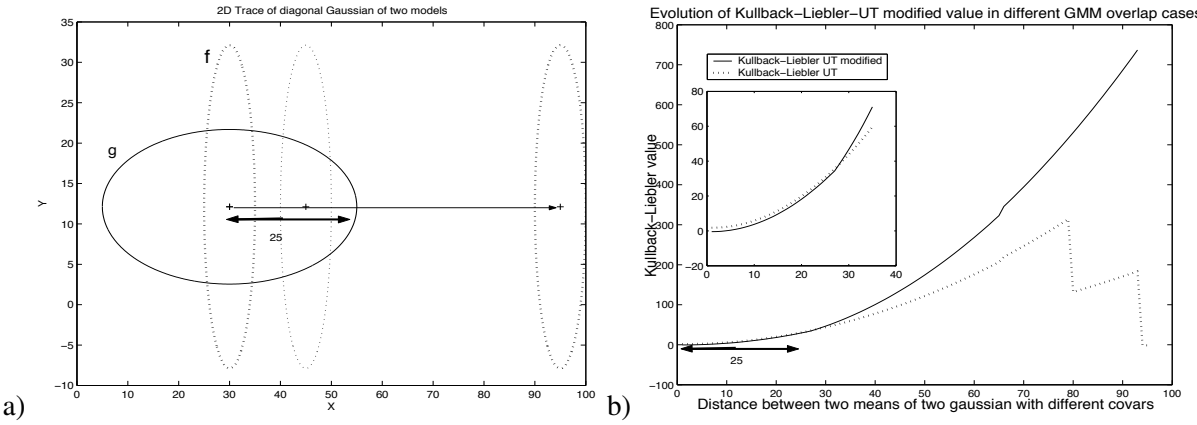


Figure 8.5 – a) L'évolution de la mesure KL_{ut} et KL_{ut-mod} entre deux composantes gaussiennes f et g au cours d'une translation horizontale de f par rapport à l'axe des X

Dans la deuxième étape, l'étude de performance de la structure proposée est réalisée à l'aide d'une simulation d'arrivée de modèles de locuteurs crées à partir d'un nombre limité de MMG. Ainsi, 20 modèles de locuteurs vont se permuter en créant par suite une simulation de changement de rôle de parole dans un document audio, le choix d'un modèle entrant est réalisé avec une fonction aléatoire et uniforme.

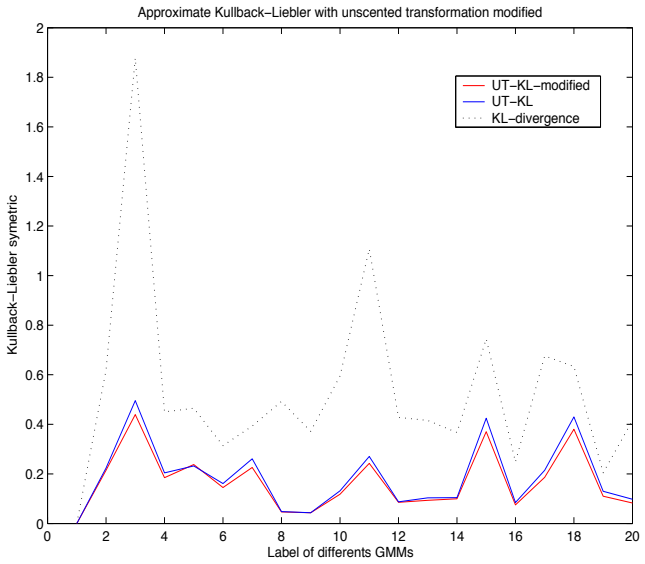


Figure 8.6 – Courbe de performance des différentes approximations de la divergence de KL, les différents mélanges de gaussiennes sont testés avec des données aléatoires

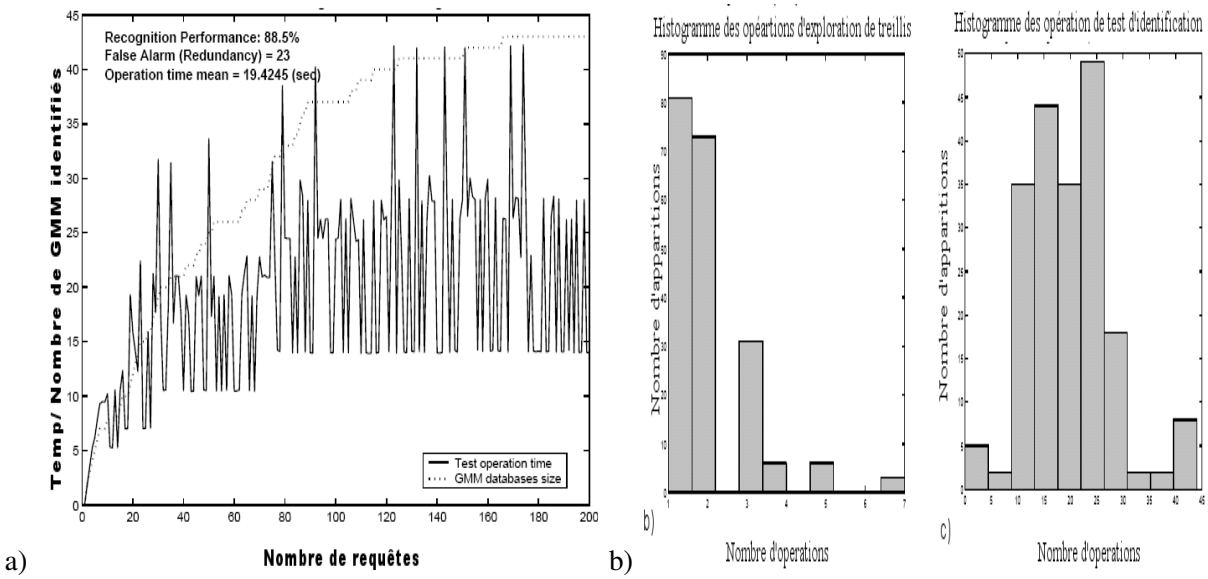


Figure 8.7 – a) L'évolution de temps de la requête d'identification en fonction du nombre de locuteurs déjà inscrits, entre 200 événements construits à partir de 20 locuteurs, 43 modèles de MMG sont insérés avec une redondance de 23, en revanche le temps d'exploration est nettement amélioré à 88.5%. b) présente l'histogramme du nombre des opérations nécessaires en utilisant la variable W par rapport à la taille de la base de données MMG. c) représente le nombre total des tests d'identification utilisés pour chaque opération de redirection

8.5 Conclusion

Après avoir donné un aperçu des techniques de segmentation et de reconnaissance automatique de locuteur ce chapitre a présenté deux contributions principales : d'un coté, dans le but d'améliorer et d'adapter la divergence de KL en une mesure plus discriminante pour l'identification et la reconnaissance et d'un autre coté, pour construire une structure de modèles MMG incrémentale à l'aide des treillis qui permet de compenser aux pertes d'informations liée à une structure basé sur un regroupement ou répartition arborescente binaire ou n'aire.

L'approximation de la divergence de KL a permis de fournir une expression robuste et discriminante entre modèle MMG de locuteur. L'expression de la divergence de KL donnée par l'approximation de *Monte Carlo* a permis de développer une nouvelle expression à l'aide d'une transformation non-linéaire appelée *Unscented Transformation* noté KL_{ut}, cette nouvelle formule de KL utilise des points de sigma égale à (2 x dim(g_i)) deux fois la dimension d'une composante gaussienne d'un MMG. Notre contribution, a permis non seulement de réduire le nombre de points de sigma, mais aussi d'augmenter la performance de cette mesure par critère de similarité (voir

tab. 8.1).

Le rôle de la structure de Treillis est très important dans la suite de ce travail, car il permet une simple implémentation des modèles et une représentation sous forme d'une architecture qui permet de réduire la zone de recherche par vue temporelle et de similarité.

Les résultats expérimentaux sont effectués sur un flux de modèles MMG déjà inscrits arrivant selon un scénario aléatoire. La complexité en terme de temps de manipulation et d'exploration est nettement améliorée avec une performance d'identification qui voisine les 88.5%.

Conclusion

Conclusion générale

Ce travail de thèse entre dans le cadre d'une collaboration où une activité a déjà été initiée dans l'équipe partenaire et où une couche logicielle offrant des services de base pour la gestion de données audio a été élaborée. Cette tâche entre dans le cadre des objectifs définis par le plan stratégique INRIA 2003-2007 et par le pôle de compétitivité Images Réseaux, auquel le laboratoire LINA est rattaché (par le biais de la FR Atlanstic).

Les technologies de navigation et de recherche d'information connaissent une évolution rapide afin de répondre aux besoins diversifiés des applications multimédia. A cet effet, un fort appel a été lancé par des spécialistes pour guider les utilisateurs à une meilleure navigation. En effet, l'analyse du contenu enrichie des méta-données au delà de ce qu'on peut faire à la main, ainsi, la recherche d'information peut évoquer des ressources tels que (archives des stations radio, télévision, podcast, etc..) ; en particulier, le flux en continu des données audios vidéo.

De nombreuses applications de reconnaissance automatique utilisent souvent une technique d'évaluation très coûteuse de données grâce au calcul de la maximum de vraisemblance, notamment dans le cas d'un grand nombre de candidats de modèles statistique, ces derniers sont utilisés afin de vérifier l'appartenance des données de test (tâche de classification). Dans un cas plus particulier, nous traitons ce problème de la reconnaissance automatique de locuteur dans le cas de la recherche par contenu via un système d'indexation audio. Plus précisément, nous proposons de réduire la complexité de temps de requête, par une organisation hiérarchique des modèles de locuteur a priori. Ceci est très classique utilisant des vecteurs multidimensionnels, en revanche, nous proposons une nouvelle technique de construction d'une structure hiérarchique des modèles dont la forme est représentée par des mélanges de gaussiennes.

Ce travail inscrit dans le cadre d'une thèse qui a permis de proposer une méthode non-supervisée d'indexation de locuteur combinant des techniques de reconnaissance automatique de locuteur dans le cas de large volume de données audio ou flux continu ne contenant que de la parole. L'objectif est de proposer une technique d'organisation hiérarchique des modèles de locuteurs dont le but est : de construire une structure d'index qui facilite la navigation et la mise à jour de bases de données et de pouvoir adapter la structure au problème incrémental.

Les différentes contributions proposées dans ce travail de thèses ont pour objectif consiste à réduire le temps de requête et la complexité de recherche du modèle de locuteur :

1. Première contribution permet le regroupement ascendant de modèles de mélanges de gaussiennes en structure n-aire. La construction de la structure est basée sur des techniques de mesure de similarité (i.e la divergence de Kullback-Leibler (KL), ou encore la vraisemblance des données).
2. Développement d'un algorithme de fusion de modèles de mélanges de gaussiennes pour optimiser le coût de construction d'un arbre de recherche de modèle de locuteurs par lot de modèles de mélanges de gaussiennes via un arbre binaire de recherche ascendant [69].
3. Création d'un algorithme incrémental d'organisation des modèles de mélanges de gaussiennes de locuteurs basé sur la mesure de divergence de KL proposant un mécanisme de mise à jour des modèles de mélanges de gaussiennes [70].
4. Une approche originale et efficace d'organisation hiérarchique des modèles de mélanges de gaussiennes de locuteurs a priori basée sur des techniques de classification (dendogram-based ou K-mean like), à partir d'une expression modifiée de la divergence de KL de noeud parent et fils des critères d'optimisation sont produits. Le schéma proposé est évalué sur des données réelles [67].
5. Dans un travail récent, nous avons proposé une nouvelle approche permet de définir des régions temporelles de similarité à l'aide de structure en treillis. Deux majeures contributions sont réalisées :
   - une transformation non-lineaire de l'approximation de Monte-Carlo de la divergence de KL avec réduction de nombre de point sigma (i.e. 'point sigma' est représenté par le vecteur de la covariance). Des composantes gaussiennes des deux modèles de mélanges de gaussiennes à comparer donne lieu à une nouvelle approximation de la divergence de KL robuste et discriminante entre MMG de locuteurs.
   - un algorithme de création d'une structure arborescente adaptée au problème incrémental à l'aide de treillis. La structure permettant de définir des régions temporelles d'historique d'apparissions des MMG des locuteurs précédemment inscrits partageant une mesure de similarité très proche entre le modèle en cours de traitement [68].

De nombreux autre aspect de l'indexation de documents multimédia font l'objet de recherche parallèles (image, vidéo, audio). Comme perspectives à notre travail : d'abord nous intéressons à définir des méthodes pour évaluer la performance des approches hiérarchiques de classification et d'avoir des techniques d'identifier le nombre de noeuds dans les couches intermédiaires des arbres. En suite, nous avons besoin de maîtriser la perte de probabilité dans le cas d'une organisation hiérarchique pour une bonne classification. La question qui reste ouverte est à quel

point peut-on tolérer en performance de reconnaissance en profit d'une recherche rapide à l'aide des algorithmes d'organisation de modèles MMG en structure hiérarchique ?

Bibliographie

[1] R. ANDRÉ-OBRECHT.
Segmentation and parole.
Document d'habilitation, 1993.

[2] S.S ANILY et A. FEDERGRUEN.
Simulated annealing mandhods with general acceptation probability.
Journal of Applied Probabilities, 24:657–667, 1968.

[3] R. AZENCOTT.
Simulated annealing.
Bourbaki, 697:1–15, 1988.

[4] M. BASSEVILLE et I. V. NIKIFOROV.
Dandection of abrupt changes: theory and application.
PTR Prentice-Hall, 1993.

[5] M. BASSEVILLE et IV. NIKIFOROV.
Dandection of abrupt changes: theory and application.
1993.

[6] H. BEIGI et S. MAES.
Speaker channel and environment change dandection.
Dans *World congress of automation*, 1988.

[7] S BERRANI, L. AMSALAG et P. GROS.
Robust content-based images searches for copyright protection.
Dans *ACM Workshop on Multimedia Databases*, pages 70–77, New Orlean, USA, November 2003.

[8] P. BERTRAND.
A local mandhod for estimating change points: the 'halt-fonction".
Statistics, 34:215–235, 2000.

[9] F. BIMBOT, J.F. BONASTRE, C. FERDOUILLE, G. GRAVIER, I. MARGIN-CHAGNOLLEAU, S. MEIGNIER, T. MERLIN, J. ORTEGA-GARCIA, Pandrovska-Delacrandaz D. et D.A REYNOLDS.
tutorial on text-independent speaker verification.
Dans *EURASIP J. Appl. Signal Process*, volume 4, pages 530–451, 2004.

[10] C. BISHOP.
Neural Nandworks for Pattern Recogntion.
Oxford University, 1995.

[11] R. BOITE et M. KUNT.
Traitement de la parole.
Presses Polytechniques Romandes,, 1987.

[12] A.V. BRANDT.
Dandection and estimating paramander jumps using algorithms and likelohood ratio tests.
Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'83)*, 1983.

[13] B. E BRODSKY et B. S DARKHOVSKY.

Nonparamandric mandhods in change-point problems.
Mathematics and its Applications. Kluwer Academic Publishers, 1993.

[14] K. BURNHAM et P.D ANDADERSON.
Model selection and multi-model inference.
Springer-Verglas New York, 2002.

[15] CALLIOPE.
La parole and son traitement automatique.
Paris, France,, 1989.

[16] J.P. CAMPBELL, D. A. REYNOLD et R. B. DUMN.
Fusing high- and low level features for speaker recognition.
Dans *EUROSPEECH*, 2003.

[17] W. M CAMPBELL, Reynold D. A. et Campbell J. P.
Fusing discriminative and generative mandhods for speaker recognition: Experiments on switchboard and nfi/tno field data.
Dans *Odyssey 2004, The sSpeaker and Lanhuage Recognition Workshop*, 2004.

[18] W. M CAMPBELL, Assaleh K. T et Broun C. C.
Speaker recognition with polynomial classifiers.
IEEE Transactions on Speech and Audio Processing, 10:205–212, 2002.

[19] B. P CARLIN, A. E. GELFAND et A. F. M. SMIYH.
Hierarchical bayesian analysis of change-point problems.
Applied Statistics, 41:389–405, 1992.

[20] D. CHEN et P. GOPALAKRISHNAN.
Speaker, environment and channel change dandection and clustering via the bayesian information criterion.
Dans *DARPA speech workshop*, 1988.

[21] M. CHEN, M. SHAO et J. IBRAHIM.
Monte Carlo Mandhods in Bayesian Computation.
Springer, 2005.

[22] S. CHEN, J.F. GALES, P. GOPALAKRISHNAN, R. GOPINATH, H. PRINTZAND, D. Kanevsky P. OLSEN et L. POLYMENAKOS.
System for transcription of broadcast news in the 1997 hub4 english evaluation.
Dans *Speech recognition workshop (DARPA '98)*, pages 389–404, 1998.

[23] M. CSÖRGO et L. HORVATH.
Limit theorems in change-point analysis.
Chichester, 1997.

[24] P. DELACOURT.
La segmentation et le regroupement par locuteurs pour lindexation de documents audio.
Thèse de doctorat, Institut Eurécom, Sophia Antipolis, *Septembre* 2000.

[25] P. DELACOURT et D. KRIZE.
Speaker based segmentation for audio data indexing.
Dans *ESCA woekshop: eccessing information in audio data*, 1999.

[26] B. DELYON, R. ANDRÉ-OBRECHT et H. Y. SU.
Expériences en vue de décodage acoustico phonétique à partir d'une recherche d'événements acoustiques and d'un acoustiques and d'un codage vectoriel.
journal of acoustique, 1:220–201, 1988.

[27] J. DESHAYES et D. PICARD.
Ruptures de modèles en statistique.
Thèse de Doctorat, Université Paris-Sud, 1983.

[28] C. FERDOUILLE.
Approche Statistique pour la Reconnaissance Automatique du Locuteur: Information Dynamiques and Normalisation Bayesienne des Vraisemnlance.
Thèse de Doctorat, Université d'avignion, Octobre 2000.

[29] T. GANCHEVAND, D. K TASOULIS, M. N. VRAHATIS et N. FAKTOAKIS.
Locally recurrent probabilistic neural nandwork for text-independant speaker verification.
Dans *EUROSPEECH*, 2003.

[30] J. L. GAUVAIN, LAMEL, L. et G. ADDA.
Audio partitionning and transcription for broadcast data indexation.
Dans *Proceedings of the first european workshop on Centent-based Multimedia Indexing CBLI'99*, pages 67–73, 1999.

[31] J. GOLDBERGER et H. ARONOWITZ.
A distance measure bandween GMMs based on the unscented transformand its application to speaker recognition.
Dans *INTERSPEECH-2005*, pages 1985–1988, 2005.

[32] J. GOLDBERGER et H. ARONOWITZ.
A distance measure bandween GMMs based on the unscented transformand its application to speaker recognition.
Dans *INTERSPEECH*, pages 1985–1988, Lisbon, Portugal, September 4-8, 2005.

[33] J. GOLDBERGER et S. ROWEIS.
Hierarchical clustering of mixture model.
Dans *Neural Information Processing Systems 17 (NIPS '04)*, pages 505–512, 2004.

[34] A. D. GORDON.
Classification 2nd Edition.
Chapman & Hall, 1999.

[35] M. A GRISHICK et H. RUBIN.
A bayes approach to quality control model.
Annals of Mathematical Statistics, 23:114–125, 1952.

[36] A. GUTTMAN.
R-trees: A dynamic index structure for spatial searching.
Dans *SIGMOD Conference*, pages 47–57, 1984.

[37] A. HIGGINS, L. BAHLER et J. PORTER.
Speaker verification using randomised phrase prompting.
Digital Signal Processing, 1:89–106, 1991.

[38] B. JACOB, J. MARIÉTHOZ, G. GRAVIER et F. BIMBOT.
Robustesse de la vérification du locuteur par mots de passe personnalisés.
XIIIèmes Journées d'étude sur la Parole (JEP), Aussois, Juin 2000.

[39] A. JAIN, A. ROSS et S. PARAHAKAR.
Biomandric-based web access.
www.citesser.nj.nec.com/433477.html, 1998.

[40] F. JELINEK.

Statistical Mandhods for Speech Recognition.
MIT Press, 2000.

[41] S. JULIER.
A general mandhod for approximating a nonlinear transformation of probability distributions.
Rapport technique, Oxford University, Department of Engineering Science, November 1996.

[42] L. KAUFMAN et P. J. ROUSSEEUW.
Finding Groups in Data.
1990.

[43] KAY et M. STEVEN.
Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory.
Prentice Hall PTR, March 1993.

[44] S. KULLER, J. NAOR et P. KLEIN.
The lattice structure of flow in planar graphs.
Rapport technique, College Park, MD, USA, 1990.

[45] S. KWON et S. NARAYANAN.
Speaker change dandection using a new weighted distance measure.
Dans *International Conference on Spoken Language Processing*, pages 2537–2540, Denver, CO, 16-20 2002.

[46] M. LASTRUCCI, Gori M. et G. SODA.
Neural autoassociators for phoneme-based speaker verification.
Esca Workshop on Speaker Recognition, Identification, and Verivifaction, pages 189–192, 1994.

[47] . E LEBARBIER.
Quelques approches pour la détection de ruptures à horizon fini.
Thèse de doctorat, Université Paris XI., 2002.

[48] E. LOISANT, R. SAINT-PAUL, J. MARTINEZ, G. RASCHIA et N. MOUADDIB.
Browsing clusters of similar images.
Dans *Actes des 19e Journées Bases de Données Avancées (BDA'2003)*, Lyon, France, 2003.

[49] Y. MAMI et D. CHARLAND.
Speaker identification by location in an optimal space of anchor models.
Dans *International Conferences on Spoken Language Processing (ICSLP '02)*, pages 1333–1336, Denver, Colorado, USA, 2002.

[50] Y. MAMI et D. CHARLAND.
Speaker identification by location in an optimal space of anchor models.
Dans *International Conferences on Spoken Language Processing (ICSLP '02)*, pages 1333–1336, Denver, Colorado, USA, 2002.

[51] S. MEIGNIER.
Indexation en locuteurs de documents sonores: Segmentation dun document et Appariement dune collection.
Thèse de doctorat, Université dAvignon, *Novembre* 2002.

[52] F. Ferdouille C. MEIGNIER S. BONASTRE, J et T. MERLIN.
Evolutive hmm for multi-speaker system.
Dans *International Conference on Acoustics Speech and Signal Processing (ICASSP '2000)*, 2000.

[53] H. MELIN.
On word boundary dandection in digit-based speaker verification.
Avril 1998.

[54] C. MONTACIÉ et M. J. CARATY.

Sound channel video indexing.
Dans *European Conference on Speech Communication and Technology (Eurospeech)*, pages 2359–2362, 1997.

[55] N. MOUADDIB et J. MARTINEZ.
Résumé de bases de données and application au données multimédias.
Dans *2e Rencontres Inter-Associations (AFIA, ARIA, EGC, INFORSID, SFC, SFDS, LMO, ASTI) : La classification and ses applications (RIAs'2006)*, Bron (Lyon), France, 20-21 2006.

[56] F. NACK et A. LINDSAY.
Everything you Wanted to Know about MPEG-7: Part 1.
IEEE Multimedia, $July \triangleleft September$ 1993 65'77.

[57] F. NACK et A. LINDSAY.
Everything you Wanted to Know about MPEG-7: Part 2.
IEEE Multimedia, $October \triangleleft Decembre$ 1993 64'73.

[58] M. NISHIDA et Y. ARIKI.
Real time speaker indexing based on subspace mandhod - application to tv news articles and debate.
Dans *International Conference on Spoken Langauge Processing (ICSLP'1998)*, pages 1347–1350, Sydney, Australia, 1998.

[59] S. OUMOUR-SAYOUD, H. SAHOUD et M. BOUDRAA.
Indexation des documents audio en vue d'un filtrage par locuteur = indexing audio documents for speaker filtering.
TALN 2002 Recital 2002 (Nancy, juin), 24:24–27, 2002.

[60] E. S PAGE.
Continous inspection schemes.
Biomandrika, 4:100–115, 1954.

[61] J. PINQUIER.
Indexation sonore : recherche de composantes primaires pour une structurationaudiovisuelle.
Thèse de Doctorat, Université Toulouse III Ecole Doctorale Informatique and Télécommunications, Toulouse, France., Dec 2004.

[62] A. E RAFTERY.
Bayesian model selection in social research (with discussion).
Rapport technique, University of Washington Demography Center Working., 1994 1994.
A revised version appeard in Sociological Mandhodology (1995).

[63] D. A. REYNOLDS.
Speaker verification using adapted gaussian mixture models.
Digital signal processing, vol 10:19–41, January 2000.

[64] D. A REYNOLDS, T.F. QUATIERI et R.B.DUNN.
Speaker verification using adapted gaussian mixture models.
igital Signal Processing (DSP), a review journal Special issue on NIST 1999speaker recognition workshop, 10(1-3):19–41, 2000.

[65] D.A REYNOLDS.
The effects of handsand variability on speaker recognition performance: experimentson the switchboard corpus.
Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'96)*, Atlanta, Georgie, USA, 1996.

[66] B. D. RIPLEY.
Pattern recognition and neural nandworks.
Campbridge University Press.

[67] J.E ROUGUI, M. GELGON, D. ABOUTAJDINE, N. MOUADDIB et M. RZIZA.
Organizing gaussian mixture models into a tree for scaling up speaker randrieval.
Pattern Recognition Landters, 28:1314–1319, 2007.

[68] J.E ROUGUI, M. GELGON, M. RZIZA, J. MARTINEZ et D. ABOUTAJDINE.
Time regions over lattice structure enable scaling up of speaker-based indexing.
Dans *Information and Communication technologies International Symposium ICTIS'07*, Fes, Maroc, 2007.

[69] J.E ROUGUI, Gelgon M., Rziza M., Martinez J. et Aboutajdine D..
Hierarchical organization of a sand of gaussian mixture speaker models forscaling up indexing and randrieval in audio documents.
Dans *ACM Symposium on Applied Computing (SAC'06), Multimedia and Visualization(MV) track*, pages 1369–1373, Dijon, France, 2006.

[70] J.E ROUGUI, M. RZIZA, D. ABOUTAJDINE, M. GELGON et J. MARTINEZ.
Fast incremental clustering of Gaussian mixture speaker models for scalingup randrieval in on-line broadcast.
Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'06)*, pages 521–524, Toulouse, France, mai 2006.

[71] R. SAINT-PAUL, G. RASCHIA et N. MOUADDIB.
Résumé généraliste de bases de donnéees.
Dans Véronique BENZAKEN, réd., *Proc. of Bases de Données Avancées (BDA 2005)*, St-Malo, France, October 2005. Université de Rennes 1.

[72] N. SAITOU et M. NEI.
The neighbor-joining mandhod: A new mandhod for reconstructing phylogenandic trees.
Molecular Biology and Evolution, vol 4:406–425, 1987.

[73] G. SCHWARZ.
Estimationg the dimension of model.
Annals of Statistics, 6:461–464, 1978.

[74] M. SECK.
Détection de ruptures et suivi de classe de sons pour lindexation sonore.
Thèse de doctorat, Université de Rennes, *Janvier* 2001.

[75] A. SHANDH et Klas W..
Multimeia Data Management: Using Mandadata to Integrate and Apply Digital Media.
McGraw-Hill, Serie on Data Warehousing and Data Management, 1998.

[76] W. A SHEWART.
Economic control of quality of manufactured product.
Prinston, 1931.

[77] D. SIEGMUND.
Sequential analysis - test and confidence intervals.
Springer, 1985.

[78] K. SÖMMEZ, L HECK et M WEINTRAUB.
Speaker tracking and dandection with multiple speakers.
Dans *European Conference on Speech Communication and Technology, Eurospeech'99*, volume 5, pages 2219–2222, 1999.

[79] A. SOLOMONOFF, Quillen C. et Campbell W. M..
Channel compensation for svm speaker recognition.
Dans *Odyssey 2004, The Speaker recognition Language Recognition Workshop*, 2004.

[80] D. A STEPHENS.
Bayesian randrospective multiple-changepoint identification.
Applied Statistics, 43(1):159–178, 1994.

[81] D.E. STURIM, D.A. REYNOLDS, D.A.SINGER et E. CAMPBELL.
Speaker indexing in large audio databases using anchor models.
Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP '01)*, pages 429–432, Salt Lake City, Utah, 2001.

[82] V. UPENDRA, J. NAVRATIL, G. N. RAMASWAMY et S.H. MAES.
Very large population text-independant speaker identification using transformationenhanced multi-grained models.
Dans *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP'01)*, Salt Lake City, Utah, mai 2001.

[83] A. J. VITERBI.
Error bounds for convolutionnal codes and an asymptotically optimum decoding algorithm.
IEEE Transactions on Information Theory, 13(2):260–269, 1967.

[84] S. VUURREN.
Comparison of text-independent speaker recognition mandhod on telephone speechwith acoustical mismatch.
pages 1788–1791, Philadelphie, pennsylvanie, USA, 1996. International Conference on Spoken Language Processing (ICSLP).

[85] V. WAN.
Speaker Verification using Support Vector Machines.
Thèse de Doctorat, University of Sheffield, United Kingdom, Juin 2003.

[86] P. ZEZULA, G. AMATO, V. DOHNAL et M. BATKO.
Similarity search the mandric space approach.
Advanced in Databases Systems, 32, 2006.

[87] B. ZHOU et J. HASEN.
Improved structural maximum likelihood engenspace mapping for rapid speaker adaptation.
Dans *International Conference on Spoken Languege Processing (ICSLP'2002)*, number 554-564, Denver, Colorado, 2002.

Liste des tableaux

— Corps du document —

Partie I — Présentation de l'indexation audio

Partie II — Vers une structuration hiérarchique des documents audio au sens du locuteur adaptée au problème incrémental

Table des matières

— Pages liminaires —

— Corps du document —

Indexation de documents audio: Cas des grands volumes de données

Jamal Eddine ROUGUI

Résumé

Cette thèse est consacrée à l'élaboration et l'évaluation des techniques visant à renforcer la robustesse des systèmes d'indexation de documents audio au sens du locuteur. L'indexation audio au sens du locuteur consiste à reconnaître l'identité des locuteurs ainsi que leurs interventions dans un flux continu audio ou dans une base de données d'archives audio, ne contenant que la parole. Dans ce cadre nous avons choisi de structurer les documents audio (restreints à des journaux radiodiffusés) selon une classification en locuteurs. La technique utilisée repose sur l'extraction des coefficients mel-cepstrales, suivi par l'apprentissage statistique de modèles de mélange de gaussiennes (MMG) et sur la détection des changements de locuteur au moyen de test d'hypothèse Bayésien. Le processus est incrémental : au fur et à mesure que de nouveaux locuteurs sont détectés, ils sont identifiés à ceux de la base de données ou bien, le cas échéant, de nouvelles entrées sont créées dans la base.

Comme toute structure de données adaptée au problème incrémental, notre système d'indexation permet d'effectuer la mise à jour des modèles MMG de locuteur à l'aide de l'algorithme fusion des MMG. Cet algorithme à été conçu à la fois pour créer une structure ascendante en regroupant deux à deux les modèles GMM jugés similaires. Enfin, à travers de deux études utilisant des structures arborescentes binaire ou n'aire, une réflexion est conduite afin de trouver une structure ordonnée et adaptée au problème incrémental. Quelques pistes de réflexions sur l'apport de l'analyse vidéo sont discutées et les besoins futurs sont explorés.

Motsclés : Reconnaissance automatique de locuteurs, bases de données multimédias, structuration audiovisuelle, classification hiérarchique, modèle de mélange de gaussiennes, divergence de Kullback-Leibler, architecture arborescente, structure incrémentale, Archivage audio.

Text-independent speaker technologies for Audio indexing and retrieval in the case of large data

Abstract

This thesis is devoted to techniques for speaker-based recognition systems to scale up to large amounts of data and speaker models. We have chosen to partition audio documents (news broadcast) according to speakers. The mel-cepstral acoustic characteristics of each speaker are model through a probabilistic Gaussian mixture model. First, speaker change detection in the stream is carried out by Bayesian hypothesis testing. The scheme is incremental : as new speakers are detected, they are either identified in the database or new entries are created in the database. First, we have examined some issues related to building a tree structure exploiting a similarity between speaker models. Several contributions were made. First, a proposal for organising a set of speaker models, based on an elementary model grouping. Then, we used an approximation of Kullback-Leibler divergence for this purpose. Finally, through two studies using binary of nary tree structures, we discuss the way of a version suitable for incremental processing. Finally, perspectives are drawn regarding joint audio/video analysis and future needs are analyzed.

Keywords: Speaker recognition, multimedia databases, audiovisual structuring, hierarchical classification, Gaussian mixture, Kullback-Leibler divergence, incremental processing.

Discipline : INFORMATIQUE Laboratoire : LINA (UMR CNRS 6241), équipe ATLASGRIM Spécialité : INFORMATIQUE LINA,

2 rue de la Houssinière, BP92208, 44322 Nantes cedex 03, FRANCE