Thèse

présentée pour obtenir le grade de docteur

de l'Ecole nationale supérieure des télécommunications

Spécialité : Signal et images

Yassine Mami

Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence

Soutenue le 21 octobre 2003 devant le jury composé de

Jean-Paul Haton	Président
Frédéric Bimbot	Rapporteurs
Jean-François Bonastre	
Régine André-Obrecht	Examinateurs
Delphine Charlet	
Christian Wellekens	Directeur de thèse

Ecole nationale supérieure des télécommunications

Résumé

Cette thèse s'inscrit dans le domaine de la reconnaissance automatique du locuteur, domaine riche d'applications potentielles allant de la sécurisation d'accès à l'indexation de documents audio. Afin de laisser le champ à un large éventail d'applications, nous nous intéressons à la reconnaissance du locuteur en mode indépendant du texte et dans le cas où nous disposons de très peu de données d'apprentissage. Nous nous intéressons plus particulièrement à la modélisation et à la représentation des locuteurs. Il s'agit d'estimer avec très peu de données un modèle suffisamment robuste du locuteur pour permettre la reconnaissance du locuteur. La modélisation par un mélange de gaussiennes (GMM), en mode indépendant du texte, fournit des bonnes performances et constitue l'état de l'art en la matière. Malheureusement, cette modélisation est peu robuste dans le cas où on ne dispose que de quelques secondes de parole pour apprendre le modèle du locuteur.
Pour tenter de remédier à ce problème, une perspective intéressante de modélisation consiste à représenter un nouveau locuteur, non plus de façon absolue, mais relativement à un ensemble de modèles de locuteurs bien appris. Chaque locuteur est représenté par sa localisation dans un espace de locuteurs de référence. C'est cette perspective que nous avons explorée dans cette thèse.
Au cours de ce travail, nous avons recherché le meilleur espace de représentation et la meilleure localisation dans cet espace. Nous avons utilisé le regroupement hiérarchique et la sélection d'un sous-ensemble pour construire cet espace. Les locuteurs sont ensuite localisés par la technique des modèles d'ancrage. Il s'agit de calculer un score de vraisemblance par rapport à chaque locuteur de référence. La proximité entre les locuteurs est évaluée par le calcul d'une distance entre leurs vecteurs de coordonnées. Nous avons proposé ensuite une nouvelle représentation des locuteurs basée sur une distribution de distances. L'idée est de modéliser un locuteur par une distribution sur les distances mesurées dans l'espace des modèles d'ancrage. Cela permet d'appliquer une mesure statistique entre l'occurrence de test et les modèles des locuteurs à reconnaître (au lieu d'une mesure géométrique). Ainsi, après avoir approfondi la modélisation d'un locuteur par sa position dans un espace de locuteurs de référence, nous avons également étudié comment cette position pouvait permettre une meilleure estimation du modèle GMM du locuteur, par exemple en fusionnant les modèles de ses plus proches voisins. Finalement, en complément à la modélisation GMM-UBM, nous avons étudié des algorithmes de fusion de décisions avec les différentes approches proposées.

Abstract

This thesis relates to the Automatic Speaker Recognition (ASR). The ASR is a field with many potential applications extending from access security to audio document indexing. In this thesis, the text-independent speaker recognition using very few training data is studied with a specific focus on speaker modeling and representation. The goal is to estimate with very few data a robust speaker model for speaker recognition. Gaussians mixture models (GMM), in text-independent speaker recognition, provide good performance and constitute the state-of-the-art. Unfortunately, this modeling is not robust enough when speaker model is trained with only a few seconds of speech.
To cope with this problem, an interesting modeling method consists in presenting a new speaker not absolutely but rather relatively, by comparing him to a set of well trained speakers. Each speaker is represented by his location in the space of the reference speakers. It is this approach which is explored in this thesis.
First the best representation space and the best location in the space are searched. Subsequently, a clustering and a subset selection to build the space are carried out. The speakers are then located by the Anchor Models technique which consists in computing a likelihood score for each reference speaker. The proximity between the speakers is evaluated by computing a distance between their coordinate vectors.
A new speaker representation based on a distance distribution is proposed in this thesis. The idea is to model a speaker by a distribution of distances measured in the anchor models space. That allows applying statistical measures between the test occurrence and the speaker models to be recognized (instead of a geometrical measure). After further investigation on speaker modeling by its position in the space of reference speakers, and the possibility to improve this position in order to obtain better estimate of the speaker GMM model is also studied. Thus, for example, merging the models of the speaker closest neighbors can be one of the solutions. Finally, in complement with modeling GMM-UBM fusion algorithms with various approaches are also proposed.

Table des matières

Table des figures

Liste des tableaux

Notations

x	Vecteur acoustique
X	Séquence de vecteurs acoustiques x
N et $n = 1, \cdots, N$	Nombre et indice de vecteurs acoustiques
D et $d = 1, \cdots, D$	Dimension et indice de l'espace acoustique
M et $m = 1, \cdots, M$	Nombre et indice de gaussiennes
t	Indice de temps (notamment dans les itérations)
$'$	Transposée
μ	Vecteur des moyennes
Σ	Matrice des covariances
σ^2	Variance
π	Poids des gaussiennes
λ	Modèle d'un locuteur
$\overline{\lambda}$	Modèle d'un locuteur de référence
$\mathscr{S}$	Nombre des locuteurs à reconnaître
S	Nombre de tous les locuteurs de référence
E et $e = 1, \cdots, E$	Nombre et indice des locuteur de référence sélectionnés
s	Indice des locuteurs
w	Vecteur des coordonnées d'un locuteur de dimension E
w_e	Composante du vecteur w

Acronymes

ACP	Analyse en Composantes Principales
ALD	Analyse Linéaire Discriminante
DMA	Distribution sur les Modèles d'Ancrage
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
NMAM	Neighborhood-Merged and Adapted Model
NMM	Neighborhood-Merged Model
REM	Ré-Estimation des Modèles par sélection de voisins
UBM	Universal Background Model

Introduction

La reconnaissance automatique du locuteur est interprétée comme une tâche particulière de reconnaissance de formes. Ce domaine regroupe les problèmes relatifs à l'identification ou à la vérification du locuteur sur base de l'information contenue dans le signal acoustique : il s'agit de reconnaître une personne à partir de sa voix. Le champ d'application est très large, allant des applications domestiques aux applications militaires, en passant par des applications judiciaires.

Au cours de ce travail, nous nous intéressons à la reconnaissance automatique de locuteur en mode indépendant du texte et dans le cas où nous disposons de très peu de données d'apprentissage. Il s'agit d'estimer avec peu de données un modèle suffisamment robuste du locuteur pour permettre la reconnaissance du locuteur. Cette thèse est donc dédiée essentiellement à la modélisation et à la représentation des locuteurs. Dans cette optique, plusieurs approches ont été proposées dans la littérature : approche vectorielle, connexionniste, statistique et prédictive. De ce large panel, l'approche statistique demeure au premier plan des systèmes de la reconnaissance automatique des locuteurs des récentes années.

En effet, la modélisation par un mélange de gaussiennes (GMM) fournit des bonnes performances, en mode indépendant du texte, et constitue l'état de l'art en la matière. Il s'agit de modéliser un locuteur par une somme pondérée de gaussienne. L'utilisation d'un modèle GMM se justifie essentiellement en faisant appel à l'interprétation des classes du mélange : chaque composante du mélange va représenter des évènements acoustiques. L'autre raison poussant à utiliser les GMM est qu'à l'aide d'une combinaison linéaire de gaussiennes, on peut représenter une large gamme de distributions. Malheureusement, cette modélisation n'est pas suffisamment robuste notamment si on dispose de peu de données d'apprentissage. Or dans la plupart des applications de reconnaissance de locuteurs, la phase d'enrôlement doit être très brève (de l'ordre de quelques secondes de parole).

Pour tenter de remédier à ce problème, une perspective intéressante de modélisation consiste à représenter un nouveau locuteur, non plus de façon absolue, mais relativement à un ensemble de locuteurs dont les modèles sont bien appris. Chaque locuteur est représenté par sa localisation dans un espace de référence. Par cette démarche, on espère obtenir une

modélisation fiable même avec peu de données.

Cette approche de modélisation a été utilisée en reconnaissance automatique de parole dans le domaine des techniques d'adaptation rapide au locuteur. Cette nouvelle approche de modélisation relative a donné naissance à la notion d'espace de locuteurs où un modèle de locuteur est représenté généralement par une combinaison linéaire des modèles de locuteurs de référence.

Dans cette thèse, nous appliquons ce principe de représentation relative à la reconnaissance automatique du locuteur. Dans un tel système, on distingue trois modules. Dans un premier temps, un espace de représentation est construit tandis que le deuxième est dédié à la localisation des locuteurs dans cet espace. Dans le dernier module, on effectue un test de reconnaissance.

Le premier module consiste à rechercher l'ensemble des locuteurs les plus représentatifs. Dans cette thèse, nous avons utilisé, essentiellement, le regroupement hiérarchique ascendant et la sélection pour construire l'espace représentatif. La deuxième phase consiste à placer chaque locuteur dans l'espace représentatif précédemment construit. Nous avons utilisé les principes des modèles d'ancrage pour localiser les locuteurs. Ainsi dans la troisième phase, les locuteurs sont représentés par des points dans l'espace et nous évaluons la proximité spatiale entre eux par des distances entre leurs vecteurs des coordonnées.

Par ailleurs, le regroupement hiérarchique ascendant ou la sélection ne fournit pas des espaces orthogonaux, c'est pourquoi nous appliquons un post-traitement sur les coordonnées des locuteurs. Il s'agit de faire une transformation géométrique afin de retrouver l'orthogonalité de l'espace et ensuite appliquer correctement les métriques entre les coordonnées des locuteurs.

Cependant, cette approche accorde une place symétrique à l'apprentissage et au test, ce qui peut être un défaut important puisque la quantité de données pour l'apprentissage et pour le test peuvent être radicalement différentes. Pour pallier ce problème, nous introduisons une asymétrie entre l'apprentissage et le test. Pour cela, nous présentons une nouvelle représentation des locuteurs basée sur une distribution de distances. L'idée est de représenter un locuteur par une densité de probabilité qui modélise ses distances à un ensemble de modèles de locuteurs de référence. En outre, nous pouvons introduire de l'information a priori dans l'estimation du modèle de locuteur.

Alors que, dans les approches relatives présentées précédemment, nous avons étudié comment la position d'un locuteur par rapport à un ensemble de locuteurs de référence pouvait être utilisée directement pour modéliser le locuteur, le positionnement relatif peut également servir à l'estimation d'un modèle GMM du locuteur. Ainsi, nous montrons que par un simple moyennage des modèles des voisins, nous capturons une information signifi-

cative dans la modélisation par GMM.

Ainsi, cet ouvrage s'articule en trois grandes parties. En introduction, nous rappelons le principe de la reconnaissance automatique du locuteur et nous présentons les différentes étapes du système de reconnaissance.
La deuxième partie de cette thèse est consacrée à la modélisation GMM où le modèle du mélange est présenté et suivi par des évaluations sur une base de données de France Télécom R&D et une base publique. Les résultats de cette évaluation constituent les résultats de base et de référence avec lesquels nous allons comparer les performances de la reconnaissance par localisation.
La troisième partie est consacrée à la représentation relative des locuteurs. Tout d'abord, nous dressons un état de l'art sur la modélisation relative. Nous présentons ensuite le principe de reconnaissance de locuteurs par placement dans un espace optimisé et nous détaillons les étapes du système de reconnaissance. Cette approche est évaluée au travers de séries d'expériences menées sur la base de données de France Télécom R&D.
Enfin, un ensemble de conclusions et de perspectives concluent le travail de cette thèse.

Chapitre 1

Système de reconnaissance automatique du locuteur

La reconnaissance automatique du locuteur (RAL) est un terme générique regroupant les problèmes relatifs à l'identification ou à la vérification du locuteur sur la base de l'information contenue dans le signal acoustique : il s'agit de reconnaître une personne à partir de sa voix. Un système de reconnaissance de locuteur procède en trois étapes : l'analyse acoustique du signal de parole, la modélisation du locuteur et une dernière étape de décision.
Dans ce chapitre, nous introduisons le principe de la reconnaissance automatique du locuteur et nous présentons les différentes étapes du système.

1.1 Introduction à la reconnaissance automatique du locuteur

Comme on peut le constater sur la figure 1.1, la reconnaissance automatique du locuteur s'inscrit dans le domaine plus général du traitement de la parole. Elle exploite la variabilité inter-locuteurs et s'intéresse aux informations extra-linguistiques du signal vocal.
Les variations individuelles entre locuteurs ont deux origines essentielles. D'abord, les caractéristiques morphologiques de l'appareil de phonation sont différentes pour chaque locuteur, indépendamment de la phrase prononcée. Ensuite, une même phrase n'est pas prononcée de la même façon par deux locuteurs ; on observe des différences dans les débits d'élocution, dans l'étendue des variations du pitch ou encore des différences liées à leur milieu socioculturel. Cette variabilité est l'essence même de la reconnaissance automatique du locuteur.
La reconnaissance automatique du locuteur est probablement la méthode la plus ergonomique pour résoudre les problèmes d'accès notamment dans le cas des transactions téléphoniques. Cependant, la voix ne peut être considérée comme une caractéristique biométrique

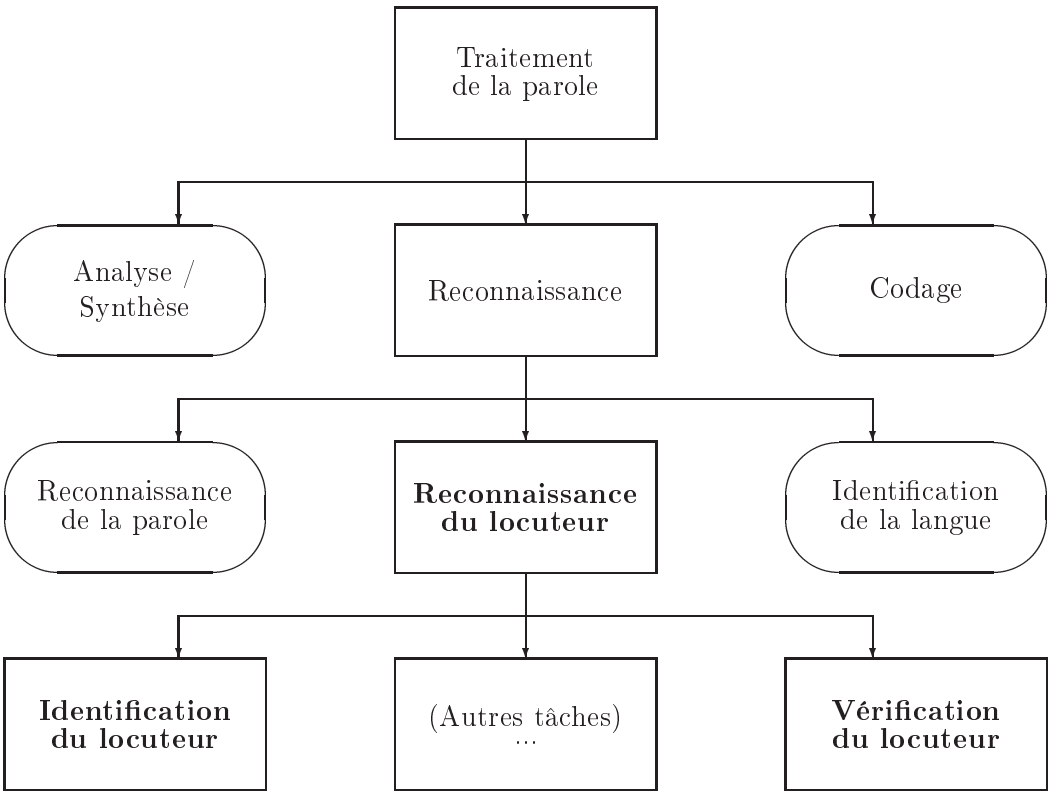


FIG. 1.1 – Traitement de la parole

d'une personne compte tenu de la variabilité intra-locuteur. Ainsi, on préfère la qualifier comme une signature vocale plutôt qu'une empreinte vocale.
Les applications potentielles des systèmes de reconnaissance sont nombreuses, incluant notamment la sécurisation accrue des cartes d'accès (par exemple, cartes de crédit et cartes téléphoniques), le contrôle d'accès à des bases de données et bâtiments protégés, commerce électronique, services d'information et de réservation, etc.

Identification et vérification du locuteur

On distingue deux tâches principales en reconnaissance du locuteur : identification et vérification. L'identification du locuteur consiste à reconnaître une personne parmi un ensemble de locuteurs en comparant son expression vocale à des références connues. Deux modes d'identification sont possibles : identification en ensemble fermé pour lequel le locuteur est identifié parmi un nombre connu de locuteurs ou bien identification en ensemble ouvert pour lequel le locuteur à identifier n'appartient pas forcément à cet ensemble.

La vérification (ou l'authentification) du locuteur consiste, après que le locuteur a décliné son identité, à vérifier l'adéquation de son message vocal avec la référence acoustique du locuteur qu'il prétend être. C'est une décision en tout ou rien.

Mode dépendant et indépendant du texte

On distingue également la reconnaissance du locuteur indépendante du contenu de la phrase prononcée (mode indépendant au texte) et la reconnaissance du locuteur qui prononce un mot ou une phrase clef (mode dépendant du texte). Les niveaux de dépendance au texte sont classés suivant les applications :

- Systèmes à texte libre (ou *free-text*) : le locuteur est libre de prononcer ce qu'il veut. Dans ce mode, les phrases d'apprentissage et de test sont différentes.
- Systèmes à texte suggéré (ou *text-prompted*) : un texte, différent à chaque session et pour chaque personne, est imposé au locuteur et déterminé par la machine. Les phrases d'apprentissage et de test peuvent être différentes.
- Systèmes dépendants du vocabulaire (ou *vocabulary-dependent*) : le locuteur prononce une séquence de mots issus d'un vocabulaire limité. Dans ce mode, l'apprentissage et le test sont réalisés sur des textes constitués à partir du même vocabulaire.
- Systèmes personnalisés dépendants du texte (ou *user-specific text dependent*) : chaque locuteur a son propre mot de passe. Dans ce mode, l'apprentissage et le test sont réalisés sur le même texte.

D'évidence, la connaissance a priori du message vocal rend la tâche des systèmes de RAL plus facile et les performances sont meilleures. La reconnaissance en mode indépendant du texte nécessite plus de durée de parole que le mode dépendant du texte.

Les sources d'erreurs

Le signal acoustique de la parole présente des caractéristiques qui rendent complexe son interprétation. L'information portée par ce signal peut être analysée de bien des façons et à plusieurs niveaux (acoustique, phonologique, morphologique, syntaxique, sémantique et pragmatique). Ce qui rend la tâche de traitement de la parole complexe. Plus particulièrement, on a vu que la variabilité inter-locuteurs est l'essence même de la reconnaissance. Il existe, cependant, plusieurs facteurs qui peuvent augmenter la variabilité intra-locuteur comme par exemple :

- L'état pathologique du locuteur (maladie, émotions, ...).
- Vieillissement (la voix d'une personne change au fur et à mesure de son vieillissement).
- Facteurs socioculturels (le locuteur peut changer d'accent).
- Locuteurs non coopératifs (notamment dans des applications judiciaires).
- Conditions de prise de son, bruit ambiant, ...

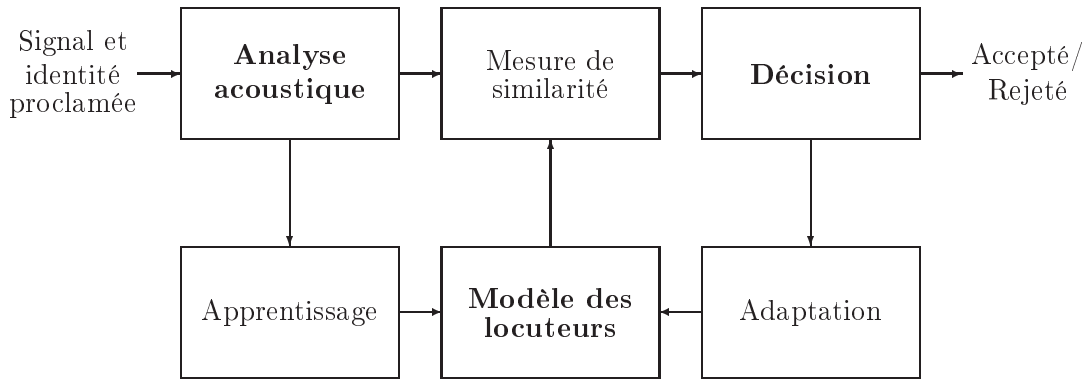


FIG. 1.2 – Schéma typique d'un système de vérification du locuteur

Systèmes de reconnaissance automatique de locuteur

La reconnaissance automatique du locuteur peut être interprétée comme une tâche de particulière de reconnaissance de formes. Différents modules sont présents dans ce système (figure 1.2). Tout d'abord, le message vocal, capté par un microphone, est converti en signal numérique. Il est ensuite analysé dans un étage d'analyse acoustique. A l'issue de cette étape, le signal est représenté par des vecteurs de coefficients pertinents pour la modélisation du locuteur. Dans l'étape d'apprentissage, on crée un modèle du locuteur. A la reconnaissance, un module de classification va mesurer la similarité entre les paramètres acoustiques du signal prononcé et les modèles de locuteurs présents dans la base. En dernier lieu, un module de décision, basé sur une stratégie de décision donnée, fournit la réponse du système. On peut également introduire un module d'adaptation pour augmenter les performances du système de reconnaissance.

1.2 Analyse acoustique du signal de parole

La mise en œuvre d'une tâche de reconnaissance de locuteur (ou de parole) est loin d'être facile, et ce pour deux raisons majeures. La première tient au fait que l'on ne maîtrise pas l'espace acoustique et en particulier la fonction de production d'un signal de parole. Aucune méthode analytique ne permet de prédire quelle va être la forme du signal de parole correspondant à l'émission d'un symbole donné par un locuteur particulier. La seconde, qui n'est qu'un corollaire de la première, est que la concrétisation acoustique d'un symbole donné n'est pas unique.

1.2.1 Production et perception du signal vocal

La phonation est réalisée au moyen d'un appareil qui n'est pas spécifique à la parole. Les organes qu'elles met en jeu sont d'abord affectés aux fonctions vitales de respiration et de nutrition. Le processus de phonation comporte trois étapes essentielles [Bartkova, 2002] :

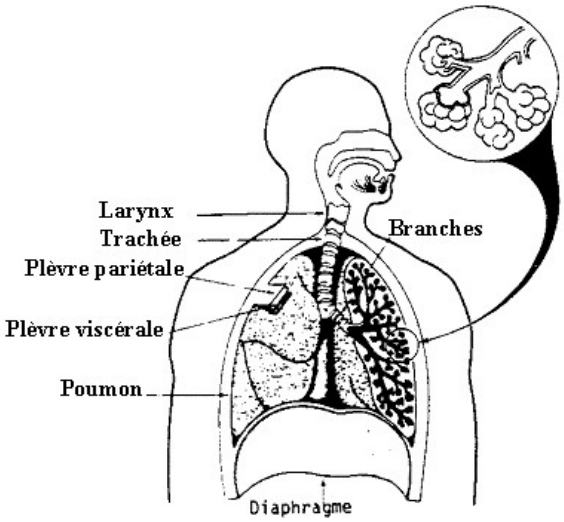


FIG. 1.3 – L'appareil phonatoire

- La génération d'une énergie ventilatoire qui va servir à mettre en mouvement oscillatoire les cordes vocales ou à les écarter afin de générer un bruit.
- La vibration des cordes vocales donnant naissance à tous les sons voisés, soit 80% du temps de phonation.
- La réalisation d'une disposition articulatoire dans ce qu'il est commode de désigner sous le nom de cavités surpra-glottiques.

Le système vocal se compose d'une soufflerie (poumons et conduit trachéo-bronchique), du larynx et du conduit vocal, lui même formé par le pharynx et les cavités orales et nasales (figure 1.3).

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production. L'appareil auditif se divise en deux parties : le système auditif périfirique correspondant à ce que l'on nomme communément l'oreille (qui se décompose en oreille externe, moyenne et interne) et le système auditif central.

Cependant, tout ce qui peut être acoustiquement mesuré ou observé par la phonétique articulatoire n'est pas nécessairement perçu. Les psychoacousticiens tentent de comprendre comment l'information auditive est traitée par le cerveau. En effet, au delà des caractéristiques mesurables (comme l'intensité et la fréquence), le son a deux qualités subjectives, la force et la hauteur, qui s'apprécient différemment. Les qualités subjectives relèvent des sensations éprouvées par un sujet qui écoute, et ne peuvent pas se mesurer sans lui. Ainsi, l'intensité perçue d'un son est égale à l'intensité physique (mesurée en décibels). Quant à sa hauteur, elle dépend de l'intensité avec laquelle ce son est transmis à l'auditeur. Par ailleurs, la prosodie assure la fonction de segmentation syntaxique et sémantique de l'énoncé [Zwicker et Feldtkeller, 1981] [Bartkova, 2002].

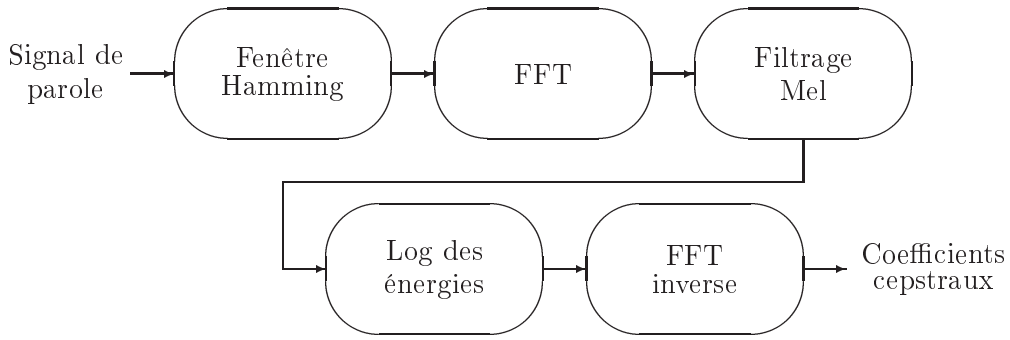


FIG. 1.4 – Calcul des coefficients cepstraux avec une échelle Mel

1.2.2 Les coefficients cepstraux

L'analyse acoustique du signal de parole consiste à extraire l'information pertinente et à réduire au maximum la redondance. Généralement, on calcule un jeu de coefficients acoustiques à des intervalles de temps réguliers, sur des blocs de signal de longueur fixe. Ce jeu de coefficients constitue un vecteur acoustique. Les techniques de paramétrisation acoustique sont nombreuses néanmoins on peut les regrouper en trois grandes familles :

- Analyse par bancs de filtres.
- Analyse par transformée de Fourier.
- Analyse par prédiction linéaire.

Les coefficients cepstraux issus d'une analyse par transformée de Fourier caractérisent bien la forme du spectre et permettent de séparer l'influence de la source de celle du conduit vocal. Ils peuvent aussi être calculés à partir d'une analyse de prédiction linéaire.
Le cepstre du signal de parole est défini comme la transformée de Fourier inverse du logarithme de la densité spectrale de puissance. Pour ce signal, la source d'excitation glottique est convoluée avec la réponse impulsionnelle du conduit vocal considéré comme un filtre linéaire :

$$s(t) = e(t) * h(t) \tag{1.1}$$

où $s(t)$ est le signal de parole, $e(t)$ est la source d'excitation glottique et $h(t)$ est la réponse impulsionnelle du conduit vocal.
L'application à l'équation 1.1 du logarithme du module de la transformée de Fourier donne :

$$\log|S(f)| = \log|E(f)| + \log|H(f)| \tag{1.2}$$

Par une transformée de Fourier inverse, on obtient :

$$s'(cef) = e'(cef) + h'(cef) \tag{1.3}$$

La dimension du nouveau domaine est homogène à un temps et s'appelle la *quéfrence* (cef), le nouveau domaine s'appelle donc le domaine *quéfrentiel*. Un filtrage dans ce domaine s'appelle *liftrage*.
Les coefficients cepstraux les plus répandus sont les MFCC (*Mel Frequency Cepstral Coefficients*). Ils présentent l'avantage d'être faiblement corrélés entre eux, et qu'on peut donc approximer leur matrice de covariance par une matrice diagonale. L'utilisation ici d'un filtrage *Mel* est justifiée par le fait qu'il reproduit la sélectivité de l'oreille qui diminue avec l'accroissement des fréquence.

1.3 Modélisation des locuteurs

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

- L'approche vectorielle : le locuteur est représenté par un ensemble de vecteurs de paramètres dans l'espace acoustique. Ses principales techniques sont la reconnaissance à base de DTW et par quantification vectorielle.
- L'approche statistique : consiste à représenter chaque locuteur par une densité de probabilité dans l'espace des paramètres acoustiques. Elle couvre les techniques de modélisation par les modèles de Markov cachés, par les mélanges de gaussiennes ou et par des mesures statistiques du second ordre.
- L'approche connexionniste : consiste principalement à modéliser les locuteurs par des réseaux de neurones.
- L'approche relative : il s'agit de modéliser un locuteur relativement par rapport à d'autres locuteurs de référence dont les modèles sont bien appris.

1.3.1 L'approche vectorielle

Reconnaissance du locuteur à base de DTW

Le reconnaissance par DTW (*Dynamique Time Warping*) repose sur le principe que chaque mot est représenté par une prononciation de référence (*template*). Compte tenu des décalages temporels entre les différentes prononciations d'un même mot, l'algorithme met en correspondance des séquences de paramètres par distortion temporelle (*Time Warping*). La programmation dynamique permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage ce qui signifie que c'est une technique exclusivement utilisée en mode dépendant du texte.
Cette approche est facile à mettre en œuvre et donne des performances relativement bonnes [Furui, 1981] [Booth et al., 1993].

Quantification vectorielle

Il s'agit de représenter l'espace acoustique par un nombre fini de vecteurs acoustiques. Cela consiste à faire un partitionnement de cet espace en régions, qui seront représentées par leur vecteur centroïde. Pour déterminer la distance d'un vecteur acoustique à cet espace, on effectue une mesure de distance avec chacun des centroïdes des régions et on retient la distance minimale. Si le vecteur acoustique provient du même locuteur pour lequel on a établit le dictionnaire de quantification, la distorsion sera en général moins grande que si ce vecteur provient d'un autre locuteur. Ainsi, on va représenter un locuteur par son dictionnaire de quantification. De nombreux articles proposent l'emploi de la quantification vectorielle en reconnaissance du locuteur. C'est le cas notamment des articles suivants : [Matsui et Furui, 1992], [Matsui et Furui, 1994] et [Yu et al., 1995]. On peut trouver une bonne description de cette méthode dans [L. Rabiner, 1993].

1.3.2 L'approche statistique

Modèles de Markov cachés

Les modèles de Markov (ou HMM pour *Hidden Markov Models*) ont été initialement introduits en reconnaissance de la parole. Puis leur utilisation s'est étendue peu à peu au domaine de la reconnaissance du locuteur. Dans cette approche, il ne s'agit plus d'une mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme acoustique ait été engendrée par le modèle de référence du locuteur. Le modèle d'un locuteur est constitué de l'association d'une chaîne de Markov, une succession d'états avec des probabilités de transition d'un état à l'autre, et de lois de probabilités (probabilités d'observation d'un vecteur acoustique dans un état). Quant à l'utilisation des modèles de Markov cachés en reconnaissance du locuteur, on peut se référer à [Savic et Gupta, 1990], [Rosenberg et al., 1991] et [Rissanen et Webb, 1993].

Les mélanges de gaussiennes

La reconnaissance du locuteur par mélanges de gaussiennes (ou GMM pour *Gaussian Mixture Models*) consiste à modéliser un locuteur par une somme pondérée de composantes gaussiennes [Reynolds, 1995]. Ainsi une large gamme de distributions peut être parfaitement représentée. Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques. L'utilisation de ce type de modèle semble être bien prometteuse. Il semble bien modéliser les caractéristiques spectrales des voix des locuteurs, et il est relativement simple à mettre en œuvre. Les mélanges de gaussiennes est considéré comme un cas particulier des HMM et une extenstion de la quantification vectorielle. Nous allons aborder cette méthode avec plus de détails dans le chapitre 2.

Mesures statistiques du second ordre

Cette partie présente une famille de mesures de similarité entre locuteurs. Ces mesures reposent sur les caractéristiques du second ordre d'une séquence de vecteurs, c'est-à-dire sur le vecteur moyen et la matrice de covariance de cette séquence. Plusieurs mesures de distance ont été utilisées, on peut citer : Le rapport de vraisemblance, la distance de *Kullbak-Leibler*, maximum de vraisemblance, test de sphéricité, déviation absolue des valeurs propres. Ces mesures donnent des résultats très encourageants sur la parole propre, et, naturellement, voient leurs performances se dégrader sur la parole téléphonique. De part leur relative simplicité, ces mesures peuvent également servir de référence pour évaluer la qualité d'une base de données [Bimbot et al., 1995] et [Magrin-Chagnolleau et al., 1995].

1.3.3 L'approche connexionniste

Les réseaux de neurones ont été assez largement utilisés en reconnaissance du locuteur. Ils offrent en effet une bonne alternative au problème de la discrimination entre les locuteurs. Ces outils de classification permettent de séparer des classes, dans un espace de représentation donné, de façon non linéaire. Pour une bonne description de cette technique, on peut lire [Oglesby et Mason, 1989], [Artières et Gallinari, 1994] et également [Homayounpour et Chollet, 1995]. L'inconvénient important de l'application de cette technique en identification du locuteur est le coût important lié à l'ajout d'un nouveau locuteur dans la base de référence (ce n'est pas le cas en vérification du locuteur). On peut aussi utiliser les réseaux de neurones en les couplant à d'autres techniques, comme par exemple les modèles de Markov cachés. On parle alors de méthodes hybrides.

1.3.4 L'approche relative

Cette nouvelle technique consiste à modéliser un locuteur non plus de façon absolue mais relativement à un ensemble de locuteurs bien appris. L'état de l'art de cette technique sera abordé avec plus de détails au paragraphe 5.1.

1.4 Décision et mesure des performances

Comme on l'a déjà vu, on distingue deux tâches principales en reconnaissance du locuteur : la vérification du locuteur et l'identification du locuteur. Cependant, un système de RAL peut aussi servir à identifier les segments de chaque locuteur dans un document audio, à la poursuite du locuteur ou à faire l'indexation des documents audio :

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

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

et

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

La vérification du locuteur consiste, après que le locuteur a décliné son identité, à vérifier l'adéquation du message vocal avec la référence acoustique du locuteur qu'il prétend être. C'est une décision en tout ou rien. Les performances de vérification de locuteur sont données en termes des faux rejets f_r et de fausses acceptations f_a :

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

$$f_a = \frac{\text{Nombre de tentatives d'imposteurs acceptées}}{\text{Nombre total de tentatives d'imposteurs}}$$

La segmentation en locuteurs consiste à découper un flux sonore avec ou sans modèle explicite des locuteurs. Dans le cas ou il n'y a pas de modèle, le processus procède généralement en deux phases : la détection de changements de locuteurs et le regroupement de locuteurs segments (appartenant au même locuteur). La première phase repose sur le principe de calcul d'une distance entre deux portions de signal consécutives. Si elle excède un seuil, on décide qu'il y a eu changement de locuteur. La deuxième phase consiste à regrouper les segments de parole par locuteur selon un critère de distance. Une bonne segmentation fournit les changements de locuteurs corrects et des segments ne contenant qu'un seul locuteur. Nous distinguons deux types d'erreur pour la détection de changements de locuteurs :

- Une fausse alarme (FA) a lieu lorsqu'un changement de locuteur est détecté alors qu'il n'existe pas.
- Une détection manquée (DM) a lieu quand un changement de locuteur existant n'est pas détecté.

Une valeur élevée de FA est significative d'une sur-segmentation et une valeur élevée de DM indique une sous-segmentation.

La poursuite du locuteur (*Speaker tracking*) se fait avec un modèle du locuteur, contrairement à la segmentation qui peut se faire sans modèle. Elle consiste à déterminer quand une personne parle dans une conversation.

Indexation des documents audio : Grâce au développement des technologies numériques, les besoins en outils d'indexation se font cruellement ressentir. Il devient donc indispensable d'indexer automatiquement les documents audio pour être exploitables. La clé d'indexation qui nous intéresse ici est l'identité du locuteur : nous voudrions savoir *qui parle et quand*. En pratique, on dispose des documents audio représentés par leurs modèles respectifs. La phase de recherche du système d'indexation consiste, généralement, à évaluer des mesures de similarité entre la requête et ces différents modèles. Par ailleurs, le système d'indexation par locuteurs peut servir également comme étape préliminaire pour des tâches de transcription ou pour le suivi de locuteurs.

1.5 Conclusion

Dans ce chapitre, nous avons introduit le principe de la reconnaissance automatique du locuteur ainsi que les différentes étapes du système. La reconnaissance automatique du locuteur est probablement la méthode la plus ergonomique pour résoudre les problèmes d'accès. Cependant, la voix ne peut être considérée comme une caractéristique biométrique d'une personne compte tenu de la variabilité intra-locuteur. Un système de reconnaissance de locuteur procède généralement en trois étapes : l'analyse acoustique du signal de parole, la modélisation du locuteur et une dernière étape de décision. En analyse acoustique, les MFCC sont les coefficients acoustiques les plus répandus. Quant à la modélisation, l'approche GMM constitue l'état de l'art en RAL, en mode indépendant du texte. La décision d'un système de reconnaissance automatique du locuteur est basée sur les deux processus d'identification et/ou de vérification de locuteur, et cela quelle que soit l'application ou la tâche visée.

Première partie

Reconnaissance des locuteurs par mélanges de gaussiennes (GMM)

Chapitre 2

Les mélanges de gaussiennes

Les mélanges de gaussiennes sont utilisés pour modéliser un locuteur donné par une somme pondérée de gaussiennes. On peut assimiler un modèle GMM (*Gaussian Mixture Models*) à un HMM (*Hidden Markov Model*) à un seul état. On ne modélise donc pas les aspects temporels du signal. Cette méthode est la plus utilisée en ce qui concerne la reconnaissance du locuteur en mode indépendant du texte. Les travaux de D. A. Reynolds [Reynolds, 1995] constituent l'état de l'art en la matière.
L'utilisation d'un modèle GMM se justifie essentiellement en faisant appel à l'interprétation des classes du mélange : il est certain que les vecteurs de paramètres vont se répartir différemment selon les caractéristiques du son de parole considéré (son voisé / non voisé, ou plus finement en fonction du phonème). Chaque composante va modéliser des ensembles sous-jacents de classes acoustiques, chaque classe représentant des évènements acoustiques (voyelles, nasales, ...). Ainsi, l'allure spectrale de la $i$ème classe pourra être représentée par la moyenne et la matrice de covariance de la $i$ème composante. Ces classes caractérisent l'espace acoustique propre à chaque locuteur.
L'autre raison poussant à utiliser les GMM est qu'à l'aide d'une combinaison linéaire de Gaussiennes, on peut représenter une large gamme de distributions.
Dans ce chapitre, nous présenterons le modèles du mélange de gaussiennes. Nous aborderons ensuite le problème d'estimation de l'ensemble de ces paramètres et enfin, nous détaillerons la phase de décision d'un système de reconnaissance de locuteur par GMM.

2.1 Modèle du mélange

Un mélange de Gaussiennes est une somme pondérée de M densités gaussiennes. Soit un locuteur s et un vecteur acoustique x de dimension D, le mélange de gaussiennes est défini comme suit :

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

où les $b_m^s(x)$ représentant des densités gaussiennes, paramétrées par un vecteur de moyenne μ_m^s et une matrice de covariance Σ_m^s :

$$b_m^s(x) = \frac{1}{(2\pi)^{D/2} \cdot |\Sigma_m^s|^{1/2}} \times \exp\left[-\frac{1}{2}(x-\mu_m^s)'(\Sigma_m^s)^{-1}(x-\mu_m^s)\right] \tag{2.2}$$

et les π_m^s représentent les poids du mélange, avec $\sum_{m=1}^{M} \pi_m^s = 1$.
Un locuteur est donc modélisé par un ensemble de paramètres noté λ_s :

$$\lambda_s = \{\pi_m^s, \mu_m^s, \Sigma_m^s\}_{m=1,\cdots,M} \tag{2.3}$$

Ce modèle peut prendre plusieurs formes, notamment en ce qui concerne les matrices de covariance. On peut assigner une matrice de covariance à chaque gaussienne, ou bien utiliser une matrice de covariance globale, commune à toutes les gaussiennes. De plus, elles peuvent être pleines ou diagonales (en raison de la faible corrélation des coefficients melcepstraux, on considèrera généralement les matrices de covariance diagonales, cf. 1.2.2).

2.2 Apprentissage du modèle

Il s'agit, lors de la phase d'apprentissage, d'estimer l'ensemble λ des paramètres d'un modèle GMM de locuteur[1]. La méthode conventionnelle est celle du Maximum de Vraisemblance (MV) dont le but est de déterminer les paramètres du modèle qui maximisent la vraisemblance des données d'apprentissage. Pour une séquence de N vecteurs d'apprentissage $X = \{x_1, x_2, \cdots, x_N\}$, la vraisemblance du modèle GMM est :

$$p(X|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda) = \prod_{n=1}^{N} \sum_{m=1}^{M} p(x_n|\pi_m, \mu_m, \Sigma_m) \tag{2.4}$$

En remplaçant l'expression de $p(x_n|\lambda)$ on obtient une expression complexe de la vraisemblance et il n'y a malheureusement pas de solution analytique à ce problème. De plus, le calcul de cette expression conduit au logarithme d'une somme et à une fonction non linéaire des paramètres du modèle λ ce qui rend la maximisation directe très difficile.
Cependant, la variable indicatrice m est une donnée constitutive du problème qui présente l'inconvénient de ne pouvoir être observée en pratique : on observe des réalisations du vecteur aléatoire x_n sans savoir de manière certaine quelle est la classe du mélange associée à chaque observation. Au sens de l'algorithme EM, la variable m constitue une donnée latente, c'est-à-dire fortement suggérée par le problème considéré (on parle également de donnée non-observée ou manquante). Nous verrons que l'introduction de ces données non-observées permet de résoudre de manière élégante un problème d'estimation relativement complexe et que ce type de problème est adapté à l'algorithme d'apprentissage EM.

[1] Pour alléger les formules, nous avons volontairement supprimé l'indice s du locuteur.

2.2.1 Apprentissage par Maximum de Vraisemblance : l'algorithme Expectation-Maximization (EM)

Le problème de l'algorithme EM (*Expectation-Maximization*) peut être considéré comme un cas particulier de gradient [Dempster et al., 1977]. Il fait intervenir à la fois des observations X et des variables manquantes (l'indice de la gaussienne $m = 1, \cdots, M$). Cet algorithme maximise, de façon itérative, la fonction de la vraisemblance. Cette maximisation n'est pas directe, elle fait intervenir la fonction auxiliaire $Q(\theta, \theta^{(t)})$ qui est définie comme étant l'espérance mathématique du logarithme de la vraisemblance jointe (incluant les variables observées et les variables cachées) sur l'ensemble complet des variables d'entraînement, calculée sur base des paramètres courants [Boite et al., 1999], à savoir :

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M} \sum_{n=1}^{N} p(m|x_n, \theta^{(t)}) \log p(x_n, m|\theta) \tag{2.5}$$

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

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\log \pi_m - \frac{D}{2} \log(2\pi) - \frac{1}{2} \log |\Sigma_m| \right] \\ &- \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\frac{1}{2} (x_n - \mu_m)' \Sigma_m^{-1} (x_n - \mu_m) \right] \end{aligned} \tag{2.6}$$

où $\gamma_{n,m}^{(t)}$ est une probabilité a posteriori estimée à l'itération t :

$$\gamma_{n,m}^{(t)} = \frac{\pi_m^{(t)} p(x_n|\mu_m^{(t)}, \Sigma_m^{(t)})}{\sum_{k=1}^{M} \pi_k^{(t)} p(x_n|\mu_k^{(t)}, \Sigma_k^{(t)})} \tag{2.7}$$

En supposant que $p(x_n|\theta)$ sont des densités gaussiennes à matrices de covariance diagonales, l'expression de la fonction auxiliaire devient :

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \log \pi_m \\ &- \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\text{Cste} + \log \sigma_m^2 + \frac{(x_n - \mu_m)^2}{\sigma_m^2} \right] \end{aligned} \tag{2.8}$$

où σ_m^2 est un élément diagonal de la matrice de covariance.
Les paramètres sont estimés en annulant la dérivée partielle de la fonction auxiliaire Q par rapport à chacun de ceux-ci. Le cas des poids des composantes de mélange π_m est assez simple puisqu'il s'agit de paramètres scalaires. Ceci dit, il faut tenir compte de la contrainte

qui existe sur ces paramètres ($\sum_{m=1}^{M} \pi_m = 1$). La maximisation sous contrainte se résout simplement en introduisant un multiplicateur de Lagrange associé à cette contrainte et l'on obtient :

$$\pi_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,m}^{(t+1)} \tag{2.9}$$

En ce qui concerne les vecteurs des moyennes, on montre que les formules de réestimations sont données par :

$$\mu_m^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} x_n}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{2.10}$$

et pour les variances :

$$\sigma_m^{2(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} (x_n - \mu_m^{(t)})^2}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{2.11}$$

2.2.2 Apprentissage par Maximum A Posteriori

L'algorithme EM est un des algorithmes les plus importants et les plus puissants en estimation statistique. De plus, il bénéficie d'une preuve de convergence garantissant que l'itération de l'étape d'estimation et de maximisation converge vers un maximum de la fonction de vraisemblance [Boite et al., 1999]. Cependant, ses limites apparaissent lorsqu'on dispose de peu de données. Donc, il est important d'introduire de l'information a priori. Par conséquent, on ne cherche plus à maximiser la vraisemblance des données mais plutôt la probabilité a posteriori.

Apprentissage incrémental des modèles

L'apprentissage incrémental est un apprentissage bayésien simplifié. Il correspond à un apprentissage MAP (*Maximum A Posteriori*) [Gauvain et Lee, 1994] avec un choix particulier des paramètres a priori [Mokbel et Collin, 1999]. Les formules de ré-estimation, pour une gaussienne m, sont les suivantes :

– Les poids des gaussiennes :

$$\pi_m = \frac{n_m^0 + n_m}{\sum_{k=1}^{M} (n_k^0 + n_k)} \tag{2.12}$$

– Les vecteurs des moyennes :

$$\mu_m = \frac{n_m^0 \overline{X_m^0} + n_m \overline{X_m}}{n_m^0 + n_m} \tag{2.13}$$

– Les variances :

$$\sigma^2{}_m = \frac{n_m^0 \overline{X_m^0 X_m^{0\prime}} + n_m \overline{X_m X_m'}}{n_m^0 + n_m} - \mu_m \mu_m' \tag{2.14}$$

où n (respectivement n^0) représente le poids, $\overline{X}$ (respectivement $\overline{X^0}$) le moment d'ordre 1 et $\overline{XX'}$ (respectivement $\overline{X^0 X^{0\prime}}$) le moment d'ordre 2 des données à adapter X (respectivement des données initiales X^0).
L'apprentissage incrémental consiste à effectuer quelques itérations d'apprentissage sur les données d'adaptation en conservant l'information apportée par les données initiales X^0. Dans le cas où de nombreuses données sont disponibles, l'apprentissage incrémental (ou plus généralement l'estimateur MAP) converge vers les estimateurs du maximum de vraisemblance. Il permet d'obtenir de nouveaux modèles avec peu de données. Ces estimées seront plus fiables que celle obtenues par MV étant donné qu'elles intègrent des connaissances a priori.
Cette approche est la plus utilisée en reconnaissance du locuteur en mode indépendant du texte. La valeur du poids initial n^0 est empirique et comprise entre (8-20) [Reynolds et al., 2000].

Initialisation

Les valeurs initiales d'une densité multi-gaussienne peuvent être obtenues par différentes méthodes comme par exemple, la QV (Quantification vectorielle) ou par éclatement de gaussiennes. Cette initialisation est suivie ensuite par un apprentissage EM ou par une adaptation incrémentale. En GMM, le modèles initial correspond au modèle du monde UBM (*Universal Background Model*) [Reynolds et al., 2000].

2.3 Décision

Toute application de reconnaissance du locuteur peut se voir comme une déclinaison des processus de décision principaux que sont l'identification et la vérification. C'est pourquoi, dans cette partie, nous allons présenter la phase de décision d'un système d'identification et de vérification par GMM.

2.3.1 Vérification du locuteur

La stratégie de décision en vérification du locuteur a fait l'objet de plusieurs travaux de recherche. Dans ce paragraphe, nous présenterons le test d'hypothèses utilisé en vérification du locuteur et les différentes approches d'estimation du modèle du rejet. Ensuite, nous introduirons quelques techniques de normalisation de scores.
En vérification du locuteur, nous utilisons souvent le test d'hypothèses suivant :

- H_0 : le segment de parole X a été prononcé par le locuteur λ.
- H_1 : le segment de parole X a été prononcé par un imposteur.

On calcule le rapport de vraisemblance LR (*Likelihood Ratio*) donné par :

$$LR = \frac{p(X|H_0)}{p(X|H_1)} \lessgtr \theta \qquad \begin{cases} \text{on accepte } H_1 \\ \text{on accepte } H_0 \end{cases} \tag{2.15}$$

où θ est un seuil qui peut dépendre du modèle du locuteur λ. L'hypothèse H_0 correspond au modèle du locuteur λ et l'hypothèse H_1 au modèle de rejet.
Généralement, l'hypothèse H_1 correspond à un modèle de rejet indépendant du locuteur appelé modèle du monde λ_{UBM} et on évalue plutôt le LLR (*Log-Likelihood Ratio*), soit :

$$LLR = \Lambda = \log p(X|\lambda) - \log p(X|\lambda_{UBM}) \tag{2.16}$$

Le LLR est ensuite comparé à un seuil.
Cependant, il est relativement facile d'estimer le modèle du locuteur λ mais il n'en est pas de même pour le modèle du rejet. Ce dernier correspond souvent au modèle du monde [Carrey et al., 1991]. Le modèle de rejet peut aussi correspondre aux cohortes de locuteurs. Cet ensemble peut être spécifique d'un locuteur λ [Rosenberg et al., 1992].
Par ailleurs, d'autres techniques de normalisation de scores ont été introduites telles que la *Z-norm*, la *T-norm* et la *D-norm*.
La *Z-norm* (pour *Zero Normalization*) consiste à ramener les scores à espace de comparabilité commun :

$$S_{Z-norm} = \frac{\log p(X|\lambda) - \mu_I}{\sigma_I}$$

où les paramètres de cette normalisation μ_I et σ_I sont estimés, off-line, à partir d'un autre corpus de données. Ils représentent la moyenne et la variance des scores du corpus sur le modèle du locuteur [Auckenthaler et al., 2000].
La normalisation *T-norm* (pour *Test Normalization*) est aussi basée sur l'estimation d'une moyenne et d'une variance. Durant la phase de test, un ensemble des modèles imposteurs est utilisé pour calculer le score de log-vraisemblance d'une occurence de test. La moyenne et la variance sont estimées à partir de ces scores [Auckenthaler et al., 2000].
La normalisation *D-norm* (pour *Distance Normalization*) ne nécessite aucune donnée de parole supplémentaire ni de population de locuteurs externes. Les scores des locuteurs sont normalisés par la distance de Kullback-Leibler entre modèles du locuteur et du non locuteur [Seck et al., 2000].

2.3.2 Identification du locuteur

Soit un groupe de $\mathscr{S}$ locuteurs, représentés par les modèles GMM : $\lambda_1, \lambda_2, \cdots, \lambda_{\mathscr{S}}$. L'objectif de la phase d'identification est de trouver, à partir d'une séquence observée X, le modèle qui a la probabilité a posteriori maximale, c'est-à-dire :

$$\widehat{s} = \arg \max_{1 \leq s \leq \mathscr{S}} p(\lambda_s|X) \tag{2.17}$$

ce qui donne, d'après la loi de Bayes :

$$\widehat{s} = \arg \max_{1 \leq s \leq \mathscr{S}} \frac{p(X|\lambda_s)}{p(X)} p(\lambda_s) \tag{2.18}$$

Supposant l'équiprobabilité d'apparition des locuteurs ($p(\lambda_s) = \frac{1}{\mathscr{S}}$) et que la probabilité $p(X)$ d'apparition d'une séquence X est la même pour tous les locuteurs, la loi de classification devient :

$$\widehat{s} = \arg \max_{1 \leq s \leq \mathscr{S}} p(X|\lambda_s) \tag{2.19}$$

En utilisant le logarithme et l'indépendance entre les observations, le système d'identification calcule le score suivant :

$$\widehat{s} = \arg \max_{1 \leq s \leq \mathscr{S}} \sum_{n=1}^{N} \log p(x_n|\lambda_s) \tag{2.20}$$

où $p(x_n|\lambda_s)$ est donnée par l'équation 2.1.
En identification en ensemble fermé, les performances se mesurent en terme du taux d'identification incorrecte I_i et se dégradent considérablement si le nombre de locuteurs $\mathscr{S}$ augmente. En identification en ensemble ouvert, on mesure en plus les faux rejets f_r et les fausses acceptations f_a (cf. paragraphe 1.4).

2.4 Conclusion

Les mélanges de gaussiennes constituent l'état de l'art en reconnaissance automatique du locuteur, en mode indépendant du texte. Il existe plusieurs techniques pour apprendre un modèle GMM. En premier lieu, l'algorithme EM permet d'estimer les paramètres du modèle tout en offrant un formalisme théorique et une preuve de convergence. Malheureusement, ses limites apparaissent lorsqu'on dispose de peu de données. Dans ce cas, il est plus judicieux d'utiliser un apprentissage par adaptation MAP. Cette estimation sera plus fiable que celle obtenue par MV étant donné qu'elle intègre des connaissances a priori.

Chapitre 3

Contexte expérimental

Ce chapitre présente le contexte expérimental des évaluations de reconnaissance de locuteur, en mode indépendant du texte. En premier lieu, nous décrivons les bases de données utilisées. Ensuite, nous rappelons l'analyse acoustique appliquée ainsi que l'algorithme d'apprentissage des modèles. Enfin, nous présentons le protocole d'évaluation en identification et vérification du locuteur.
Notons que ce chapitre décrit les conditions expérimentales de toutes les évaluations de reconnaissance tant par GMM que par l'approche relative.

3.1 Bases de données

3.1.1 Base de données de France Télécom R&D

La base de données de parole utilisée est une base téléphonique interne à France Télécom R&D. Elle comporte 1107 locuteurs. Cette base est divisée en quatre sous-ensembles :

- L'ensemble $\mathscr{E}_1$ composé de 50 locuteurs à reconnaître (33 femmes et 17 hommes), utilisés comme corpus de test.
- L'ensemble $\mathscr{E}_2$ composé de 57 locuteurs (23 femmes et 34 hommes, différents des locuteur de l'ensemble $\mathscr{E}_1$) utilisés comme corpus de développement pour estimer certaines transformations[1].
- L'ensemble $\mathscr{E}_3$ composé de 500 locuteurs (219 femmes et 281 hommes) utilisés pour estimer les informations a priori sur l'espace des locuteurs[2].
- L'ensemble $\mathscr{E}_4$ composé de 500 (différents de $\mathscr{E}_3$) utilisés pour les tests d'impostures en vérification du locuteur.

Les locuteurs des ensembles $\mathscr{E}_1$ et $\mathscr{E}_2$ ont suivi le même protocole de collecte. Pour chaque locuteur de $\mathscr{E}_1$ et $\mathscr{E}_2$, on dispose de :

[1] Les transformation ACP et ALD seront expliquées plus loin (cf. paragraphe 7.2.3).
[2] Pour plus de détails, consulter le chapitre 6.

- 5 appels de 25 phrases d'une durée moyenne de 4 secondes réservées à l'apprentissage des modèles (soit 125 phrases d'apprentissage) ;
- 25 appels de 5 phrases d'une durée moyenne de 4 secondes réservées au test enregistrées durant plusieurs mois (soit 125 phrases de test).

La durée moyenne des phrases est de l'ordre de 4 secondes, soit environ 250 trames. Les phrases de cette base sont lues et extraites du journal *Le Monde*.
Etant donné que les 57 locuteurs de l'ensemble $\mathscr{E}_2$ disposent de beaucoup de données, ils seront utilisés pour construire les matrices de transformation de l'ACP et de l'ALD. En effet, pour estimer les matrices inter et intra-classes, chaque classe (ou chaque locuteur) doit disposer d'un nombre suffisant d'observations.
En revanche, les locuteurs de l'ensemble $\mathscr{E}_3$ disposent de moins de données mais ils sont suffisamment nombreux pour construire un espace de locuteurs. Chaque locuteur de l'ensemble $\mathscr{E}_3$ (et de $\mathscr{E}_4$) dispose d'une quinzaine de phrases, relativement courtes. Elles peuvent être :

- des phrases lues et extraites du journal *Le Monde*.
- une suite de chiffres ;
- des réponses à des questions ;
- les cinq phrases suivantes : "*Zazou va en avril au gymnase", "Ma mère m'a donné l'autorisation", "Isabelle a les yeux bleus", "On n'a pas eu de neige l'année dernière" et "Les oiseaux voyagent en hiver".*

Ce qui fait un total d'environ 95 secondes de parole par locuteur.
Les conditions de prise de son de l'ensemble des locuteurs varient d'une phrase à une autre, mais la qualité générale des enregistrements est de type Réseau Téléphonique Commuté (RTC).

3.1.2 NIST

Afin de pouvoir comparer nos résultats, nous évaluons également notre système de vérification du locuteur par GMM sur une base de données publique. Pour cela, nous utilisons une partie de la base Switchboard, qui a servi aux évaluations de NIST 2000 (*National Institute of Standards and Technology*).

Les données utilisées sont extraites de Switchboard-2 Phase-3[3]. Elle contient 2728 appels et comporte 640 locuteurs (348 femmes et 292 hommes). Les appels sont enregistrés dans plusieurs conditions, avec un microphone électrique ou charbon. La durée des phrases d'apprentissage est comprise entre 110 et 130 secondes. Quant aux phrases de test, la durée est comprise entre 15 et 45 secondes.

[3]Ce corpus a été collecté par le consortium LDC (*Linguistic Data Consortium*)

3.2 Analyse acoustique

Dans nos expériences, une analyse est appliquée toutes les 16 ms sur des fenêtres d'analyse de 32 ms (par glissement et recouvrement des fenêtres d'analyse). A chaque trame, on associe un vecteur de représentation acoustique, composé de l'énergie temporelle de la trame et des 13 premiers MFCC. A cela on rajoute leurs dérivées première et seconde ; ce qui donne 42 coefficients acoustiques. On applique ensuite un module de soustraction cepstrale.

3.3 Détection d'activité vocale

Un module de détection d'activité vocale est appliqué sur les vecteurs acoustiques. On ne retient que les trames étiquetées "Parole". Ce module est une solution propriétaire de France Télécom R&D. Sur les données de France Télécom et NIST, cette sélection conserve en moyenne 75% des trames.

3.4 Apprentissage des modèles

L'apprentissage des modèles GMM des locuteurs est un apprentissage incrémental. Il s'agit d'un algorithme itératif qui adapte un modèle de référence λ_{UBM}, indépendant du locuteur, aux données d'apprentissage du locuteur considéré (paragraphe 2.2.2). Le poids donné aux données initiales est fixé à 10 [Reynolds et al., 2000]. D'autre part, le modèle du monde λ_{UBM} est appris à partir de l'ensemble $\mathscr{E}_3$ (approximativement 5 heures de parole). Le noyau de reconnaissance sur lequel nous allons nous appuyer est le moteur de reconnaissance *PhilSoft*® de France Télécom R&D. Cependant l'implémentation des GMM dans ce décodeur diffère légèrement de celle présenté dans le chapitre 2 (équation 2.1). Le GMM est implémenté de la façon suivante :

$$p(x|\lambda_s) = \max_{i=1}^{M} \pi_i^s b_i^s(x) \tag{3.1}$$

3.5 Protocole d'évaluation

3.5.1 Evaluation des performances

En identification de locuteur, nous allons évaluer les performances d'identification des 50 locuteurs de l'ensemble $\mathscr{E}_1$ (de la base de France Télécom), i.e. il s'agit d'identifier un locuteur parmi les 50 locuteurs (ensemble fermé) et de calculer les taux d'identification incorrecte (cf. paragraphe 1.4). Chacun de ces locuteurs dispose de 125 phrases de test. Nous effectuons un test par phrase soit plus de 6000 tests.
En vérification de locuteur, nous allons déterminer le taux d'égale erreur *EER*, qui est

obtenu quand $f_a = f_r = EER$ (cf. paragraphe 1.4), pour un seuil commun à tous les locuteurs. Pour cela, nous effectuons plus de 6000 tests d'abonnés et autant de tests d'imposteurs (tirés aléatoirement du corpus $\mathscr{E}_4$).

3.5.2 Intervalle de confiance

La précision d'un résultat se mesure par la probabilité de se tromper. Nous appelons "intervalle de confiance à $x\%$", l'intervalle de la forme $[P-\epsilon, P+\epsilon]$, où P est le pourcentage du succès, dans lequel nous sommes sûr à $x\%$ que se trouve la bonne valeur du taux de reconnaissance. Si on considère que les succès dans un test de reconnaissance suivent une loi binomiale et que le nombre N de tests est important, les bornes de l'intervalle de confiance à $x\%$ sont :

$$P \pm z_x \sqrt{\frac{P(1-P)}{N}} \tag{3.2}$$

où la valeur de z_x se lit dans une table.
Le tableau 3.1 donne les intervalles de confiance pour différentes valeurs de taux d'identification incorrecte.

80%	70%	60%	50%	40%	30%	20%	10%
±1.01	±1.16	±1.24	±1.26	±1.24	±1.16	±1.01	±0.76

TAB. 3.1 – Valeurs des intervalles de confiance

Chapitre 4

Evaluation du système de reconnaissance par GMM

Cette partie a pour but d'évaluer les performances de reconnaissance par GMM, en mode indépendant du texte. Dans ce chapitre, nous présenterons l'impact des différents paramètres : ordre des modèles, activation de l'apprentissage de différents paramètres et la quantité de données d'apprentissage.

4.1 Protocole expérimental

Identification du locuteur (sur les données de France Télécom R&D)

Les étapes de l'évaluation de l'identification du locuteur sont les suivantes :
- Apprentissage : génération des 50 modèles de locuteur (de l'ensemble $\mathscr{E}_1$) par apprentissage incrémental (paragraphe 2.2.2).
- Test : identification du locuteur selon les modalités du paragraphe 2.3.2.

Vérification du locuteur

Les étapes de l'évaluation de la vérification du locuteur sont les suivantes :
- Apprentissage : génération des modèles de locuteur par apprentissage incrémental.
- Test : détermination a posteriori de l'EER avec un seuil commun à tous les locuteurs.

Notons que les résultats des expériences d'identification et de vérification constituent les résultats de base et de référence avec lesquels nous allons comparer les performances de la reconnaissance par localisation.

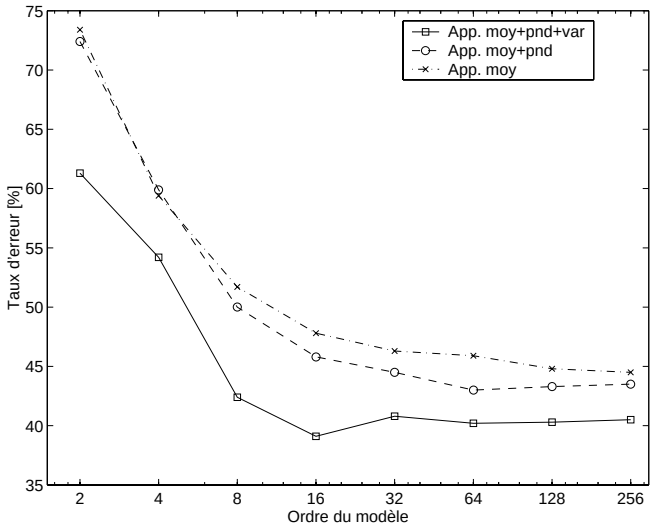


FIG. 4.1 – GMM : taux d'erreur d'identification en fonction de l'ordre des modèles (pour un apprentissage à partir de 4 secondes de parole)

4.2 Evaluation

4.2.1 Influence de l'ordre des modèles

Dans cette expérience, nous étudions l'impact de l'ordre des modèles (ou nombre de gaussiennes) sur les performances d'identification dans le cas où on a très peu de données d'apprentissage (4 secondes de parole). Nous allons aussi les comparer lorsque nous autorisons l'apprentissage des différents paramètres. La figure 4.1 trace les variations des taux d'erreur d'identification des 50 locuteurs de l'ensemble $\mathscr{E}_1$ en fonction de l'ordre des modèles et de l'activation de l'apprentissage des moyennes et/ou des variances et/ou des pondérations. Trois cas sont étudiés :

- Apprentissage de tous les paramètres des gaussiennes.
- Apprentissage des moyennes et des pondérations des gaussiennes, les variances étant inchangées.
- Apprentissage uniquement des moyennes des gaussiennes, les variances et les pondérations étant inchangées.

Les modèles GMM des locuteurs sont appris avec un apprentissage incrémental avec 4 secondes de parole.
La figure 4.1 montre que l'augmentation du nombre de gaussiennes dans la représentation du locuteur apporte une amélioration des performances. Cependant, le gain est peu significatif au-delà de 32 gaussiennes, et le temps de calcul augmente considérablement.
Le choix de l'ordre de modèle dépend de sa finesse et de la quantité de données d'apprentissage. Choisir un ordre trop peu élevé va nuire à la précision du modèle. Choisir trop

de composantes engendrera une charge de calcul plus importante. En général, 16 composantes suffisent pour représenter un locuteur disposant de très peu données d'apprentissage.

D'autre part, une autre possibilité d'action sur les paramètres du modèle est d'activer ou non l'apprentissage de la variance et des pondérations des gaussiennes. Si nous n'autorisons pas cet apprentissage, les valeurs des variances (ou des pondérations) du modèle d'initialisation sont affectées aux gaussiennes du modèle final. C'est-à-dire que les variances (ou les poids) de la $i$ème gaussienne du locuteur X auront la même valeur que les variances (ou les poids) de la $i$ème gaussienne du locuteur Y. Dans ce cas, les variances ne sont donc pas représentatives du locuteur, elles servent simplement de facteur d'échelle sur l'axe acoustique, pour donner à chacun des axes le même poids. Par contre, dans le cas où nous choisissons d'apprendre ces valeurs, chaque locuteur aura des variances différentes, et ces paramètres deviendront discriminants.
La figure 4.1 montre que les performances d'identification sont meilleures dans le cas où tous les paramètres sont appris. Il semble donc que chacun de ces paramètres apporte sa discrimination. Cependant, le modèle du locuteur sera plus complexe. A titre d'exemple, si l'espace acoustique est de dimension 42, à 256 gaussiennes le nombre de paramètres est de 21760 (42×256 moyennes + 42×256 variances + 256 pondérations).
En se basant sur ces résultats, pour la suite du travail, *les modèles de tous les locuteurs seront estimés en apprenant tous les paramètres des gaussiennes.*

4.2.2 Influence de la quantité d'apprentissage

Le but de cette expérience est d'évaluer les performances d'identification et d'authentification des 50 locuteurs en fonction de la quantité d'apprentissage. Les locuteurs sont modélisés par 256 gaussiennes et leurs modèles sont appris avec un apprentissage incrémental qui adapte tous les paramètres.
Sur les figures 4.2 et 4.3, nous avons représenté les variations des taux d'erreur ainsi que les taux EER des 50 locuteurs de l'ensemble $\mathscr{E}_1$ obtenus avec la modélisation GMM, en fonction de la quantité d'apprentissage. Ces figures donnent un aperçu des variations des taux d'erreur que l'on peut attendre en fonction de la quantité de données d'apprentissage. Les taux d'identification incorrecte décroissent significativement jusqu'à 40 secondes de données de parole. Au-delà de cette valeur, les performances d'identification ont tendance à saturer. Ces figures montrent aussi que, pour atteindre des taux d'erreur d'identification inférieurs à 10%, il faut disposer d'au moins d'une minute de données d'apprentissage. Quant à l'authentification, il suffit de 8 secondes de parole pour avoir un taux EER proche des 10%. Cependant, avec 100 secondes d'apprentissage, le système de reconnaissance par GMM atteint un taux d'erreur d'identification de 7.9% et un $EER = 4.3\%$.

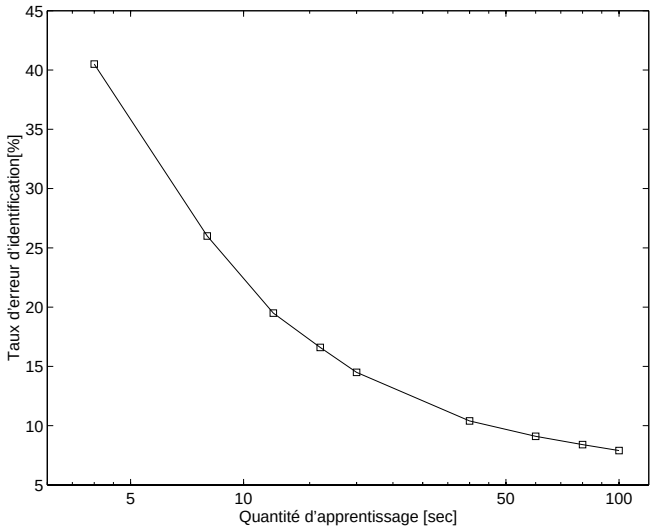


FIG. 4.2 – GMM : performances d'identification par GMM (modèles GMM à 256 gaussiennes)

4.3 Evaluations NIST de la vérification du locuteur

Le but de cette expérience est d'évaluer les performances de vérification sur les données de parole utilisées dans la campagne d'évaluations NIST 2000.
Dans cette expérience, les vecteurs acoustiques sont composés de 13 coefficients MFCC. A cela on rajoute leurs dérivées premières et secondes ; ce qui donne 42 coefficients acoustiques. On applique ensuite un module de soustraction cepstrale et un module de détection d'activité vocale.

En ce qui concerne les données NIST, leur modèle UBM est appris sur les données de parole utilisées dans la campagne d'évaluations NIST 1999. Parmi les données NIST 2000, nous considérons uniquement l'ensemble des locuteurs hommes / imposteurs hommes. La quantité de données d'apprentissage est environ 120 secondes de parole et les locuteurs sont modélisés avec 256 gaussiennes. Nous effectuons 2028 tests client et 19334 tests imposture. Le microphone utilisé est électrique en apprentissage et en test.
Pour les données de France Télécom, le modèle UBM est appris avec les données des locuteurs de l'ensemble $\mathscr{E}_3$. Afin de comparer les performances avec les données NIST, nous ne considérons que l'ensemble des locuteurs hommes / imposteurs hommes. Pour cela, nous utilisons 43 hommes (des ensembles $\mathscr{E}_1$ et $\mathscr{E}_2$).

La figure 4.4 trace les variations des taux de faux-rejet f_r en fonction des taux de

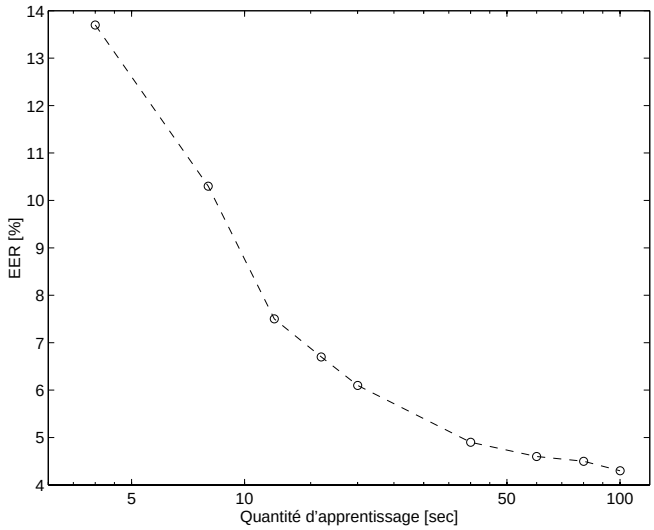


FIG. 4.3 – GMM : performances de vérification par GMM (modèles GMM à 256 gaussiennes)

fausse-acceptation f_a. Cette courbe DET[1] permet de représenter les performances d'un système de vérification de locuteur. Le taux d'égal erreur EER est obtenu quand $f_a = f_r$ et égale 13.6%. Dans des conditions à peu près similaires (100 secondes d'apprentissage), le taux d'erreur EER sur les données de France Télécom R&D est de 11.0%. La nature des données utilisées (parole spontanée dans la base de NIST et lue dans la base de France Télécom) peut expliquer ce léger écart de performances.

4.4 Conclusion

Dans ce chapitre, nous avons présenté les évaluations GMM sur la base de données de France Télécom et NIST. En ce qui concerne l'influence de l'ordre des modèles, l'augmentation du nombre de gaussiennes apporte une amélioration des performances peu significatives au-delà de 32 gaussiennes. Les expériences d'évaluation ont montré aussi que les performances d'identification sont meilleures dans le cas où tous les paramètres sont appris. Avec une modélisation à 256 gaussiennes et avec 4 secondes d'apprentissage, le système de reconnaissance par GMM atteint un taux d'erreur d'identification de 40.5% et un $EER = 13.7\%$.

[1] DET pour *Detection Error Trade-off*.

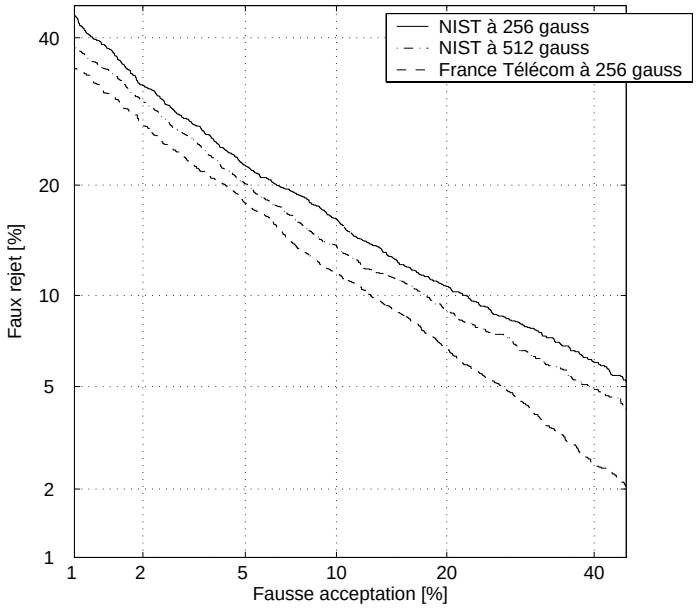


FIG. 4.4 – GMM : performances de vérification obtenues sur la base NIST et sur la base France Télécom

Deuxième partie

Reconnaissance du locuteur par localisation dans un espace de locuteurs de référence

Chapitre 5

Système de reconnaissance par placement dans un espace de locuteurs de référence

Dans la plupart des applications de reconnaissance de locuteurs, la phase d'enrôlement doit être très brève (de l'ordre de quelques secondes de parole). Il faut donc estimer avec très peu de données un modèle suffisamment robuste du locuteur pour permettre la reconnaissance du locuteur même lorsque les conditions de prise de son ou le contenu phonétique de la phrase de test ne sont pas les mêmes qu'à l'apprentissage. Nous avons vu que la modélisation par un mélange de gaussiennes (GMM), en mode indépendant du texte, fournit des bonnes performances et constitue l'état de l'art en la matière[1]. Malheureusement, cette modélisation n'est pas suffisamment robuste et la complexité des modèles reste importante. Pour tenter de remédier à ce problème, une perspective intéressante de modélisation consiste à représenter un locuteur, non plus de façon absolue, mais relativement à un ensemble de locuteurs dont les modèles sont bien appris. Chaque locuteur est représenté par sa localisation dans un espace de référence. Par cette démarche, on espère hériter des connaissances pour la modélisation qu'on ne pouvait pas estimer avec peu de données.
Dans ce chapitre, nous présenterons quelques repères bibliographiques sur la modélisation relative des locuteurs. Nous expliquerons ensuite le principe de reconnaissance de locuteurs par placement dans un espace de locuteurs de référence et nous introduirons les deux étapes de notre système de reconnaissance.

[1] Voir les évaluations des GMM au paragraphe 4.2.2

5.1 Etat de l'art de la représentation relative des locuteurs

Les techniques d'adaptation globale, telles que MAP (*Maximum A Posteriori*) ou MLLR (*Maximum Likelihood Linear Regression*), permettent d'obtenir un modèle qui a des performances similaires au modèle obtenu avec un apprentissage classique tout en nécessitant une quantité de données moins grande que celle utilisée pour un apprentissage classique. Cependant, cette quantité de données reste assez importante.
Il existe des techniques d'adaptation rapide, basées sur le principe de la représentation relative des locuteurs, qui ont été développées initialement dans le cadre de la reconnaissance automatique de la parole. Ces nouvelles approches ont donné naissance à la notion d'espace de locuteurs (*speaker space*) où un modèle de locuteur est représenté généralement par une combinaison linaire des modèles de référence ce qui réduit considérablement le nombre de paramètres. L'objet principal est de pouvoir hériter, à partir de cet espace représentatif, quelques connaissances pour la modélisation qu'on ne pouvait pas avoir avec le peu de données disponibles.
Dans les paragraphes suivants, nous allons dresser un état de l'art des techniques de représentation relatives utilisées en reconnaissance de parole et de locuteur.

5.1.1 Représentation relative en reconnaissance de la parole

Le principe de la représentation relative des locuteurs a été initialement appliqué en reconnaissance de parole dans des techniques d'adaptation rapide. Ces techniques reposent sur le principe d'utiliser des connaissances a priori obtenues à partir d'un ensemble de locuteurs de référence. Les principales techniques sont : RMP (*Regression-Based Model Prediction*), *Speaker Clustering*, RSW (*Reference Speaker Weighting*) et les voix propres (ou *eigenvoices*).

Modèle de prédiction basé sur la régression (RMP)

Cette technique d'adaptation, initialement utilisée en reconnaissance automatique de la parole, est une variante de la MAP [Ahadi-Sarkani, 1996]. Elle exploite les corrélations qui peuvent exister entre les paramètres des modèles. En effet, une phrase prononcée par différents locuteurs doit être reconnue comme étant la même phrase, et un locuteur doit être reconnu même s'il prononce différentes phrases. A cet effet, la régression linéaire est utilisée pour exprimer certains paramètres en fonction des autres. Par exemple, dans le cas où l'espace est de dimension 2, on peut déterminer une relation linéaire entre les deux paramètres d'un modèle :

$$\gamma_2 = b_1\gamma_1 + b_0 + \epsilon \tag{5.1}$$

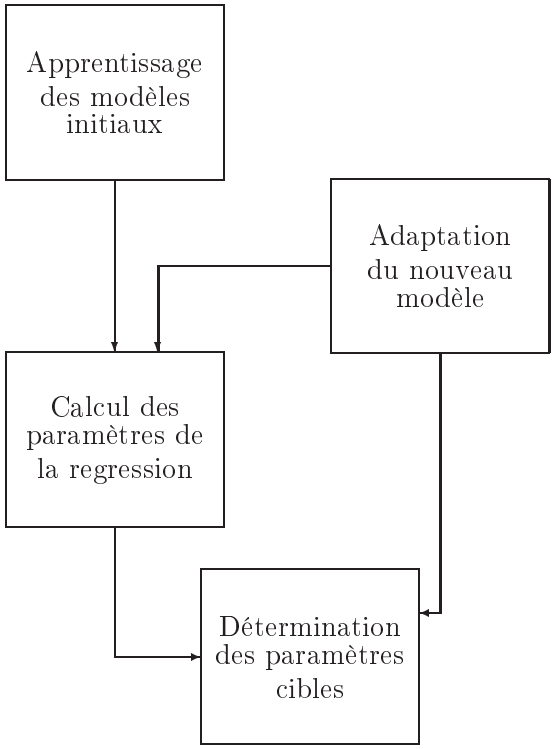


FIG. 5.1 – Adaptation par RMP (*Regression-Based Model Prediction*)

où b_1 et b_0 sont les paramètres de la régression et ϵ est l'erreur de modélisation donnée par :

$$\sum_{s=1}^{S} \epsilon_s^2 = \sum_{s=1}^{S} [\gamma_2^s - b_1\gamma_1^s - b_0]^2 \tag{5.2}$$

où S est le nombre total des locuteurs initiaux.
Dans ce cas, γ_2 est appelé un paramètre *cible* et γ_1 est appelé un paramètre *source*. Les paramètres cibles sont exprimés en fonction des paramètres sources.
Cette approche fonctionne en quatre étapes (figure 5.1). Une première étape, off-line, consiste à apprendre les modèles des S locuteurs initiaux. Ces modèles permettront de déterminer une combinaison linéaire entre leurs différents paramètres. Un grand nombre de locuteurs est nécessaire pour estimer, de manière fiable, les paramètres de la régression.
La deuxième étape consiste à adapter le modèle d'un nouveau locuteur (par exemple, par MAP). Durant l'adaptation, on retient la quantité de données qui a servi à apprendre chaque état du modèle. Cela permet de déterminer les paramètres cibles (ceux qui sont mal estimés) et les paramètres sources.
Dans la troisième étape, on détermine les paramètres de la régression à partir des modèles des S locuteurs et on exprime les paramètres cibles en fonction des paramètres sources. La

dernière étape consiste à apprendre le nouveau modèle du locuteur. Il s'agit d'adapter les valeurs des paramètres mal estimés à partir de ceux qui ont été bien estimés (par simple application de l'équation 5.1).
La RMP conserve les avantages de la MAP et permet, en plus, de bien apprendre les paramètres mal estimés par régression. Cependant, l'apprentissage par MAP converge plus rapidement, la phase de recherche des corrélations entre les paramètres est assez complexe et fastidieuse.
Il existe une autre variante étendue de la MAP appelée EMAP qui repose sur le principe d'exploitation des corrélations entre les paramètres des modèles. Elle suppose que les vecteurs des moyennes sont corrélés et réalisent une distribution normale $\mu \sim \mathcal{N}(\mu_0, \Sigma_0)$. Ensuite elle maximise la probabilité a posteriori du vecteur des moyennes μ par rapport aux observations [Rozzi, 1991] [Jom et al., 2001].

Pondération des locuteurs de référence (RSW)

La RSW [Hazen, 1998] est une technique très proche des voix propres. Chaque modèle de locuteur est représenté par une somme pondérée des modèles de référence. Dans l'approche des voix propres, ces derniers correspondent aux vecteurs des moyennes des locuteurs de référence. En RSW, ils correspondent à des vecteurs centroïdes (centre de masses des paramètres).
On suppose que chaque locuteur de référence s peut être représenté par M vecteurs centroïdes $\mathscr{C}_{m,s}$ où chacun d'eux correspond à une classe phonétique m ($m = 1, \cdots, M$). Ainsi pour chaque classe m, la matrice des centroïdes de tous les locuteurs de référence est donnée par :

$$\Gamma_m = [\mathscr{C}_{m,1}, \mathscr{C}_{m,2}, \ldots, \mathscr{C}_{m,S}] \tag{5.3}$$

où S est le nombre total des locuteurs de référence.
Durant la phase d'adaptation, le modèle d'un nouveau locuteur est représenté par une combinaison linéaire des vecteurs centroïdes et la moyenne est exprimée, pour une gaussienne m, par l'expression suivante :

$$\mu_m = \Gamma_m w \tag{5.4}$$

où w est le vecteur des poids du nouveau locuteur. Ce vecteur est déterminé par maximum de vraisemblance, soit :

$$\widehat{w} = \arg\max_{w} p(X|w) \tag{5.5}$$

Notons que la démarche est la même que celle de la MLED (cf. paragraphe 7.1.2).

Clustering des locuteurs

Le clustering repose sur le principe de création de plusieurs clusters de référence représentant l'ensemble des locuteurs. Ces clusters peuvent être déterminés avec des mesures de similarité sur les coefficients acoustiques des locuteurs ou sur leurs modèles (voir paragraphe 6.3) : chaque cluster contient des locuteurs similaires.

Durant la phase d'adaptation, le modèle d'un nouveau locuteur est représenté soit par une combinaison linéaire des modèles associés aux clusters finaux soit par sélection du modèle correspondant au plus proche cluster.
Il existe deux variantes principales du clustering : le clustering hiérarchique et la CAT (*Cluster Adaptative Training*). La première consiste à regrouper deux à deux les locuteurs les plus proches et à construire un arbre de classification des locuteurs (cf. paragraphe 6.3). Cette approche a été aussi combinée et utilisée avec la MLLR [Padmanabhan et Nahamoo, 1998] dans le cadre de la reconnaissance de parole à grand vocabulaire.
La CAT est une autre variante du clustering [Gales, 1998] : au lieu d'affecter un locuteur à un seul cluster, les paramètres de son modèle sont déterminés par une combinaison linéaire des paramètres de plusieurs clusters ([Gales, 1998] adapte uniquement les moyennes des modèles). La CAT est un algorithme d'adaptation itératif : pour un ensemble de S locuteurs initiaux, on détermine les E clusters de référence en deux étapes :

- D'abord, on calcule les vecteurs des poids (ou vecteur des coordonnées) de chaque locuteur.
- A partir de ces vecteurs, on détermine un nouvel ensemble de clusters de référence.

Ce processus est itéré jusqu'à satisfaction d'un critère de convergence.
Le calcul des vecteurs de poids ainsi que les clusters de référence passe par résolution d'un système d'équations linéaires (le principe de calcul est proche de celui de la MLED, cf. paragraphe 7.1.2). Par ailleurs, contrairement aux voix propres, l'approche de la CAT ne permet pas de garantir l'obtention d'un nombre E de locuteurs de référence.

Les voix propres

Cette approche [Kuhn et al., 2000] a été développé initialement dans le cadre de l'adaptation en locuteur. Elle s'inspire largement du concept des *eigenfaces*. Les voix propres sont générées par des algorithmes de réduction de dimensionnalité. A partir de la matrice des paramètres HMM des locuteurs, on applique ces algorithmes et on ne conserve que les axes à grande inertie (cf. paragraphes 6.1.2 et 6.1.4). Les voix propres sont aussi calculées par une méthode itérative appelée MLES (cf. paragraphe 6.2) qui consiste à rechercher un espace optimal en maximisant la vraisemblance des données.
Les locuteurs sont localisés par maximum de vraisemblance. Cette technique de localisation est très proche de celles présentées dans les paragraphes précédents. En identification du locuteur, on applique la distance euclidienne ou l'angle entre les vecteurs des coordonnées.

5.1.2 Représentation relative en reconnaissance du locuteur

En reconnaissance du locuteur, la représentation relative a été récemment appliquée. Dans les paragraphes suivants, nous présentons les principales techniques proposées.

"Non Directly Acoustic Process"

Cette technique consiste à caractériser un locuteur par rapport à un ensemble de locuteurs dont les modèles sont bien appris [Merlin et al., 1999]. Les trames acoustiques des locuteurs sont "projetées" dans un nouvel espace de représentation. La technique de projection consiste à évaluer un score de vraisemblance entre une portion du signal et l'ensemble des locuteurs de référence. Dans [Merlin et al., 1999], les locuteurs de référence sont tirés aléatoirement.
Chaque locuteur correspond à un ensemble de points dans le nouvel espace. Son modèle est appris par un algorithme de classification tel que les *K-means* et représenté par deux vecteurs : le vecteur centre de gravité et le vecteur des variances. Dans la phase de test, on mesure l'angle entre le vecteur du signal de test et les centres de gravité des modèles de locuteurs à identifier.

Les modèles d'ancrage

Il s'agit de représenter et de caractériser un locuteur par rapport à un ensemble de modèles de locuteurs bien appris appelés modèles d'ancrage (ou *Anchor Models*). Cette technique est largement inspirée de l'approche présentée dans le paragraphe précédent. Elle peut être utilisée en reconnaissance du locuteur, en détection du locuteur ou encore en regroupement de locuteurs. Dans [Sturim et al., 2001], les modèles d'ancrage sont utilisés essentiellement en indexation en locuteurs.
Pour caractériser un locuteur, on évalue un score de vraisemblance entre les données du locuteur et chaque modèle de référence A_i (modèles GMM-UBM). Le locuteur λ est donc représenté par le vecteur V suivant :

$$V = \begin{bmatrix} p(x|A_1) \\ p(x|A_2) \\ \vdots \\ p(x|A_N) \end{bmatrix} \tag{5.6}$$

La phase de test consiste à appliquer la distance euclidienne entre les vecteurs caractéristiques d'un locuteur cible et d'un locuteur de test.

Les voix propres

Les voix propres ont été appliquées en reconnaissance automatique du locuteur (notamment par [Thyes et al., 2000]). Un locuteur est modélisé par ses coefficients de combinaison des voix propres. On mesure ensuite la similarité entre les locuteurs par une distance entre les coefficients.

5.2 Principe de reconnaissance par placement dans un espace de locuteurs de référence

Notre travail s'inscrit dans le domaine de la représentation relative en reconnaissance du locuteur. Comme nous l'avons déjà vu dans les paragraphes précédents, ces systèmes de représentation et de modélisation des locuteurs exploitent la position d'un locuteur par rapport à un ensemble de locuteurs de référence. Dans notre étude, nous nous baserons sur le même principe et nous l'appliquerons en reconnaissance automatique du locuteur. Nous rechercherons le meilleur espace de représentation et la meilleure manière de localiser un locuteur dans cet espace.
Notre système de reconnaissance se déroule en deux phases : la première phase est dédiée à la construction d'un espace de représentation tandis que dans la deuxième, on localise des nouveaux locuteurs dans cet espace et on évalue la proximité spatiale entre eux. Selon la technique de construction de l'espace et de localisation, plusieurs variantes peuvent être explorées.

La représentation des locuteurs par localisation est une nouvelle technique de reconnaissance de locuteur. L'une des premières voies explorées est celle des voix propres. Cette nouvelle approche peut être formulée par la question suivante : parmi un ensemble de S locuteurs, peut-on sélectionner ou fabriquer un certain nombre de locuteurs (ou de voix) dont chacun reflète une caractéristique de la voix humaine ?
L'idéal serait d'avoir un espace orthogonal où chacun des axes évalue et détermine le sexe du locuteur, son âge, l'intensité de la voix, sa qualité, le débit et le rythme de la parole, les variations mélodiques, etc. Chaque locuteur peut être représenté dans cet espace et son modèle λ approximé par la relation :

$$\lambda \approx \sum_{e=1}^{E} w_e \overline{\lambda}_e \tag{5.7}$$

où $\overline{\lambda}_e$ représentent les vecteurs propres de l'espace représentatif ou les voix propres, E est la dimension de l'espace c'est-à-dire le nombre de locuteurs ou de voix propres.
Ainsi, on associe à chaque locuteur λ un vecteur caractéristique w :

$$w = \{w_e\}_{e=1,\ldots,E}$$

En plaçant plusieurs locuteurs dans l'espace, on peut évaluer la proximité spatiale et dire que tel ou tel locuteur est proche ou loin d'un autre. Il s'agit de représenter tout simplement un locuteur relativement à d'autres locuteurs.
En effet, la motivation principale de cette approche repose sur le fait que la dimension du problème (le nombre de paramètres) est très grande par rapport à celle qu'on peut estimer de façon fiable. Ainsi, plutôt que d'estimer les nombreux paramètres d'un modèle absolu du locuteur, on cherche à estimer des paramètres moins nombreux d'un modèle relatif à

d'autres modèles de locuteurs, ces derniers étant estimés de façon absolue avec suffisamment de données. Par conséquent, *on ne modélise plus de façon absolue mais relativement à des locuteurs de référence.*

5.3 Composantes du système

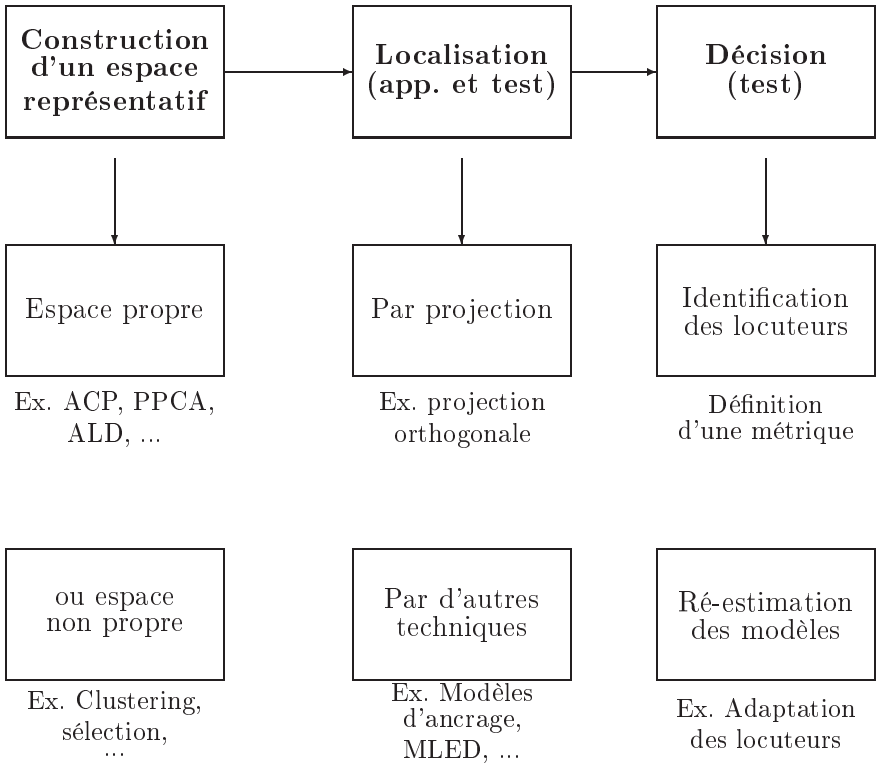


FIG. 5.2 – Système de reconnaissance par placement dans un espace de référence

L'équation 5.7 traduit la nouvelle représentation des locuteurs et distingue clairement deux principales tâches pour le système. D'abord, la recherche des $\overline{\lambda}_e$ c'est-à-dire la construction d'un espace représentatif et ensuite, la détermination des w_e c'est-à-dire la localisation dans cet espace. Ainsi le système d'identification de locuteurs par localisation se déroule en deux phases (figure 5.2) :

Construction de l'espace représentatif des locuteurs

Cette étape consiste à créer une base de locuteurs de référence (ou un nouvel espace de représentation) : on dispose des modèles GMM d'un ensemble de locuteurs bien appris qui serviront à la construction de l'espace représentatif. Plusieurs techniques peuvent être

utilisées pour la réduction et la construction de l'espace des locuteurs. Dans ce travail, nous avons exploré trois voies :

- on construit un espace propre (voix propres ou orthogonales) ;
- on réalise un regroupement hiérarchique des locuteurs ;
- on sélectionne les plus dispersés d'entre eux selon un critère de sélection donné.

Les voix propres s'obtiennent en appliquant des algorithmes de réduction de dimensionnalité de l'espace, telle que l'ACP (Analyse en Composantes Principales), la PPCA (Probabilistic ACP) ou la ALD (Analyse Linéaire Discriminante). Il s'agit de réajuster le nuage de points de l'espace et, dans le cas de l'ACP, rechercher les axes principaux à grande inertie ou rechercher les axes les plus discriminants dans le cas de l'ALD. Le regroupement hiérarchique et la sélection d'un sous-ensemble constituent une alternative aux voix propres. Le principe du regroupement repose sur l'agrégation deux à deux les données des locuteurs les plus proches. Quant à la sélection, on recherche un sous-groupe de locuteurs supposés le plus représentatif de l'ensemble de locuteurs.
Notons finalement que les locuteurs qui serviront à la construction de l'espace sont différents des locuteurs à reconnaître et que cette phase de construction se fait une fois pour toute.
Toutes ces techniques seront présentées et discutées dans le chapitre suivant (chapitre 6).

Localisation et décision

La deuxième étape du système s'applique à tout nouveau locuteur (d'apprentissage ou de test) et se décompose en deux étages :

Localisation : elle consiste à localiser des nouveaux locuteurs dans cet espace de représentation. Chaque locuteur est associé à un vecteur de caractéristiques propres $w = [w_1, \ldots, w_E]'$, qui représentent ses coordonnées dans l'espace des locuteurs de référence. La localisation est réalisée soit par une simple projection (qui n'a de sens que si l'espace est orthogonal), soit en maximisant la vraisemblance du problème (MLED) ou encore en utilisant la technique des modèles d'ancrage. Le principe de ces derniers consiste à représenter et caractériser un signal de parole d'un locuteur en estimant sa vraisemblance par rapport à un ensemble de modèles de locuteurs (ou modèles d'ancrage).

Décision : dans cet étage, on exploite les coefficients caractéristiques des locuteurs w_e ainsi calculés. Ils peuvent servir à faire de l'identification de locuteurs comme ils peuvent aussi donner lieu à une meilleure estimation des modèles et faire de l'adaptation des locuteurs.

L'identification de locuteurs est l'application d'une simple distance. En effet, la représentation intuitive d'un locuteur par sa localisation dans l'espace de représentation présume que plus des locuteurs sont similaires plus leurs points de projection sont proches et la distance entre eux est petite. Donc, pour exploiter la notion du voisinage et évaluer la proximité dans l'espace, on utilise une métrique entre les coordonnées des locuteurs d'apprentissage et de test.

Soit un locuteur inconnu X représenté par $\lambda_X = \{w_e^X\}_{e=1,...,E}$, le locuteur reconnu, $\hat{R}$, est celui dont le modèle $\lambda_R = \{w_e^R\}_{e=1,...,E}$ donne la plus petite distance :

$$\hat{R} = \arg\min_R d(\lambda_X, \lambda_R)$$

Les techniques de localisation des locuteurs ainsi que les métriques utilisées pour l'identification feront l'objet du chapitre 7.

5.4 Conclusion

Dans ce chapitre, nous avons introduit le principe de reconnaissance par placement dans un espace de locuteurs de référence. Nous avons commencé par dresser un état de l'art de la représentation relative.
Les techniques d'adaptation rapide sont présentées comme une alternative aux approches d'adaptation classiques dans le cas où nous disposons de peu de données d'apprentissage. La RMP est une technique d'adaptation qui exploite les corrélations existant entre les paramètres des modèles. Elle conserve les avantages de la MAP et permet, en plus, de bien apprendre les paramètres mal estimés par régression. Quant aux autres techniques de représentation relative, chaque modèle de locuteur est représenté par une somme pondérée des modèles de référence. En RSW, ces derniers correspondent à des vecteurs centroïdes. Dans l'approche des voix propres, ils correspondent aux vecteurs des moyennes des locuteurs de référence. Dans le clustering, les modèles de références sont des clusters déterminés avec des mesures de similarité sur les coefficients acoustiques des locuteurs ou sur leurs modèles. La représentation relative a été aussi appliquée en reconnaissance automatique du locuteur. Les locuteurs de référence sont soit tirés aléatoirement soit correspondent aux voix propres. Ces dernières sont généralement obtenues par les méthodes de réduction de dimensionnalité ou par maximum de vraisemblance. La localisation des locuteurs dans l'espace de référence consiste généralement à évaluer un score de vraisemblance par rapport à chaque axe de l'espace. Dans le cas des voix propres, la localisation est réalisée par MLED, qui est similaire à des techniques de localisation utilisées dans le cadre de la représentation relative en reconnaissance de parole.
Notre système de reconnaissance exploite le positionnement relatif en reconnaissance de locuteur. Il se déroule en deux phases : la première phase est dédiée à la construction d'un espace de représentation tandis que la deuxième est consacrée à la localisation des locuteurs dans cet espace. Selon la technique de construction de l'espace et de localisation, plusieurs variantes peuvent être explorées.

Chapitre 6

Construction de l'espace représentatif

La construction d'un espace consiste à rechercher un ensemble de locuteurs les plus représentatifs. Les méthodes et les approches à explorer sont multiples mais on peut les classer en quatre grandes familles :

Construction de l'espace par les méthodes d'analyse de données : à partir des paramètres GMM des locuteurs, on applique un algorithme de réduction de dimensionnalité. Cette approche permet d'obtenir un espace orthogonal et a été déjà utilisée par [Nguyen et al., 1999] pour construire des voix propres.

Construction de l'espace par maximum de vraisemblance : il s'agit de rechercher les voix propres itérativement afin de maximiser la vraisemblance des données.

Construction de l'espace par regroupement hiérarchique ascendant : c'est une méthode originale [Mami et Charlet, 2002a] qui utilise les scores GMM pour fusionner les voix les plus proches et construire ainsi un arbre de locuteurs.

Construction de l'espace par sélection : cette approche également originale recherche, selon un critère donné, un sous-ensemble de locuteurs qui est supposé le plus représentatif de l'ensemble des locuteurs [Mami et Charlet, 2002b].

Dans les paragraphes suivants, on présentera ces techniques et on expliquera comment on peut les utiliser pour construire les locuteurs de référence.

6.1 Construction d'un espace propre par les méthodes d'analyse de données

La construction d'un espace propre se fait, en général, par les méthodes d'analyse de données qui recouvrent un grand nombre de méthodes qui ont pour objectif de décrire, synthétiser, expliquer l'information contenue dans de vastes tableaux de données. Parmi elles, on retrouve les méthodes factorielles qui se proposent de figurer géométriquement

dans un espace euclidien de faible dimension les informations les plus diverses. Il existe de nombreuses techniques d'analyse factorielle parmi lesquelles, l'analyse en composantes principales (notée ACP), l'analyse des correspondances (simple ou multiple) (notée AFC), l'analyse discriminante et l'analyse canonique.

6.1.1 Notations

On suppose qu'on dispose de S locuteurs, chaque locuteur est modélisé par M gaussiennes dans un espace acoustique de dimension D.
Pour appliquer les méthodes d'analyse de données, on se propose de travailler sur un tableau de données R à $n = S$ lignes (individus ou locuteurs)[1] décrits par un ensemble de $p = M \times D$ colonnes (paramètres GMM des locuteurs) ; de terme général r_{ij}. L'application des méthodes factorielles sur ce tableau permet d'obtenir un espace orthogonal Y de faible dimension $E \times (M \times D)$ où les nouveaux paramètres des locuteurs seront décorrélés dans un espace à E voix propres.

$$R \rightarrow Y$$

Dans le cas où on ne considère que les moyennes des modèles, chaque ligne i représente les $p = M \times D$ moyennes GMM du locuteur i et les matrices, avant et après transformation de l'espace, s'écrivent :

$$R = \begin{pmatrix} \mu_1(1) & \mu_2(1) & \cdots & \mu_{M\times D}(1) \\ \mu_1(2) & \mu_2(2) & \cdots & \mu_{M\times D}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1(S) & \mu_2(S) & \ldots & \mu_{M\times D}(S) \end{pmatrix} \rightarrow Y = \begin{pmatrix} \overline{\mu}_1(1) & \overline{\mu}_2(1) & \cdots & \overline{\mu}_{M\times D}(1) \\ \overline{\mu}_1(2) & \overline{\mu}_2(2) & \cdots & \overline{\mu}_{M\times D}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\mu}_1(E) & \overline{\mu}_2(E) & \ldots & \overline{\mu}_{M\times D}(E) \end{pmatrix}$$

où E est le nombre des locuteurs propres qui est inférieur au nombre initial des locuteurs S.

6.1.2 Construction de l'espace par l'analyse en composantes principales (ACP)

La réduction de l'espace des paramètres par un algorithme de réduction de dimensionnalité constitue une des approches les plus simples et les plus intuitives. On recherche des sous-espaces de faible dimension qui ajustent au mieux le nuage de points-individus et celui des points-variables, de façon à ce que les proximités mesurées dans ces sous-espaces reflètent autant que possible les proximités réelles.
L'analyse en composantes principales (ACP)[2] (par exemple [Jollife, 1986]) répond à ce

[1] Plus exactement, S est le nombre d'observations car on peut avoir plusieurs réalisations pour un même locuteur.

[2] En anglais PCA pour *Principal Component Analysis*

type de problème. Elle s'applique aux tableaux de type "variables-individus". Les proximités entre variables s'interprètent en terme de corrélations ; les proximités entre individus s'interprètent en terme de similitudes globales des valeurs observées.
Dans notre étude, nous pouvons appliquer l'ACP sur le tableau de données R de deux façons différentes :

- Si nous considérons que les variables sont les paramètres des modèles GMM et les individus sont les locuteurs, nous travaillerons sur un tableau de dimension $S \times (M \times D)$ et la réduction se fera sur le nombre de paramètres GMM : il s'agit tout simplement d'effectuer une orthogonalisation des GMM et non pas de construire un espace propre.
- Si nous considérons que les variables sont les locuteurs et les individus sont les paramètres de leurs modèles GMM, nous travaillerons sur un tableau de dimension $(M \times D) \times S$ et la réduction se fera, cette fois ci, sur le nombre de locuteurs. L'espace résultant est constitué d'un nombre réduit de locuteurs propres décrits par le même nombre de paramètres.

Dans cette partie, c'est le deuxième cas qui nous intéresse.
Si on suppose que les valeurs d'un paramètre GMM pour différents locuteurs sont hétérogènes du point de vue de leurs moyennes et de leurs dispersions, on les ramène à un espace de comparabilité commun. Pour ce faire, on analyse le tableau dit centré plutôt que le tableau des données d'origine R (de terme général r_{ij}).
Par $\overline{r_i}$ on désigne la moyenne de l'individu (ou locuteur) i donnée par :

$$\overline{r_i} = \frac{1}{p} \sum_{j=1}^{p} r_{ij} \tag{6.1}$$

où $p = M \times D$ est le nombre des paramètres GMM des locuteurs.
Dans le but de faire jouer à chaque variable i un rôle identique, on effectue une translation de l'origine au centre de gravité du nuage et on change les échelles sur les différents axes. On parle de l'analyse en composantes principales normées et ça se traduit par la diagonalisation de la matrice de covariance C de terme général $c_{ii'}$ donné par :

$$c_{ii'} = \frac{1}{p} \sum_{j=1}^{p} (r_{ij} - \overline{r}_i)(r_{i'j} - \overline{r}_{i'}) \tag{6.2}$$

La matrice C est symétrique ; ses valeurs propres $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ sont positives ou nulles et ses vecteurs propres normés ordonnés forment une base orthonormée. Notons $u_1, u_2, \ldots, u_n$ les vecteurs propres ordonnées suivant les valeurs propres décroissantes et rangées dans la matrice U.
Les composantes principales recherchées représentent les colonnes de la matrice Y :

$$Y = RU\Lambda^{-1/2} \tag{6.3}$$

où Λ est la matrice diagonale des valeurs propres de la matrice de covariance C.
La matrice Y correspond à l'espace propre recherché. Chaque ligne est une voix propre

modélisée par $M \times D$ paramètres décorrélés. Cette matrice est de dimension $E \times (M \times D)$[3] où E correspond à la dimension de l'espace. Généralement, on ne retient que les composantes à inertie $\geq 1/n$. La qualité globale de représentation est le rapport entre l'inertie de l'espace réduit et celle de l'espace d'origine soit :

$$\frac{\lambda_1 + \cdots + \lambda_E}{\lambda_1 + \cdots + \lambda_n}$$

où $n = S$ représente le nombre de locuteurs initiaux.

6.1.3 Construction de l'espace par l'ACP probabiliste (PPCA)

Bien que l'ACP soit une technique très utilisée dans le domaine de l'analyse des données, elle présente l'inconvénient de se baser uniquement sur une approche géométrique très étroite. En effet, son rôle consiste simplement à réajuster le nuage des points ou de réordonner les axes de l'espace de façon à ne donner de l'importance qu'aux axes de grandes variances.
L'introduction d'un modèle probabiliste sur l'ACP permet de modéliser les données par une ou plusieurs gaussiennes [Tipping et Bishop, 1997]. Par conséquent, on peut estimer leur vraisemblance et appliquer les différentes méthodes d'inférence bayésienne.
L'application de la PPCA[4] pour construire un espace propre représentatif des locuteurs est une idée originale.
Soit r_i le vecteur des p variables ou paramètres du locuteur i et $i = 1, \ldots, S$. On sait par ailleurs qu'à partir des vecteurs des composantes principales y_i et des vecteurs d'axes principaux u_i, on peut reconstruire d'une manière approximative le vecteur r_i par la relation :

$$r_i \approx \mathscr{U}\, y_i \tag{6.4}$$

Par analogie avec l'ACP, $\mathscr{U}$ correspond à la matrice des vecteurs propres U. Le but de la PPCA est d'estimer la matrice $\mathscr{U}$ par maximum de vraisemblance (au lieu qu'elle corresponde simplement aux vecteurs propres de la matrice de covariance). On doit aussi montrer que $\mathscr{U}$ converge vers U sous certaines conditions.
Dans l'approche de la PPCA, on considère que le terme $\mathscr{U}\, y_i$ est la vraie valeur inconnue de r_i et on écrit :

$$r_i = \mathscr{U}\, y_i + \epsilon_i \tag{6.5}$$

où ϵ_i désigne l'erreur affectant la reconstruction du $i$ème vecteur de données. Il s'agit là d'un cas particulier élémentaire de modèle linéaire : les données sont supposées être la somme d'un terme dépendant linéairement du paramètre inconnu ($\mathscr{U}$) et d'une erreur de

[3] On ne retient que les E premières lignes au lieu de garder S lignes (E est $\leq$ au rang de la matrice initiale R).
[4] Pour *Probabilistic Principal Components Analysis*

modélisation. Dans ce modèle, on suppose que y_i est une variable latente qui suit une distribution normale $y \sim \mathcal{N}(0, I)$, c'est-à-dire :

$$p(y_i) = (2\pi)^{-n/2} \exp\left\{-\frac{1}{2} y_i' y_i\right\} \tag{6.6}$$

et que le bruit est un processus gaussien centré de variance σ^2 :

$$p(r_i|y_i) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(r_i - Uy_i)'(r_i - Uy_i)\right\} \tag{6.7}$$

La matrice $\mathscr{U}$ est déterminée en maximisant le log de la vraisemblance marginale $\mathscr{L}$, soit :

$$\mathscr{L} = \sum_{i=1}^{n} \log p(r_i) \tag{6.8}$$

avec

$$\begin{aligned} p(r_i) &= \int p(r_i|y)p(y)dy \\ &= (2\pi)^{-n/2}|\mathscr{C}|^{-1/2} \exp\left\{-\frac{1}{2}(r_i - \mu)'\mathscr{C}^{-1}(r_i - \mu)\right\} \end{aligned} \tag{6.9}$$

où $\mathscr{C}$ est la nouvelle matrice de covariance donnée par :

$$\mathscr{C} = \sigma^2 I + \mathscr{U}\mathscr{U}' \tag{6.10}$$

En remplaçant toutes ces expressions dans 6.8 et en considérant que l'estimateur de $\mathscr{U}$ correspond à :

$$\frac{\delta \mathscr{L}}{\delta \mathscr{U}} = 0,$$

l'estimateur $\mathscr{U}$ est donnée par (après simplification) :

$$\mathscr{U} = U_E(\Lambda_E - \sigma^2 I)^{1/2} \tag{6.11}$$

où U_E et Λ_E sont, respectivement, la matrice des E premiers vecteurs propres et la matrice diagonale des E premières valeurs propres de la matrice de covariance C donnée par l'équation 6.2.
Par ailleurs les paramètres μ et σ^2 sont estimés par maximum de vraisemblance et donnés par :

$$\begin{aligned} \mu_{ML} &= \frac{1}{n}\sum_{i=1}^{n} r_{ij} \\ \sigma^2_{ML} &= \frac{1}{n-E}\sum_{i=E+1}^{n} \lambda_i \end{aligned} \tag{6.12}$$

Comme les valeurs propres de la matrice de covariance C sont classées en ordre décroissant, les $\lambda_{E+1}, \cdots, \lambda_d$ représentent les plus petites valeurs propres et mesurent la perte d'inertie quand on passe de n à E axes.
Ainsi, les voix propres sont obtenues par l'équation matricielle suivante :

$$Y = R\mathscr{U} \tag{6.13}$$

L'approche de la PPCA peut être étandue au cas multigaussien et la maximisation de la vraisemblance $\mathscr{L}$ est réalisée avec l'algorithme EM.

6.1.4 Construction de l'espace par l'analyse linéaire discriminante (ALD)

Comme dans les paragraphes précédents, on dispose d'un tableau de données R à n lignes (observations) et p colonnes (paramètres GMM). On considère que les n observations sont partitionnées en q classes de locuteurs C_k, $(k = 1, \cdots, q)$ connues a priori. On définit une classe par locuteur et donc q représente aussi le nombre de locuteurs.
Il s'agit de définir des fonctions permettant de séparer aux mieux ces classes de locuteurs. En effet, l'analyse en composantes principales permet de définir les directions de variabilité principale mais qui ne correspondent pas nécessairement aux directions de meilleure discrimination. L'analyse linéaire discriminante (ou LDA[5]) répond à ce type de problème. Elle regroupe une famille de techniques destinées à classer (affecter à des classes *préexistantes*) des individus caractérisés par un certain nombre de variables numériques ou nominales. Cette analyse s'appuie sur des critères de séparabilité entre classes (généralement les critères de *Fisher*).
Pour cela, la matrice de covariance totale des données est décomposée à l'aide du théorème de *Huygens* en deux matrices différentes, l'une donnant la variabilité dans chacune des q classes et l'autre la variabilité entre les q classes. Le critère couramment utilisé consiste à rechercher, dans l'espace des paramètres des locuteurs, les axes d'inertie qui ont la plus grande variance inter-classe tout en ayant une variance intra-classe constante (généralement égale à l'unité). Il s'agit de maximiser la quantité $J = tr(W^{-1}B)$ [Fukunaga, 1990] où W est la matrice de covariance intra-classe et B est celle de covariance inter-classe. Leurs termes généraux sont donnés par les expressions suivantes [Lebart et al., 2000] :

$$W_{jj'} = \frac{1}{n} \sum_{k=1}^{q} \sum_{i \in I_k} (r_{ij} - \overline{r}_{kj})(r_{ij'} - \overline{r}_{kj'}) \tag{6.14}$$

$$B_{jj'} = \sum_{k=1}^{q} \frac{n_k}{n} (\overline{r}_{kj} - \overline{r}_j)(\overline{r}_{kj'} - \overline{r}'_j) \tag{6.15}$$

[5]Pour *Lineair Discriminant Analysis*

où chaque classe de locuteur C_k caractérise un sous-nuage I_k, de n_k individus (nombre d'observations du locuteur k), de centre de gravité :

$$\overline{r}_{kj} = \frac{1}{n_k} \sum_{i \in I_k} r_{ij}$$

et leurs moyennes sur l'ensemble des individus correspondent au centre de gravité du nuage global :

$$\overline{r}_j = \frac{1}{n} \sum_{i=1}^{n} r_{ij}$$

La maximisation du critère J se traduit géométriquement par une double rotation. La première transforme la matrice intra-classe W en la matrice identité et le tableau des données s'écrit dans le nouvel espace comme :

$$R \longrightarrow R U_W \Lambda_W^{-1/2} \tag{6.16}$$

où U_W et Λ_W sont respectivement la matrice des vecteurs propres et la matrice diagonale des valeurs propres de W.
La deuxième rotation permet de retrouver les axes de projection qui ont la plus grande inertie inter-classe, tout en gardant la variance intra-classe constante. Cela veut dire qu'on diagonalise la matrice de covariance inter-classe B après l'avoir estimée dans le nouvel espace issu de la diagonalisation et du blanchiment de la matrice W :

$$R \longrightarrow R U_B \tag{6.17}$$

Finalement, la transformation finale est donnée par :

$$Y = R U_W \Lambda_W^{-1/2} U_B \tag{6.18}$$

Il s'agit ainsi d'une analyse en composantes principales (où l'on intègre une opération qui donne la même inertie intra-classe à chaque axe), suivie d'une rotation supplémentaire pour trouver les axes de projection qui ont la plus grande variance inter-classe [Charlet, 1997].
La matrice Y correspond donc aux E voix propres recherchées. Elle est de dimension $E \times p$ dont chacune est modélisée par $p = M \times D$ paramètres. Notons que E est $\leq q$ car il n'existe que q valeurs propres non nulles.
Par ailleurs, il est possible d'estimer les matrices d'inter- et d'intra-classe avec le tableau initial R et de faire la transformation sur d'autres données T. L'espace engendré est donné par :

$$Y_T = T U_W \Lambda_W^{-1/2} U_B \tag{6.19}$$

Cela permet de classer des observations à partir des classes de locuteurs crées d'un autre corpus de données (voir paragraphe 7.2.3).
Notons finalement que cette approche n'a jamais été proposée ou utilisée pour la construction d'un espace de locuteurs de référence.

6.2 Espace propre à maximum de vraisemblance (MLES)

La MLES (*Maximum-Likelihood EigenSpace*) est une méthode qui consiste à rechercher un espace optimal en maximisant la vraisemblance des données [Nguyen et al., 1999]. Cependant, cette approche fait intervenir des connaissances a priori sur la base de données utilisée qu'on ne peut pas toujours formuler. Rappelons qu'on ne considère que les moyennes des modèles GMM des locuteurs c'est-à-dire que le tableau initial des paramètres ne contient que les moyennes GMM des locuteurs : chaque ligne i représentera les $(D \times M)$ moyennes GMM (désignées par μ) du locuteur i où M est le nombre de gaussiennes du modèle GMM et D est la dimension de l'espace acoustique, soit :

$$R = \begin{pmatrix} \mu_1(1) & \mu_2(1) & \cdots & \mu_{M\times D}(1) \\ \mu_1(2) & \mu_2(2) & \cdots & \mu_{M\times D}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1(S) & \mu_2(S) & \dots & \mu_{M\times D}(S) \end{pmatrix}$$

Par conséquent, l'espace résultant est constitué des nouvelles moyennes recherchées désignées par $\overline{\mu}$:

$$Y = \begin{pmatrix} \overline{\mu}_1(1) & \overline{\mu}_2(1) & \cdots & \overline{\mu}_{D\times M}(1) \\ \overline{\mu}_1(2) & \overline{\mu}_2(2) & \cdots & \overline{\mu}_{D\times M}(2) \\ \vdots & \vdots & \ddots & \vdots \\ \overline{\mu}_1(E) & \overline{\mu}_2(E) & \dots & \overline{\mu}_{D\times M}(E) \end{pmatrix}$$

où Y est l'espace propre à optimiser.
Vu la nature cachée des paramètres des modèles de Markov ainsi que l'expression de la vraisemblance $\mathscr{L}(X|Y)$, la maximisation se fait via l'algorithme EM, soit :

$$\hat{Y} = \arg\max_{Y} \mathscr{L}(X|Y)$$

où $\hat{Y}$ est l'estimé de l'espace Y et X sont l'ensemble des observations.
On définit la fonction auxiliaire $Q(Y, Y^{(t)})$ en fonction des paramètres de l'espace Y et une estimation courante $Y^{(t)}$ (à l'itération t) de ceux-ci, les variables observées X et des variables cachées ζ :

$$Q(Y, Y^{(t)}) = \sum_{\zeta} P(\zeta|X, Y^{(t)}) \log P(X, \zeta|Y)$$

Cette fonction représente tout simplement l'espérance mathématique du logarithme de la vraisemblance jointe $P(X, \zeta|Y)$ (incluant les variables observées X et cachées ζ) sur l'ensemble complet des variables d'entraînement, calculées sur la base des paramètres courants.
On peut considérer que la variable s (indice des locuteurs) ainsi que les coefficients de

localisations des locuteurs w sont des variables cachées. En utilisant les formulations de l'algorithme EM, l'espace optimal correspond à l'expression suivante :

$$\hat{Y} = Y^{(t+1)} = \arg\max_{Y} \sum_{s=1}^{S} \int p(w,s) \log p(X,w|Y) dw \tag{6.20}$$

$p(w,s)$ est une fonction de pondération qui porte l'information a priori sur le locuteur s (c'est-à-dire la probabilité d'observer une personne de tel ou tel sexe, dialecte, niveau d'études, ..., etc). Supposons pour la simplicité de l'exposé que $p(w,s) = p(s) \prod_{e=1}^{E} p(w_e|s)$. Un exemple d'une telle fonction serait [Nguyen et al., 1999] :

$$p(w_e|s) = \begin{cases} 1 \text{ si } w_e > 0 \text{ et le locuteur } s \text{ est masculin} \\ 1 \text{ si } w_e < 0 \text{ et le locuteur } s \text{ est féminin} \\ 0 \text{ ailleurs} \end{cases}$$

Comme l'algorithme EM est un algorithme itératif, les voix propres calculées par ACP ou ALD peuvent servir à des valeurs d'initialisation. Les formules de ré-estimation des voix propres sont obtenues classiquement par dérivation de la fonction auxiliaire, soit :

$$\overline{\mu}_m(e) = \frac{\sum_{s=1}^{S} \mathscr{L}^s w_s(e) \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \{x_n^s - \mu_m^s(e)\}}{\sum_{s=1}^{S} \mathscr{L}^s w_s^2(e) \sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{6.21}$$

où $\gamma_{n,m}^{(t)}$ est une probabilité a posteriori estimée à l'itération t (cf. équation 2.7). N représente le nombre total des vecteurs acoustiques, m une distribution dans un modèle de mélange de M gaussiennes, e une voix propre et $\mathscr{L}^s = \mathscr{L}(X^s|w_s(e))p(w_s)$ la probabilité a posteriori des données du locuteurs s.
La MLES nous offre un estimateur de maximum de vraisemblance ou de maximum a posteriori alors que les méthodes d'analyse de données donnent l'estimateur des moindres carrés.

6.3 Regroupement hiérarchique ascendant

La réduction de l'espace par un algorithme de réduction de dimensionnalité constitue une des approches les plus simples et les plus intuitives. En revanche, les sous-espaces recherchés ne sont pas significatifs s'ils n'ont pas été générés par un grand nombre de données. Les matrices de covariance, d'inter-classe et d'intra-classe risquent d'être très mal estimées.
Indépendamment de ces problèmes d'estimation, il est intéressant de construire un espace de locuteurs virtuels pour lesquels on dispose de trames de parole associées à chacun d'eux. L'intérêt majeur est qu'on peut travailler directement sur les trames en appliquant des métriques sur les coefficients acoustiques ou bien en les projetant dans un autre espace acoustique. Le regroupement hiérarchique répond à ce type de problème : il s'agit de créer,

à chaque étape, une partition obtenue en agrégeant deux à deux les données des locuteurs les plus proches.
Le regroupement hiérarchique ascendant (ou *hierarchical clustering*) est très utilisé dans plusieurs applications. Il peut servir à grouper les messages d'une personne spécifiée laissés sur un répondeur téléphonique ou une boîte vocale [Reynolds et al., 1998] [Charlet, 2002]. Cette classification des messages peut être une étape d'un système de transcription ou d'indexation. Cela correspond à regrouper les données d'un même locuteur et les mettre dans un cluster. Ces données ainsi regroupées servent à adapter les modèles de parole au locuteur considéré dans le but d'améliorer les taux de reconnaissance. D'autre part, le clustering est aussi utilisé pour regrouper des segments de parole du même locuteur.
Dans le cadre de ce travail, le regroupement hiérarchique est utilisé pour regrouper deux à deux les locuteurs les plus proches. C'est une méthode originale pour construire un ensemble qui soit le plus représentatif possible de tous les locuteurs [Mami et Charlet, 2002a]. L'algorithme est illustré dans la figure 6.1.

Dans les paragraphes suivants, nous présentons les aspects théoriques du clustering et nous décrivons chaque étape de son algorithme.

6.3.1 Calcul des distances

On suppose au départ que l'ensemble des locuteurs à classer est muni d'une distance. Il s'agit parfois simplement d'une mesure de dissimilarité où l'inégalité triangulaire[6] n'est pas exigée[7]. Cela correspond à un critère de regroupement : à chaque itération, les deux locuteurs les plus proches au sens de ce critère, sont réunis. Soit un locuteur i et un locuteur j, la distance $d(i,j)$ entre ces deux locuteurs peut être définie par le rapport de vraisemblance généralisé, la distance de *Kullback-Leibler* ou encore par le rapport de vraisemblance croisé.

Rapport de vraisemblance généralisé

Soit deux locuteur i et j, leurs données X_i et X_j sont modélisées par deux processus gaussiens $\mathcal{N}(\mu_i, \Sigma_i)$ et $\mathcal{N}(\mu_j, \Sigma_j)$. Le rapport de vraisemblance généralisé repose sur le test d'hypothèses suivant :

- H_0 : les données X_i et X_j sont prononcées par deux locuteurs proches. Alors l'ensemble des données X_{ij} est supposé généré par un unique processus gaussien multi-dimensionnel $\mathcal{N}(\mu_{ij}, \Sigma_{ij})$ où X_{ij} est la concaténation des données X_i et X_j.
- H_1 : les données sont prononcées par deux locuteurs très éloignés (distincts) donc sont modélisées par deux processus gaussiens multi-dimensionnels différents.

Si $\mathscr{L}(X_i|\lambda_i)$ désigne la vraisemblance de la séquence X_i et $\mathscr{L}(X_j|\lambda_j)$ la vraisemblance de la séquence X_j (où λ_i, λ_j correspondent respectivement aux modèles des gaussiennes $\mathcal{N}(\mu_i, \Sigma_i)$ et $\mathcal{N}(\mu_j, \Sigma_j)$), alors la vraisemblance que les deux segments aient été générés

[6] Soit i, j et k trois locuteurs alors l'inégalité triangulaire est $d(i,j) \leq d(i,k) + d(k,j)$
[7] La condition $d(i,i) = 0$ n'est pas garantie.

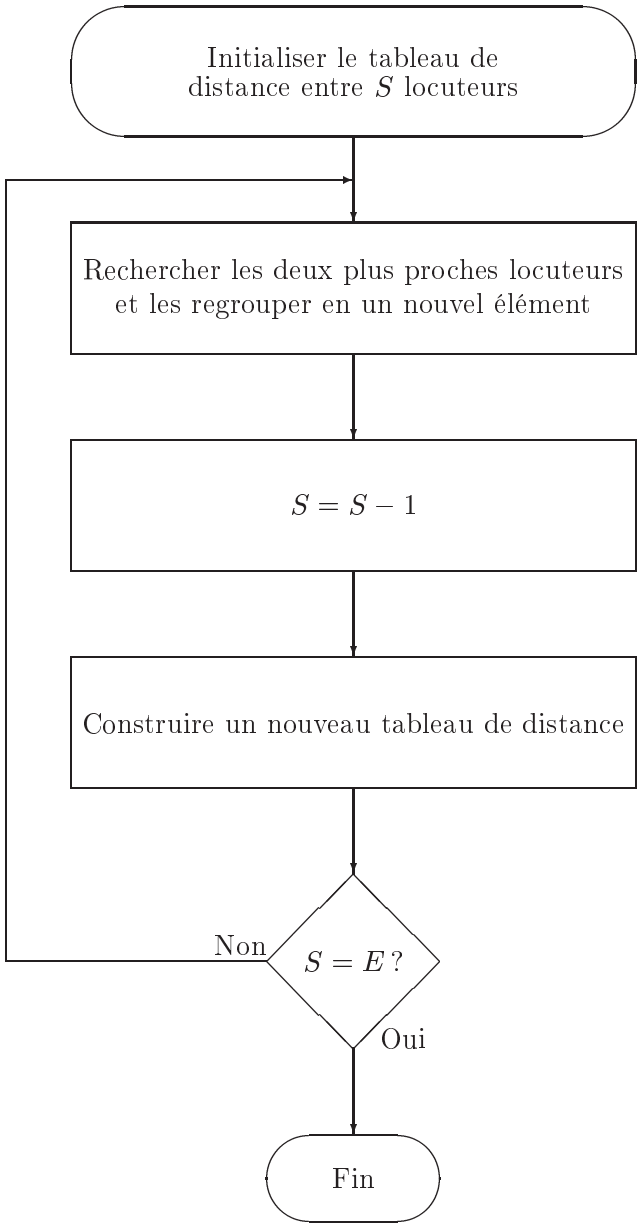


FIG. 6.1 – Regroupement hiérarchquie ascendant des locuteurs

par des locuteurs très éloignés $\mathscr{L}_1(i,j)$ est :

$$\mathscr{L}_1(i,j) = \mathscr{L}(X_i|\lambda_i) \times \mathscr{L}(X_j|\lambda_j)$$

La vraisemblance $\mathscr{L}_0(i,j)$ que les deux séquences aient été générées par deux locuteurs proches est $\mathscr{L}_0(i,j) = \mathscr{L}(X_{ij}|\lambda_{ij})$ où λ_{ij} est le modèle des données concaténées.
Le rapport de vraisemblance R_{GLR} correspondant est donné par[8] [Gish et al., 1991] :

$$\begin{aligned} R_{GLR} &= \frac{\mathscr{L}_0(i,j)}{\mathscr{L}_1(i,j)} \\ &= \frac{\mathscr{L}(X_{ij}|\lambda_{ij})}{\mathscr{L}(X_i|\lambda_i) \times \mathscr{L}(X_j|\lambda_j)} \end{aligned} \tag{6.22}$$

La distance est finalement obtenue en prenant l'opposé du logarithme de ce rapport de vraisemblance :

$$d_{GLR}(i,j) = -\log R_{GLR} \tag{6.23}$$

Cette distance, ou plutôt mesure de similarité, est symétrique mais elle ne respecte pas l'inégalité triangulaire. Dans le cas où on veut tester si les données X_i et X_j appartiennent ou pas au même locuteur, elle s'avère une bonne métrique de dissimilarité.

Distance de *Kullback-Leibler*

La distance de *Kullback-Leibler* [Solomonoff et al., 1998] [Mami, 2000] (ou entropie croisée relative) mesure la distance entre les distributions de probabilité des deux variables. Soit deux locuteurs i et j modélisés respectivement par les modèles λ_i et λ_j, la distance *Kullback-Leibler* est symétrisée et donnée par :

$$d_{KL}(i,j) = \log \frac{p(x_i|\lambda_i)}{p(x_i|\lambda_j)} + \log \frac{p(x_j|\lambda_j)}{p(x_j|\lambda_i)} \tag{6.24}$$

Cette distance est positive ou nulle[9] mais elle ne vérifie pas l'inégalité triangulaire. Une faible valeur (respectivement une forte valeur) de la métrique $d_{KL}(i,j)$ indique que les deux locuteurs i et j sont proches (respectivement éloignés).

Rapport de vraisemblance croisé

Pour calculer le rapport de vraisemblance croisée $d_{SCL}(i,j)$[10] [Reynolds et al., 1998] entre un locuteur i et un locuteur j, on évalue la vraisemblance $p(X_i|\lambda_j)$ de l'ensemble des trames acoustiques X_i par rapport au modèle λ_j du locuteur j. La distance $d_{SCL}(i,j)$ s'écrit (après normalisation) :

$$d_{SCL}(i,j) = \frac{1}{N_i} \log \frac{p(X_i|\lambda_{UBM})}{p(X_i|\lambda_j)} + \frac{1}{N_j} \log \frac{p(X_j|\lambda_{UBM})}{p(X_j|\lambda_i)} \tag{6.25}$$

[8] GLR pour *Generalized Likelihood Ratio*

[9] A condition que λ_i et λ_j soient les modèles optimaux pour X_i et X_j.

[10] SCL pour *Symmetric Cross Likelihood*

où N_i, N_j représentent, respectivement, le nombre de trames acoustiques du locuteur i et du locuteur j. λ_{UBM} est le modèle du monde (*Universal Background Model*) à partir duquel les modèles des locuteurs ont été appris.
Cette distance est symétrique $d_{SCL}(i,j) = d_{SCL}(j,i)$; en revanche $d_{SCL}(i,i) \neq 0$ et l'inégalité triangulaire n'est pas forcément vérifiée.

6.3.2 Construction du dendrogramme

La technique de regroupement décrite dans cette section est un regroupement hiérarchique par agglomération : après avoir estimé la matrice des distances $\mathscr{D}$ (de terme général $d(i,j)$), on recherche les deux locuteurs les plus proches qui correspondent à la plus petite distance. On peut se demander ensuite, sur quelle base on calcule la distance entre un locuteur donné et un groupe de locuteurs. Ceci revient à définir une stratégie de regroupement des locuteurs. Cette distance de regroupement peut être simplement exprimée en fonction des distances entre ce locuteur et chacun des locuteurs du cluster.
Pour fixer les idées, soit i, j, k trois locuteurs, si i et j sont regroupés en un seul élément noté "ij", la distance de ce groupement au locuteur k peut être définie de plusieurs manières :

- Le saut minimal (ou *single linkage*) : $d(k,ij) = \min\{d(k,i), d(k,j)\}$.
- Le saut maximal (ou *complete linkage*) : $d(k,ij) = \max\{d(k,i), d(k,j)\}$.
- La distance moyenne (ou *linkage*) : $d(k,ij) = \dfrac{N_i d(k,i) + N_j d(k,j)}{N_i + N_j}$.

où N_i, N_j représentent respectivement le nombre de trames acoustiques du locuteur i et du locuteur j.
Dans une autre approche, plus robuste, le locuteur hybride peut être modélisé à partir des données des deux locuteurs et son modèle GMM est ré-estimé. Si la distance d_{SCL} est utilisée, le critère d'agrégation est donné par :

$$d(k,ij) = \frac{1}{N_k} \log \frac{p(x_k|\lambda_{UBM})}{p(x_k|\lambda_{ij})} + \frac{1}{N_i + N_j} \log \frac{p(x_{ij}|\lambda_{UBM})}{p(x_{ij}|\lambda_k)} \tag{6.26}$$

où x_{ij} est la concaténation des données x_i et x_j. λ_{ij} est le modèle du locuteur hybride "ij".
Le processus est réitéré jusqu'à un niveau de partition donné.

6.3.3 Critère d'arrêt

Le critère d'arrêt permet de choisir à quel niveau de partition on peut stopper la construction de l'arbre ou du dendrogramme. Dans la littérature, le niveau de coupe est souvent choisi en fonction de la qualité du clustering et on utilise par exemple la métrique I_{BBN} ou I_{rand}[11] [Solomonoff et al., 1998]. Ces dernières ne sont pas adaptées à notre cas

[11] En générale, ces métriques évaluent le taux d'occurrences provenant du même locuteur et qui ne sont pas dans le même cluster ou bien le taux d'occurrences qui sont dans un même cluster mais qui proviennent des locuteurs différents.

de figure, elles reposent sur le principe qu'une partition est parfaite si chaque cluster ne contient que les données d'un seul locuteur, alors que notre clustering se propose de regrouper deux à deux les locuteurs les plus proches. Une alternative consiste à prendre le critère d'information bayésien (BIC) (ou le critère de *Schwarz*) comme un critère d'arrêt [Chen et Gopalakrishnan, 1998].
Le BIC est un critère fondé sur une approche bayésienne, qui pénalise le log-vraisemblance par une fonction linéaire de l'ordre du modèle. Il en existe bien d'autres, mais la forme générale de ces critères d'information peut se résumer à :

$$BIC(i) = 2 \log \mathscr{L}(X) - \alpha l(i) \log N \tag{6.27}$$

où :
- $l(i)$ est le nombre de paramètres des modèles de locuteurs à la partition i ;
- $\mathscr{L}(X)$ est la vraisemblance de tous les N trames des données de tous les locuteurs à l'itération i ;
- α est un poids de pénalité, en théorie égal à 1.

Le premier terme du BIC reflète l'ajustement du modèle aux données et le deuxième terme correspond à la complexité du modèle.
En pratique, ce critère est évalué à chaque itération. Il permet de sélectionner la meilleure partition (c'est-à-dire le meilleur espace) et dont la complexité des modèles reste raisonnable.

6.4 Sélection d'un sous-ensemble de locuteurs

La sélection d'un sous-ensemble de locuteurs est une approche alternative aux voix propres et au clustering. Il s'agit de rechercher, parmi un ensemble de locuteurs potentiels S, un sous-ensemble de E locuteurs qui optimise un certain critère. La sélection exhaustive devient vite insurmontable vu la complexité et la taille des problèmes. La recherche de ce sous-ensemble est un problème d'optimisation qu'on peut diviser en deux parties :

1. Définir un critère de sélection.
2. Sélectionner le meilleur sous-ensemble au sens du critère.

Dans les paragraphes suivants, on présentera quelques critères de sélection et on parlera ensuite des principaux algorithmes de sélection.

6.4.1 Critères de sélection

Généralement, à partir de S de locuteurs, les procédure et les algorithmes de sélection opèrent en deux phases :
- on forme des sous-ensemble de $S - 1$ locuteurs ;
- on évalue une mesure (exemple figure 6.2).

Le rôle du critère de sélection est de désigner le meilleur sous-ensemble.
Soit un sous-ensemble $\mathcal{N}$ de n locuteurs. On peut citer, à titres d'exemples, les critères de sélection suivants :

Critère du *F-ratio* : il permet de "quantifier" la discrimination d'un sous-ensemble de n locuteurs. Il est donné par :

$$F\text{-}ratio_{\mathcal{N}} = \frac{variance\ des\ moyennes\ des\ n\ locuteurs}{moyenne\ des\ variances\ des\ n\ locuteurs} \tag{6.28}$$

Les moyennes et les variances des locuteurs peuvent être les moyennes et les variances de leurs vecteurs acoustiques ou bien les moyennes et les variances de leurs modèles GMM. Le *F-ratio* est un critère à maximiser.

Critère du taux d'erreur : c'est le critère le plus intuitif. Il consiste à maximiser directement le taux de reconnaissance. Si le sous-ensemble $\mathcal{N}$ donne un taux de reconnaissance meilleur que les autres sous-ensembles, alors il est certainement le "meilleur" au sens de ce critère. La fonction d'évaluation se résume aux seuls tests de reconnaissance. Par conséquent, le coût de la procédure devient vite très important.

Critère de la divergence : comme on recherche le meilleur sous-ensemble, on peut estimer les matrices de covariance intra-classe (d'un sous-ensemble) $W_{\mathcal{N}}$ et inter-classe (entre différents sous-ensembles) B et maximiser la divergence donnée par :

$$D_{\mathcal{N}} = tr(W_{\mathcal{N}}^{-1} \cdot B) \tag{6.29}$$

Maximisation de la vraisemblance des données : Il s'agit de sélectionner, à chaque itération, le sous-ensemble qui maximise la vraisemblance totale des données.

Mesure de la dispersion : parmi S locuteurs, on recherche le sous-ensemble des locuteurs les plus dispersés. On suppose que plus les locuteurs sont dispersés meilleur est l'espace de représentation. Pour chaque sous-ensemble $\mathcal{N}$ de n locuteurs, la mesure de dispersion est donnée par :

$$\mu_{\mathcal{N}} = \frac{1}{n^2} \sum_{i,j=1}^{n} d(i,j) \tag{6.30}$$

où i et j sont des locuteurs appartenant au sous-ensemble $\mathcal{N}$ et $d(i,j)$ est la distance entre eux donnée par l'équation 6.25.

6.4.2 Algorithmes de sélection

Les algorithmes de sélection d'un sous-ensemble sont multiples et leur regroupement en différentes catégories est forcément subjectif. On distingue souvent les trois familles d'algorithmes [Fredouille, 2000] : exponentiels, aléatoires [Yang et Honavar, 1998] et séquentiels. Notons que dans la littérature, les méthodes de sélection du sous-ensemble optimal sont toujours appliquées sur les coefficients acoustiques des locuteurs et n'ont jamais été utilisées pour la sélection d'un sous-ensemble de locuteurs..

Algorithmes exponentiels

Il s'agit d'une recherche brutale du sous-ensemble E où tous les sous-ensemble possibles et imaginables sont évalués selon un critère de sélection. Ainsi, on évalue $2^S - 1$ sous-ensembles au lieu de 2^S car on n'évalue pas le sous-ensemble vide. Ce type d'algorithmes devient impossible à mettre en œuvre si le nombre de locuteurs potentiels S est important (> 10), leur complexité est de type exponentiel $O(2^S)$. La nécessité de rechercher des heuristiques de sélection s'impose très vite.

Algorithmes aléatoires : algorithmes génétiques

Ces algorithmes d'optimisation se situent dans un cadre stochastique et font appel à des méthodes aléatoires. On trouve à leur tête les algorithmes génétiques. Ces derniers sont inspirés des mécanismes de la sélection naturelle et de la génétique. Ils consistent à faire évoluer une population (ensemble d'individus) à l'aide de différents opérateurs : sélection, croisement et mutations.
La première étape est de définir et de coder convenablement le problème. A chaque variable d'optimisation nous faisons correspondre un gène. Nous appelons chromosome un ensemble de gènes. Chaque individu est constitué d'un ou plusieurs chromosomes. Nous appelons population un ensemble d'individus.
Pour commencer, l'algorithme génétique génère aléatoirement une population initiale (comme solutions possibles). Il opère, ensuite un croisement des meilleurs chromosomes (les meilleurs sont choisis par une fonction d'évaluation). Ce croisement ou hybridation consiste en l'échange d'un certain nombre de bits (gènes) entre les deux parents. Les meilleurs enfants obtenus seront croisés, à leur tour, pour obtenir encore une meilleure génération. L'algorithme crée des mutations (change la valeur de quelques bits aléatoirement) pour bien imiter le processus naturel. On répète ces étapes jusqu'à ce que la population soit proche de la solution recherchée. Les algorithmes génétiques ont été introduits par [Charlet, 1997] et aussi par [Demirekler et Haydar, 1999] pour optimiser les paramètres acoustiques des locuteurs.

Algorithmes séquentiels

Cette famille d'algorithmes se caractérise par une complexité polynomiale de type $O(n^2)$ et semble être la plus intéressante du point de vue de complexité calculatoire. Trois techniques fréquemment utilisées sont présentées :

Sélection des N-meilleurs : il s'agit de classer les locuteurs selon un certain critère puis de sélectionner les E meilleurs au sens de ce critère. Cette procédure requiert $2S - 1$ évaluations.

Procédure d'élimination pas à pas (knock-out) : cette approche proposée dans [Sambur, 1975] consiste à évaluer tous les sous-ensemble de $S - 1$. On élimine ensuite le locuteur qui

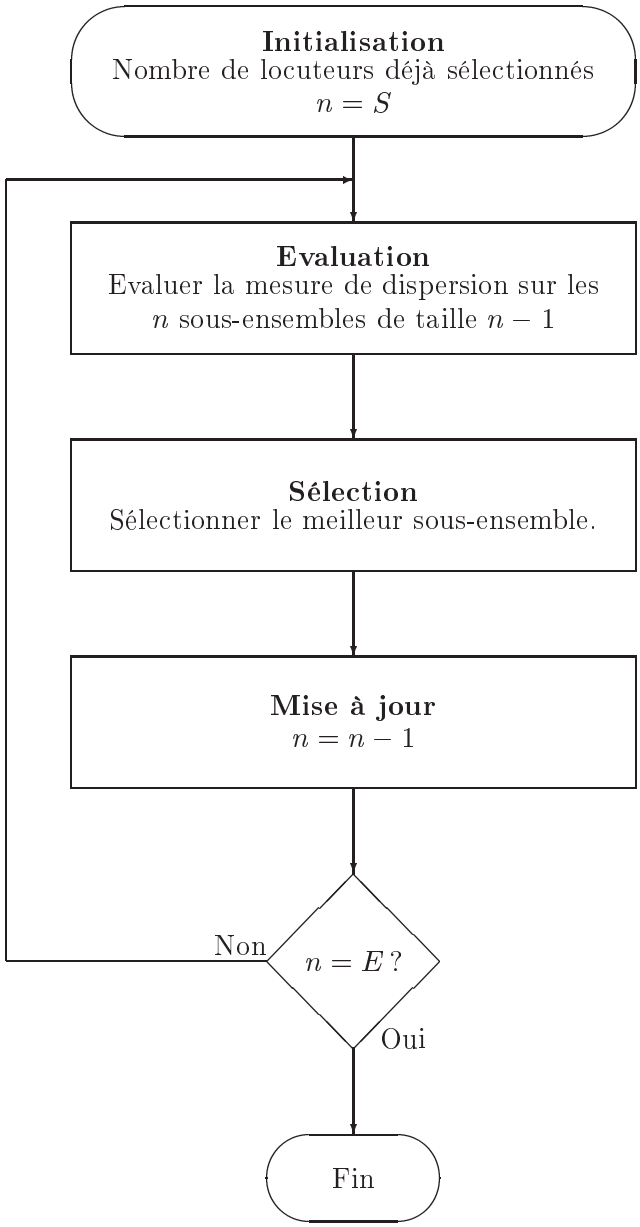


FIG. 6.2 – Procédure du knock-out

n'appartient pas au meilleur sous-ensemble sélectionné (figure 6.2). Cette procédure est ré-itérée jusqu'à n'avoir que E locuteurs finaux. A chaque itération, on suppose que le meilleur sous-ensemble de E locuteurs inclut le meilleur sous-ensemble de $E-1$ locuteurs. Le nombre d'évaluations à effectuer est ramené à $\frac{1}{2}S(S+1)$.

Procédure de sélection ascendante : il s'agit d'une approche tout à fait symétrique au knock-out. A chaque itération, on évalue tous les sous-ensembles constitués en ajoutant un locuteur et on retient le meilleur d'eux au sens du critère de sélection. Cette procédure requiert $\frac{1}{2}S(S+1)$.

6.4.3 Procédure et critères de sélections utilisés

Dans nos expériences (paragraphe 8.2.1), nous avons appliqué la procédure du knock-out pour sélectionner le sous-ensemble de E locuteurs les plus dispersés (parmi $n = S$ locuteurs). Pour commencer, nous prenons comme mesure de dispersion la moyenne des distances inter-locuteurs deux à deux (équation 6.30) et nous appliquons ensuite la procédure du knock-out comme illustré dans la figure 6.2. L'algorithme utilisé est un algorithme itératif : à chaque itération, on calcule la dispersion de tous les sous-ensembles de $(n-1)$ locuteurs et *on supprime l'individu (ou le locuteur) sans lequel la dispersion est maximale.*

6.5 Conclusion

Dans ce chapitre nous avons présenté différentes approches pour la construction d'un espace représentatif des locuteurs.
Les méthodes d'analyse de données offrent un espace orthogonal et permettent d'une part un meilleur ordonnancement de l'espace des locuteurs et d'autre part une meilleure discrimination entre les classes de locuteurs. En général, il s'agit de minimiser une erreur quadratique ou de maximiser un critère de séparabilité entre classes. La PPCA, en revanche, maximise la vraisemblance des données et se base ainsi sur une approche statistique plutôt que géométrique.
Dans une autre approche, plus robuste, la MLES cherche à estimer l'espace en maximisant directement la vraisemblance des vecteurs acoustiques et donne un cadre théorique très élégant. En revanche, elle fait intervenir des connaissances a priori sur la base de données utilisée qu'on ne peut pas toujours formuler.
Le clustering ascendant, quant à lui, fusionne les voix et permet d'avoir des trames acoustiques associées à chaque voix virtuelle. C'est un grand avantage car on peut envisager de travailler directement sur les paramètres acoustiques.
Les méthodes évoquées précédemment créent des "locuteurs virtuels".
D'autre part, il existe des méthodes moins lourdes à mettre en œuvre telles que les méthodes heuristiques, en particulier le knock-out, qui sélectionne le groupe de locuteur le plus dispersé selon un critère donné. La qualité de ces techniques se mesure par la qualité

et la complexité de la fonction d'évaluation.
Ces méthodes de construction d'espace seront appliquées et comparées dans le chapitre 8.

Chapitre 7

Localisation et décision

La localisation des locuteurs consiste à placer chaque locuteur dans l'espace représentatif précédemment construit (chapitre 6). Les techniques de localisation présentées dans ce chapitre sont :

Projection orthogonale : c'est l'approche la plus intuitive pour localiser un locuteur dans un espace. Elle ne peut être utilisée que si l'espace est orthogonal.

Localisation par maximum de vraisemblance : cette approche a été proposée et utilisée par [Kuhn et al., 1998]. Il s'agit de déterminer les coefficients de projection en maximisant la vraisemblance des locuteurs. Cette approche est très similaire à des précédentes techniques comme la RSW ou la CAT (cf. chapitre 5).

Localisation par les modèles d'ancrage : cette technique a été initialement proposée dans l'indexation en locuteurs des documents audio.

Dans la phase de décision, nous évaluons la proximité spatiale entre les locuteurs par l'application d'une métrique entre les coordonnées des locuteurs. Ces métriques doivent être utilisées dans un espace orthogonal, ce qui n'est pas toujours le cas. Pour rectifier cette démarche, nous avons proposé de retrouver l'orthogonalité de l'espace et ainsi appliquer correctement les métriques [Mami et Charlet, 2003a].

7.1 Localisation des locuteurs

Dans le chapitre 5, on a montré que chaque locuteur peut être représenté dans l'espace représentatif et son modèle λ approximé par la relation :

$$\lambda \longrightarrow \{w_e\}_{e=1,\ldots,E} \tag{7.1}$$

où E est la dimension de l'espace représentatif c'est-à-dire le nombre de locuteurs de référence $\{\overline{\lambda}_{e=1,\ldots,E}\}$. Ainsi, on associe à chaque locuteur λ un vecteur caractéristique $w = \{w_e\}_{e=1,\ldots,E}$.

Dans cette section, il s'agit justement de calculer ces coefficients caractéristiques. Autrement dit, localiser les locuteurs dans l'espace représentatif. La localisation est réalisée soit par une projection orthogonale classique, soit par maximum de vraisemblance ou encore par les modèles d'ancrage.

7.1.1 Projection orthogonale

L'idée la plus intuitive pour placer un locuteur dans l'espace représentatif est de projeter son vecteur de paramètres dans cet espace. Cette projection n'a de sens que si l'espace est orthogonal.
Soit un locuteur s représenté par un vecteur de paramètres V^s de dimension $M \times D$, si le locuteur est modélisé par M gaussiennes, et soit un espace Y de dimension $E \times (M \times D)$. La représentation spatiale w^s (ou le vecteur des coordonnées) de ce locuteur est tout simplement une projection du vecteur de paramètres dans cet espace :

$$w^s = Y \cdot V^s \tag{7.2}$$

Le nouveau modèle du locuteur s est un vecteur de coordonnées de dimension réduite E.

7.1.2 Localisation par maximum de vraisemblance (MLED)

Il s'agit d'estimer les coefficients propres des locuteurs $\{w_e\}_{e=1,\cdots,E}$ en maximisant la vraisemblance des observations X [Kuhn et al., 1998], soit :

$$\widehat{w} = \arg\max_{w} \mathscr{L}(X|w)$$

Comme on l'a déjà vu au paragraphe 6.2, la maximisation de la vraisemblance des données passe par la maximisation et la définition de la fonction auxiliaire $Q(w, w^{(t)})$ où w est le vecteur des coefficients propres à estimer et $w^{(t)}$ est son estimation courante (à l'itération t).

$$Q(w, w^{(t)}) = \sum_{m=1}^{M} P(m|X, w^{(t)}) \log P(X, m|w)$$

où m est une distribution dans un modèle de M gaussiennes (et qui également joue le rôle de la variable cachée).
On introduit le terme $\gamma_{n,m}^{(t)}$ (équation 2.7) qui représente la probabilité a posteriori des données à l'itérations t, et on suppose que chaque distribution m est une gaussienne de moyennes μ_m et de matrice de covariance Σ_m et qu'elle est pondérée par un poids π_m.

Pour un locuteur s donné, la fonction auxiliaire s'écrit :

$$\begin{aligned} Q(w, w^{(t)}) &= \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{n,m}^{(t)} \log \pi_m \\ &+ \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{n,m}^{(t)} \left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_m| \right] \\ &+ \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{n,m}^{(t)} \left[-\frac{1}{2}(x_n - \mu_m)'\Sigma_m^{-1}(x_n - \mu_m) \right] \end{aligned} \tag{7.3}$$

Si on ne considère que les moyennes des modèles GMM des locuteurs, c'est-à-dire que le tableau initial des paramètres ne contient que les moyennes GMM des locuteurs (paragraphe 6.2), alors, les moyennes des distributions s'écrivent :

$$\mu_m = \sum_{e=1}^{E} w_e \overline{\mu}_m(e) \tag{7.4}$$

où $\overline{\mu}_m(e)$ est une voix propre, E est le nombre de voix propres et w_e est le coefficient correspondant à estimer. Dans ces conditions, seul le troisième terme de l'expression précédente de la fonction auxiliaire $Q(w, w^{(t)})$ dépend du coefficient w_e. En supposant que les voix propres sont indépendantes $\left(\frac{\partial w(i)}{\partial w(j)} = 0 \text{ pour } i \neq j\right)$[1], la condition d'optimisation $\frac{\partial Q(w,w^{(t)})}{\partial w_e} = 0$ s'écrit :

$$\sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{n,m}^{(t)} \overline{\mu}'_m(k)\Sigma_m^{-1} x_n = \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{n,m}^{(t)} \sum_{e=1}^{E} w_e \overline{\mu}'_m(e)\Sigma_m^{-1}\mu_m(k) \tag{7.5}$$

Avec $k = 1, 2, \cdots, E$. L'équation précédente peut se mettre sous forme matricielle :

$$\Psi = \Phi w$$

avec :

$$\begin{aligned} \psi_i &= \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{n,m}^{(t)} \overline{\mu}'_m(i)\Sigma_m^{-1} x_n \\ \phi_{ij} &= \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_{n,m}^{(t)} \overline{\mu}'_m(j)\Sigma_m^{-1}\overline{\mu}_m(i) \\ w &= [w_1, \cdots, w_E]' \end{aligned}$$

[1] Cette condition indique que les vecteurs $w(i)$ et $w(j)$ sont linéairement indépendants c'est-à-dire que l'espace des locuteurs est orthogonal.

avec $i, j = 1, \cdots, E$.
La déterminations des coefficients propres w_e revient à la résolution d'un système d'équations linéaires, c'est-à-dire, l'inversion d'une matrice $E \times E$ qui est réalisée par le biais de la SVD (pour *Singular Value Decomposition*) ou par la méthode d'élimination gaussienne.

7.1.3 Localisation par les modèles d'ancrage

Il s'agit de caractériser et de représenter un signal de parole par rapport un ensemble de modèles de locuteurs bien appris [Merlin et al., 1999] [Sturim et al., 2001]. Les modèles d'ancrage (ou *Anchor Models*) ont été utilisés dans le regroupement, la détection et l'indexation des locuteurs. Dans notre étude, ils sont utilisés pour localiser les locuteurs dans l'espace représentatif [Mami et Charlet, 2002b]. En effet, si l'espace construit n'est pas orthogonal, on ne peut pas localiser les locuteurs par une simple projection. Les coefficients caractéristiques sont estimés par MLED ou par les modèles d'ancrage. En pratique ces derniers correspondent aux modèles de locuteurs de référence.
Le principe repose sur le calcul d'un score de vraisemblance dans chaque direction de l'espace c'est-à-dire qu'on évalue la vraisemblance des données par rapport à chaque locuteur de référence. L'ensemble des scores constitue ainsi le vecteur des coordonnées du locuteur, soit :

$$w = \begin{bmatrix} \widetilde{p}(x|\overline{\lambda}_1) \\ \widetilde{p}(x|\overline{\lambda}_2) \\ \vdots \\ \widetilde{p}(x|\overline{\lambda}_E) \end{bmatrix} \tag{7.6}$$

où $\widetilde{p}(x|\overline{\lambda}_e)$ est un score de vraisemblance normalisée des données x (de N trames acoustiques) sachant le modèle GMM du locuteur de référence $\overline{\lambda}_e$. Il correspond à la vraisemblance normalisée par un modèle universel :

$$\widetilde{p}(x|\overline{\lambda}_e) = \frac{1}{N} \log \frac{p(x|\overline{\lambda}_e)}{p(x|\lambda_{UBM})} \tag{7.7}$$

où λ_{UBM} est le modèle du monde (*Universal Background Model*) qui a servi à apprendre tous les modèles des locuteurs [Reynolds, 1997].

7.2 Décision : identification des locuteurs par localisation

A présent, chaque locuteur dispose de son vecteur de coordonnées. La tâche de notre système de reconnaissance est ensuite de mesurer la similarité entre ces locuteurs. Cette étape fait l'objet de cette section.

7.2.1 Définition d'une métrique

Après avoir représenté tous les locuteurs par des points dans un espace de locuteurs de référence, on évalue la proximité entre eux. Le locuteur dont le point de référence le plus

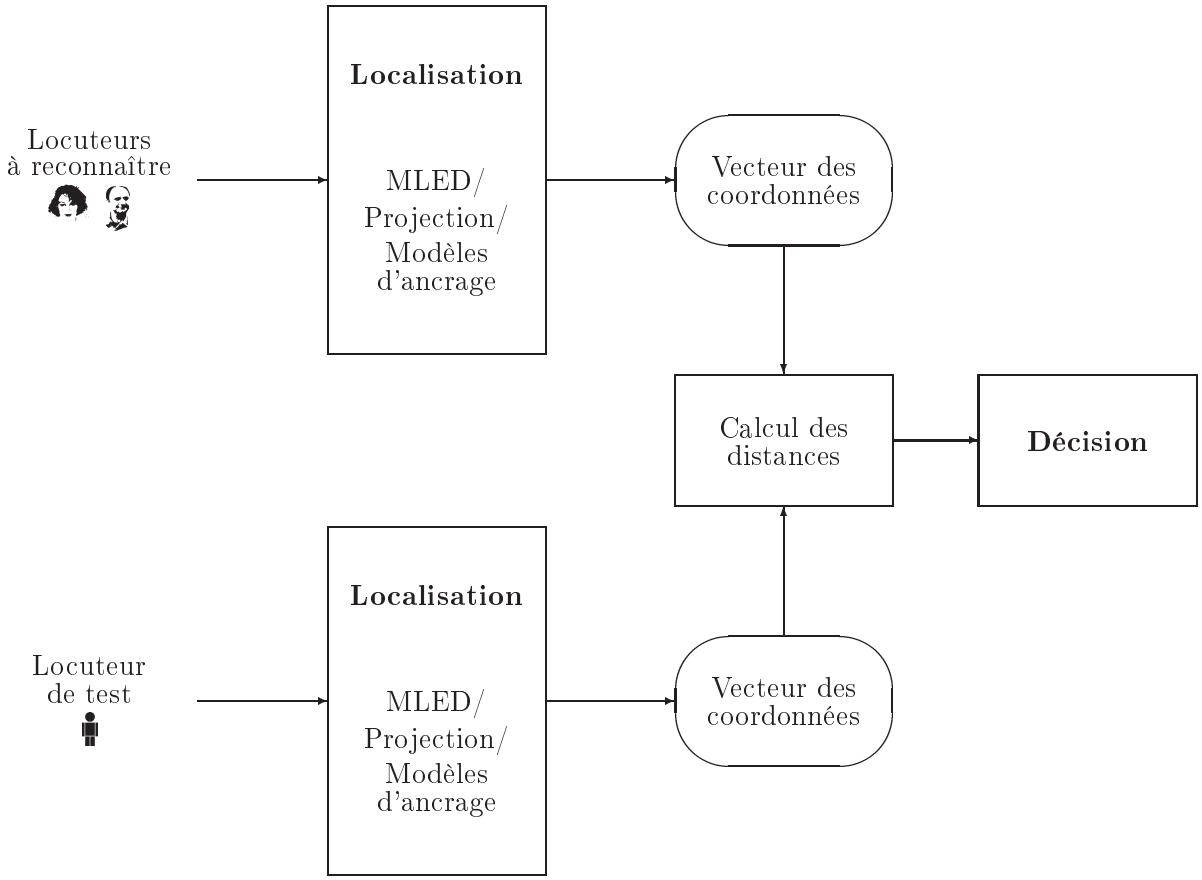


FIG. 7.1 – Système d'identification des locuteurs

proche du locuteur inconnu constitue le locuteur reconnu. Il est donc nécessaire de définir une distance entre deux points ou deux locuteurs.
Soit R un locuteur à reconnaître de modèle λ_R et T un locuteur de test de modèle λ_T représentés respectivement par les vecteurs $[r_1, \ldots, r_E]^T$ et $[t_1, \ldots, t_E]^T$. Le score d'identification correspond simplement à une mesure de distance entre ces deux points de l'espace (figure 7.1). Le locuteur reconnu $\hat{R}$ est celui dont le modèle donne la plus petite distance :

$$\hat{R} = \arg\min_R d(R, T) \tag{7.8}$$

On définit un certain nombre de distances classiques :

la distance de Hamming : $d_1(R,T) = \sum_{i=1}^{E} |r_i - t_i|$;

la distance euclidienne : $d_2(R,T) = \sqrt{\sum_{i=1}^{E} (r_i - t_i)^2}$;

la distance du maximum : $d_\infty(R,T) = \max_{i=1,\ldots,E} |r_i - t_i|$;

la distance d_n : $d_n(R,T) = \left[\sum_{i=1}^{E} |r_i - t_i|^n \right]^{1/n}$.

La distance d_n correspond à la définition générale de la distance. En effet, avec $n = 1$ on retrouve la distance de Hamming, avec $n = 2$, on retrouve la distance euclidienne, et avec $n = \infty$, on retrouve la distance du maximum.
Par ailleurs, on peut penser que l'orientation des vecteurs caractéristiques des locuteurs est importante et on peut mesurer l'angle entre deux vecteurs.

angle : $\delta(R,T) = \arccos \left[\frac{r^T t}{\sqrt{r^T r \cdot t^T t}} \right]$.

7.2.2 Distances pondérées

Pour améliorer les résultats, on introduit des pondérations dans les distances définies dans le paragraphe précédent. La distance d_n devient :

$$d_n^{pond}(R,T) = \left[\sum_{i=1}^{E} pond(i) \cdot |r_i - t_i|^n \right]^{1/n}$$

Ce traitement est particulièrement intéressant si l'espace est construit par ACP ou ALD. Trois cas de figures sont possibles :

- $pond(i) = 1$: ne pas tenir compte des poids des axes.
- $pond(i) = \frac{1}{\lambda(i)}$: pondérer avec l'inverse des valeurs propres de la matrice de covariance. Ce cas concerne particulièrement l'ACP et l'ALD.

– Appliquer la distance Mahalanobis :

$$d_{Mahalanobis}(R,T) = (r-t)^T \Sigma_0^{-1}(r-t)$$

où Σ_0 est la matrice de covariance soit des coordonnées elles mêmes c'est-à-dire du même corpus, soit d'un autre corpus de données.

7.2.3 Post-traitement ACP et ALD

Les distances présentées précédemment sont toutes des distances qui doivent être utilisées dans des espace orthogonaux. Dans le cas ou l'espace représentatif des locuteurs ne l'est pas (le cas où l'espace est construit par regroupement hiérarchique ou par knock-out), on peut ré-ajuster les axes et faire une rotation pour retrouver l'orthogonalité et appliquer ensuite les métriques précédentes [Mami et Charlet, 2003a].
En pratique, cela se traduit par l'application d'une ACP ou d'une ALD sur les vecteurs des coordonnées w des locuteurs :

$$w \overset{orthogonalisation}{\longrightarrow} w_\perp$$

où $w_\perp$ représente les vecteurs des coordonnées transformés par ACP ou ALD.

Orthogonalisation par ACP

Dans le paragraphe 6.1.2, on a montré que, pour transformer des données hétérogènes et pour les ramener dans un espace orthogonal, on effectue une rotation de sorte à donner la même inertie à tous les axes. La matrice de transformation ACP est donnée par :

$$T_{ACP} = U\Lambda^{-1/2}$$

où U et Λ sont respectivement la matrice des vecteurs propres et la matrice diagonale des valeurs propres de la matrice de covariance C.
La matrice de covariance est estimée à partir d'un autre corpus de données (par rapport aux données d'apprentissage). L'idée est d'estimer la matrice de transformation T_{ACP} à partir d'un corpus de données différent et créer des nouveaux axes de locuteurs indépendamment des données initiales.
Ainsi, chaque vecteur de coordonnées est transformé de la façon suivante :

$$w_\perp = T_{ACP} w \tag{7.9}$$

On applique ensuite les métriques habituelles (définies dans 7.2.1) entre les coordonnées transformées :

$$d_n(R,T) = \left[\sum_{i=1}^{E} |r_{\perp i} - t_{\perp i}|^n\right]^{1/n}$$

$$\delta(R,T) = \arccos\left[\frac{r_\perp^T t_\perp}{\sqrt{r_\perp^T r_\perp \cdot t_\perp^T t_\perp}}\right]$$

Orthogonalisation par ALD

Pour effectuer une transformation ALD sur les coordonnées des locuteurs, on procède de la même manière que pour la transformation ACP. Les vecteurs de coordonnées sont transformés de la façon suivante :

$$w_{\perp} = T_{ALD} w \tag{7.10}$$

Cette transformation permet de retrouver l'orthogonalité de l'espace et donc d'appliquer correctement les métriques (définies dans le paragraphe précédent).
La matrice de transformation ALD est donnée par (paragraphe 6.1.4) :

$$T_{ALD} = U_W \Lambda_W^{-1/2} U_B$$

où U_W et Λ_W sont respectivement la matrice des vecteurs propres et la matrice diagonale des valeurs propres de la matrice intra-classe W. U_B est la matrice des vecteurs propres de la matrice inter-classe B.
Les matrices de covariance intra- et inter-classe sont estimées à partir d'un corpus de développement ce qui permet de créer des classes de locuteurs complètement indépendantes des locuteurs de référence et des locuteurs à reconnaître.

7.3 Conclusion

Dans ce chapitre, nous avons décrit la phase de décision d'un système de reconnaissance par localisation dans un espace de locuteurs de référence. Nous avons présenté les techniques de localisation suivantes : projection orthogonale, MLED et modèles d'ancrage. La projection orthogonale est l'approche la plus intuitive et elle est facile à mettre en œuvre. La MLED est une technique de localisation optimale. Cependant, la détermination du vecteur des coordonnées est une phase fastidieuse (nombre d'opérations important). La localisation par les modèles d'ancrage semble bien caractériser le signal de parole d'un locuteur. La localisation est un simple calcul d'un score de vraisemblance par rapport à chaque modèle de référence.
Une fois les locuteurs localisés, nous appliquons des métriques (distance d'Hamming, euclidienne, du maximum ou l'angle) entre leurs vecteurs de coordonnées. Un post-traitement ACP ou ALD sur ces vecteurs permet de retrouver l'orthogonalité de l'espace et d'appliquer correctement les métriques.

Chapitre 8

Evaluation des systèmes d'identification du locuteur par localisation

Dans ce chapitre, nous allons évaluer les systèmes d'identification par localisation dans un espace de locuteurs de référence, en mode indépendant du texte. Nous étudierons principalement les deux systèmes suivants :

- Les locuteurs sont localisés par les modèles d'ancrage dans un espace construit par regroupement hiérarchique.
- Les locuteurs sont localisés par les modèles d'ancrage dans un espace construit par sélection.

Le but étant d'évaluer les performances d'identification par localisation et l'impact de différents paramètres tels que la taille de l'espace, la quantité de donnée d'apprentissage, etc.

Nous rappelons que les bases de données utilisées ainsi que le protocole expérimental sont décrits dans le chapitre 3

8.1 Evaluation de la localisation par les modèles d'ancrage dans un espace construit par regroupement hiérarchique

Dans cette expérience, les étapes du système de reconnaissance sont :

Construction de l'espace par regroupement hiérarchique. Les 500 locuteurs de l'ensemble $\mathscr{E}_3$ sont regroupés deux à deux pour générer E locuteurs virtuels (cf. 6.3)[1]. La distance utilisée est le rapport de vraisemblance croisé donnée par l'équation 6.25. A chaque niveau de partition, nous recherchons les deux plus proches locuteurs, nous

[1] On parle de locuteurs virtuels car les locuteurs générés par regroupement ne représentent pas des "vrais locuteurs".

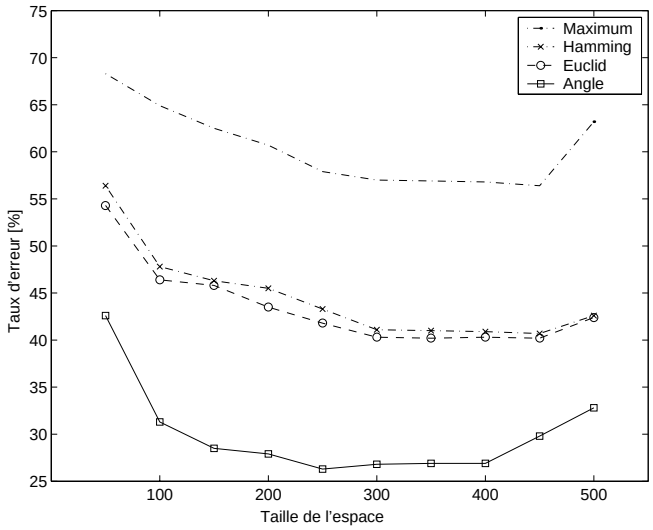


FIG. 8.1 – Regroupement hiérarchique : taux d'erreur en fonction de la taille de l'espace

les regroupons en un seul locuteur et nous estimons son modèle GMM. La distance entre ce locuteur hybride et un autre locuteur est donnée par l'équation 6.26.

Localisation et décision. L'espace représentatif des locuteurs n'étant pas orthogonal, la localisation des locuteurs est réalisée par les modèles d'ancrage (7.1.3). Nous appliquons ensuite des métriques pour faire l'identification.

Le but de cette manipulation est d'évaluer les performances de reconnaissance par localisation dans un espace construit par regroupement hiérarchique ascendant et l'impact de différents paramètres tels que la taille de l'espace, la métrique d'identification, la quantité de donnée d'apprentissage, etc.

8.1.1 Influence de la taille de l'espace et de la métrique

Dans ce paragraphe, nous allons étudier l'impact de la taille de l'espace et de la métrique appliquée. Les paramètres de cette expérience sont :

- Les locuteurs de l'espace (de l'ensemble $\mathscr{E}_3$) sont modélisés par 256 gaussiennes.
- La localisation des locuteurs est réalisée par les modèles d'ancrage et l'apprentissage se fait avec 100 secondes de parole[2].

La figure 8.1 représente les taux d'erreur des 50 locuteurs (de l'ensemble $\mathscr{E}_1$) en fonction de la taille d'espace pour différentes métriques. L'angle semble toujours la métrique la plus discriminante entre locuteurs et nous pouvons distinguer clairement trois régions de la courbe de variation :

[2]La quantité d'apprentissage est la quantité de signal avec laquelle les locuteurs sont localisés par les modèles d'ancrage.

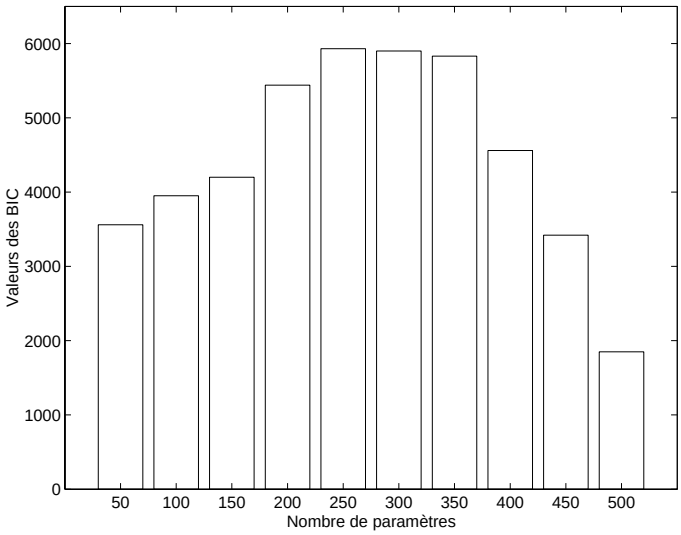


FIG. 8.2 – Regroupement hiérarchique : estimation du meilleur espace par critère BIC

- Pour un espace inférieur à 150 locuteurs virtuels, les taux d'erreur diminuent progressivement jusqu'à atteindre les 28.5% d'identification incorrecte.
- Les taux d'erreur gardent ensuite des valeurs, à peu près constantes et atteignent une valeur minimale de 26.3% à un espace de 250 locuteurs virtuels.
- Pour un espace de dimension $\geq$ à 400, les performances se dégradent et les taux d'erreur d'identification atteignent 32.8% lorsque l'espace est constitué de l'ensemble des 500 locuteurs.

Compte-tenu de la métrique qui donne à tous les modèles d'ancrage la même importance, il apparaît intéressant de regrouper certains modèles très proches pour obtenir un pavage plus homogène de l'espace.
Par ailleurs, les courbes de variation des autres distances Hamming, euclidienne et du maximum suivent pratiquement les mêmes variations que la courbe de variation de l'angle et les remarques précédentes sont toujours valables.
Le regroupement hiérarchique permet ainsi d'obtenir un espace "optimal" de locuteurs virtuels parmi les 500 locuteurs de l'ensemble $\mathscr{E}_3$.

8.1.2 Estimation théorique de la taille du meilleur espace

L'expérience précédente a montré que les performances d'identification atteignent une valeur maximale pour un espace de 250 locuteurs virtuels. Dans ce paragraphe, nous allons calculer les valeurs du critère BIC qui correspondent aux différents niveaux de partition et sélectionner le meilleur espace au sens de ce critère. Notons que les locuteurs qui servent au regroupement hiérarchique (ensemble $\mathscr{E}_3$) sont modélisés par 256 gaussiennes.

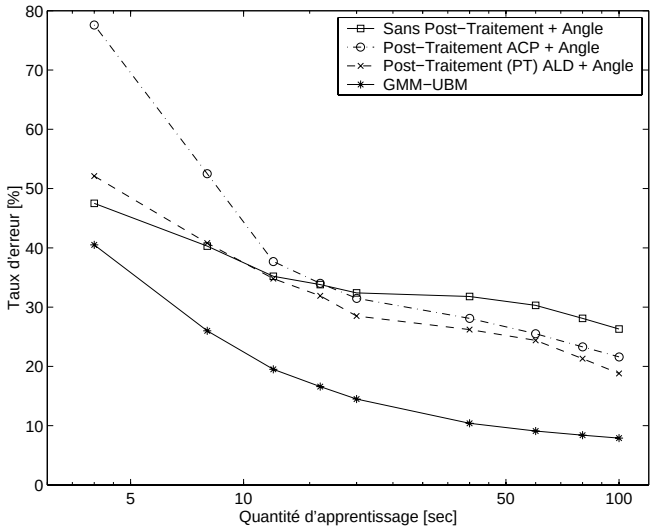


FIG. 8.3 – Regroupement hiérarchique : taux d'erreur en fonction de la quantité de données d'apprentissage

Nous rappelons l'expression théorique du critère d'information bayésien (paragraphe 6.3.3) :

$$BIC(i) = 2 \log \mathscr{L}(X) - \alpha l(i) \log N$$

où :

- $l(i)$ est le nombre de paramètres qui modélisent un locuteur. Par exemple, pour un espace de 200 locuteurs $l(200) = 200$;
- $\mathscr{L}(X)$ est la vraisemblance des N trames des 500 locuteurs de l'ensemble $\mathscr{E}_3$ à l'itération i ;
- α est un facteur de pénalisation ($= 1$).

La figure 8.2 montre que le BIC atteint des valeurs maximales pour des espaces de 200 à 350 locuteurs virtuels. Dans cet intervalle, les taux d'erreur sont les plus bas (voire figure 8.1).
La valeur maximale correspond à un espace de 250 locuteurs virtuels. Bien que ce calcul soit une approximation grossière, cette valeur est confirmée lorsque l'espace est construit par regroupement ou par knock-out (paragraphe 8.2.1).

8.1.3 Post-traitement ACP/ALD

Le regroupement hiérarchique ascendant des locuteurs ne fournit pas un espace orthogonal, c'est pourquoi nous allons appliquer un post-traitement selon les modalités du paragraphe 7.2.3. L'expérience est réalisée avec les paramètres suivants :

- Les modèles de tous les locuteurs sont appris avec 256 gaussiennes.

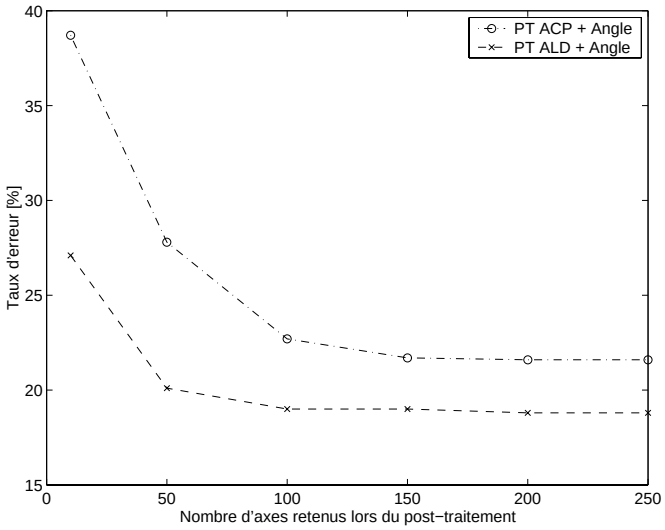


FIG. 8.4 – Regroupement hiérarchique : taux d'erreur en fonction du nombre d'axes retenus lors du post-traitement ACP/ALD

– L'identification des locuteurs se fait en appliquant l'angle entre les vecteurs des coordonnées transformées.

La matrice de covariances (pour l'ACP) et les matrices d'inter- et d'intra-classe (pour l'ALD) sont estimées à partir des 57 locuteurs de l'ensemble $\mathscr{E}_2$ complètement différents des 50 locuteurs à identifier (ensemble $\mathscr{E}_1$) et des 500 locuteurs qui ont servi à construire l'espace (ensemble $\mathscr{E}_3$).
Cette transformation permet de retrouver l'orthogonalité de l'espace et d'appliquer correctement les métriques.

La figure 8.3 trace les variations des taux d'erreur des 50 locuteurs en fonction de la quantité de données d'apprentissage dans les cas suivants :

– Aucun post-traitement n'est appliqué sur les vecteurs des coordonnées.
– Application d'une ACP sur les vecteurs des coordonnées.
– Application d'une ALD sur les vecteurs des coordonnées.

Dans cette manipulation, les post-traitements sont appliqués uniquement sur des espaces à 250 locuteurs (cf. paragraphe 8.1.1). Les locuteurs sont ensuite identifiés en appliquant l'angle entre les vecteurs des coordonnées.
La figure 8.3 donne un aperçu des variations des taux d'erreur que l'on peut attendre en fonction de la quantité de données. Cette figure montre que l'orthogonalisation ACP/ALD améliore les performances au-delà 10 secondes pour l'ALD et 20 secondes pour l'ACP.
Cette figure montre aussi que l'application de l'ALD donne des meilleures performances que l'application de l'ACP. En effet, l'ACP effectue une rotation pour retrouver l'ortho-

gonalité de l'espace ce qui permet d'appliquer correctement les métriques. En revanche, l'ALD effectue deux rotations qui rendent l'espace orthogonal et ses axes discriminants ce qui explique l'écart de performances entre l'ACP et l'ALD. Rappelons que ces classes de locuteurs ont été générées par un ensemble de locuteurs $\mathscr{E}_2$ totalement différent de l'ensemble de locuteurs $\mathscr{E}_1$ à identifier. Il y a donc une bonne généralisation des classes de locuteurs déterminées sur l'ensemble $\mathscr{E}_2$.
Par ailleurs, nous avons remarqué que l'augmentation des performances est peu significative au-delà de 57 axes. Ceci est dû au fait que nous ne disposons que de 57 classes de locuteurs de l'ensemble $\mathscr{E}_2$ (voir figure 8.4).
Notons que l'approche de GMM-UBM fournit des meilleures performances.

8.2 Evaluation de la localisation par les modèles d'ancrage dans un espace construit par sélection

Dans cette expérience, les étapes du système de reconnaissance sont :

Construction de l'espace par knock-out. Nous recherchons un sous-ensemble de locuteurs les plus dispersés parmi les 500 de l'ensemble $\mathscr{E}_3$. La procédure du knock-out est illustrée dans la figure 6.2. A chaque itération, nous évaluons la mesure de dispersion 6.30 sur tous les sous-ensembles et nous éliminons le locuteur sans lequel la dispersion est maximale. Le processus est ré-itéré jusqu'à n'avoir que E locuteurs sélectionnés.

Localisation et décision. L'espace représentatif des locuteurs n'est pas orthogonal, la localisation des locuteurs est réalisée par les modèles d'ancrage. Nous appliquons ensuite des métriques pour évaluer la proximité entre locuteurs.

Le but de cette manipulation est d'évaluer les performances de reconnaissance par localisation dans un espace construit par sélection et l'impact de différents paramètres tels que la taille de l'espace, la quantité de donnée d'apprentissage, etc.

8.2.1 Influence de la taille de l'espace et de la métrique

Dans ce paragraphe, nous allons étudier l'influence de la taille de l'espace dans les conditions suivantes :

- Les locuteurs de l'espace (ensemble $\mathscr{E}_3$) sont modélisés par 256 gaussiennes.
- La localisation des locuteurs est réalisée par les modèles d'ancrage et l'apprentissage se fait avec 100 secondes de parole.

La figure 8.5 représente les taux d'erreur d'identification des 50 locuteurs en fonction de la taille de l'espace et pour différentes métriques. Comme sur la figure 8.1, l'angle fournit toujours les meilleures performances d'identification et sa courbe de variations diminue progressivement jusqu'à atteindre une valeur optimale de 26% de taux d'erreur lorsque l'espace est constitué de 250 locuteurs. Au-delà de cette valeur, les taux d'erreur croissent et tendent vers les 32.3% d'identification incorrecte.

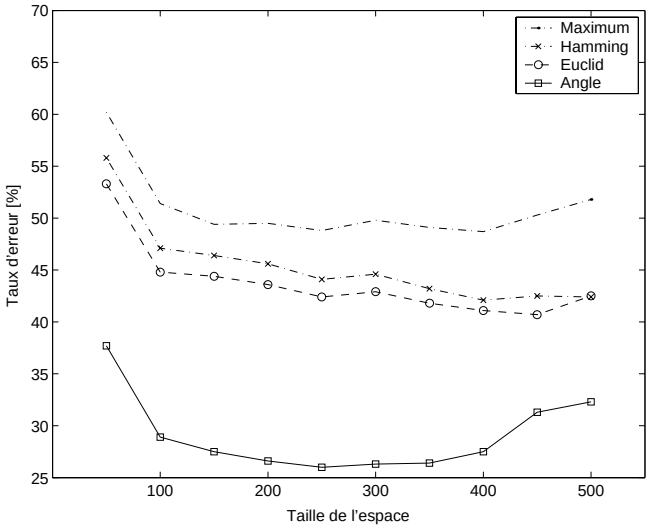


FIG. 8.5 – Knock-out : taux d'erreur en fonction de la taille de l'espace

Analyse des 250 locuteurs sélectionnés

La figure 8.6 donne une idée sur la composition sexe/âge des 250 locuteurs sélectionnés. Sur les 219 femmes et 281 hommes, nous avons sélectionné 127 femmes et 123 hommes. Quant à la composition d'âge, les 250 locuteurs sélectionnés respectent bien la répartition initiale des 500 locuteurs.

8.2.2 Post-traitement ACP/ALD

La sélection par knock-out ne fournit pas un espace propre. Afin de retrouver l'orthogonalité de l'espace et d'appliquer correctement les métriques, nous avons effectué (comme dans le paragraphe 8.1.3) une orthogonalisation des coordonnées des locuteurs (application d'une ACP ou d'une ALD). L'expérience est réalisée dans les conditions suivantes :

- Les modèles de tous les locuteurs sont appris avec 256 gaussiennes.
- L'identification des locuteurs se fait en appliquant l'angle entre les vecteurs des coordonnées transformées.

Nous rappelons aussi que les matrice de covariances, d'inter- et d'intra-classe sont estimées à partir des 57 locuteurs de l'ensemble $\mathscr{E}_2$ complètement différents des 50 locuteurs à identifier (ensemble $\mathscr{E}_1$) et des 500 locuteurs qui ont servi à construire l'espace (ensemble $\mathscr{E}_3$).

La figure 8.7 trace les variations des taux d'erreur d'identification des 50 locuteurs en fonction de la quantité de données d'apprentissage dans les cas suivants :

- Aucun post-traitement n'est appliqué sur les vecteurs des coordonnées.
- Application d'une ACP sur les vecteurs des coordonnées.

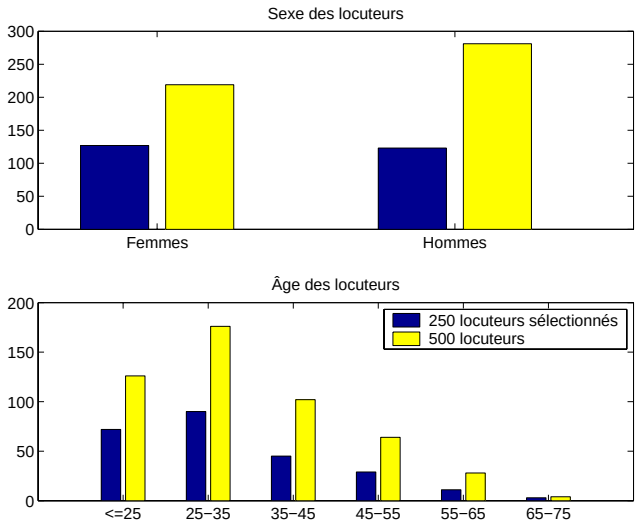


FIG. 8.6 – Knock-out : composition de l'ensemble des 250 locuteurs sélectionnés

– Application d'une ALD sur les vecteurs des coordonnées.

L'application de l'ACP et de l'ALD s'effectue sur un espace construit par knock-out à 250 locuteurs. Ainsi, l'observation de la figure 8.7 donne des résultats à peu près similaires à ceux obtenus au paragraphe 8.1.3. Nous observons d'une part que l'application du post-traitement permet d'améliorer les performances d'identification notamment si les données disponibles sont supérieures à 8 secondes. D'autre part, l'application de l'ALD donne des meilleures performances que l'application de l'ACP quelle que soit la quantité de données disponible. Notons aussi que l'approche de GMM-UBM fournit des meilleures performances.

8.2.3 Sélection du sous-ensemble de locuteurs les plus proches

Dans les paragraphes précédents, nous avons recherché le sous-ensemble des locuteurs les plus dispersés. Nous nous sommes appuyés sur l'hypothèse qui présume que plus les locuteurs sont dispersés, meilleur est l'espace de représentation. Afin de vérifier cette hypothèse a contrario, nous nous sommes proposés de rechercher les locuteurs les plus proches (parmi les 500 de l'ensemble $\mathscr{E}_3$), c'est à dire ceux qui minimisent la dispersion et qui maximisent "la compacité". La procédure de sélection utilisée est, à peu près, l'inverse de la procédure de sélection du sous-ensemble le plus dispersé : à chaque itération, nous éliminons le locuteur sans lequel la dispersion est minimale. Les locuteurs sélectionnés fournissent le nouvel espace de représentation. Nous avons réalisé cette expérience dans les conditions suivantes :

– Les locuteurs sont modélisés avec 256 gaussiennes.

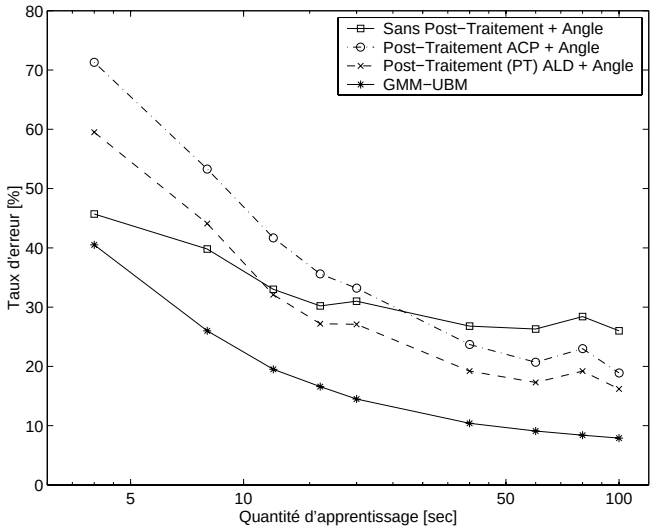


FIG. 8.7 – Knock-out : taux d'erreur en fonction de la quantité de données d'apprentissage

- La localisation à l'apprentissage est réalisée par les modèles d'ancrage avec 100 secondes de parole.
- L'identification des locuteurs se fait en utilisant la mesure de l'angle entre leurs vecteurs de coordonnées.

Sur la figure 8.9, nous avons représenté les taux d'erreur dans les trois cas suivants :

- Sélection des locuteurs les plus dispersés.
- Sélection des locuteurs les plus proches.
- Sélection aléatoire des locuteurs (plusieurs tirages aléatoires).

Nous observons un écart significatif des performances entre la sélection des locuteurs les plus dispersés et la sélection des locuteurs les plus proches. La sélection aléatoire donne des performances intermédiaires entre les deux. Ce résultat est très important car il montre qu'effectivement plus les locuteurs sont dispersés, meilleur est l'espace de représentation.

8.3 Sélection d'un sous-ensemble de voisins propres à chaque locuteur

Dans les expériences précédentes, nous avons étudié un espace représentatif des locuteurs construit par sélection. Dans cet espace, tous les locuteurs avaient le même voisinage. Une autre façon de procéder est de sélectionner un voisinage propre à chaque locuteur : tous les locuteurs n'ont pas le même espace de représentation. Pour chaque locuteur à identifier, nous sélectionnons N plus proches voisins parmi les 500 locuteurs de l'ensemble $\mathscr{E}_3$. Cela revient à un espace de représentation propre à chaque locuteur. Ainsi, les étapes de la manipulation sont les suivantes :

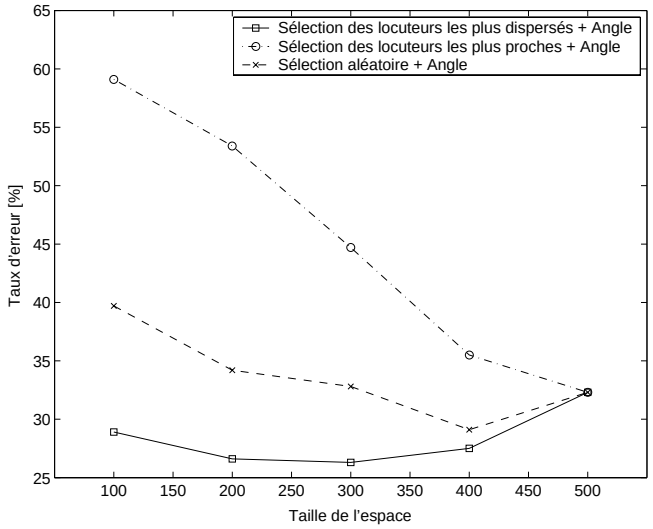


FIG. 8.8 – Sélection d'un sous-ensemble de locuteurs les plus proches : taux d'erreur en fonction de la taille de l'espace

1. Localiser (à l'apprentissage) par les modèles d'ancrage.

2. Identifier les N plus proches voisins selon la métrique de l'angle.

3. Faire le test uniquement avec ces N plus proches voisins.

Cette expérience est réalisée dans les conditions suivantes :
- Les locuteurs sont modélisés avec 256 gaussiennes ;
- leur localisation à l'apprentissage est réalisée par les modèles d'ancrage avec 100 secondes de parole.

Nous avons tracé sur la figure 8.9 les variations des taux d'erreur en fonction de la taille de l'espace. Lorsque nous sélectionnons 500 N-best, nous retrouvons évidemment les performances du knock-out à 500 locuteurs.
Cette figure montre que la sélection d'un sous-ensemble de voisins propres à chaque locuteur permet d'obtenir des bonnes performances voire meilleures que celles obtenues avec l'espace initial à condition d'en garder suffisamment. En effet, il est plus intéressent de sélectionner 350 locuteurs ou voisins propres que de conserver les 500 de l'espace initial.
Cependant, la localisation par le voisinage proche est toujours moins bonne que par le voisinage commun. Pour caractériser un locuteur, il est intéressant, non seulement de savoir de qui il est proche, mais également de qui il est éloigné.

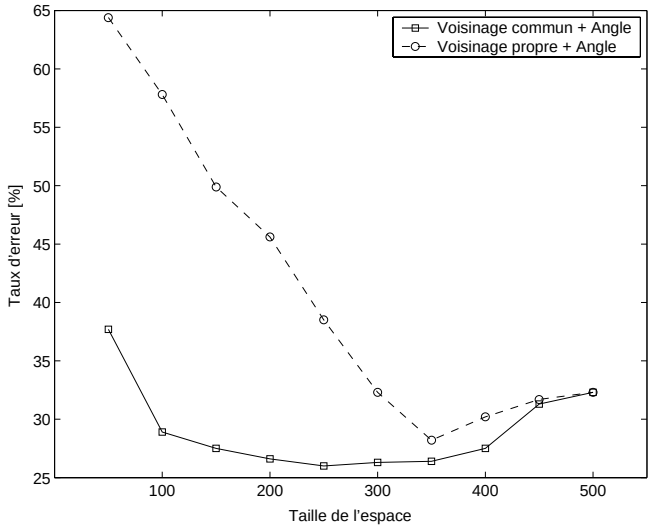


FIG. 8.9 – Sélection d'un sous-ensemble de voisins propres à chaque locuteur : taux d'erreur en fonction des N plus proches voisins

8.4 Conclusion

Dans ce chapitre, nous avons évalué les performances d'identification du locuteur par localisation dans un espace de référence. Les séries d'évaluations sur la base de France Télécom ont permis de tirer les conclusions suivantes :

- L'angle semble la métrique la plus discriminante pour évaluer la proximité entre les locuteurs.
- Que l'espace de locuteurs de référence soit construit par regroupement ou par sélection, il existe toujours un espace (à 250 locuteurs) meilleur que l'espace initial (à 500 locuteurs). Une estimation par un critère d'information bayésienne confirme ces résultats.
- L'espace de locuteurs de référence construit par sélection est meilleur que celui construit par regroupement hiérarchique.
- Le post-traitement ACP/ALD apporte une nette amélioration sur les performances d'identification. Le post-traitement ALD est meilleur que le post-traitement ACP. L'ALD permet une bonne généralisation des classes de locuteurs (obtenues à partir d'un corpus de développement).
- Les locuteurs les plus représentatifs sont les locuteurs les plus dispersés (selon un critère de dispersion).

Les modèles d'ancrage permettent de réduire le nombre de paramètres pour caractériser un locuteur (par un simple vecteur de distances par rapport aux E locuteurs de référence) mais les performances demeurent encore limitées. L'orthogonalisation par ALD a amélioré les

performances des systèmes mais elles restent insuffisantes (par rapport aux GMM-UBM).

Chapitre 9

Représentation compacte des locuteurs par distribution sur les modèles d'ancrage

Dans le chapitre précédent nous avons étudié des systèmes d'identification de locuteurs par localisation. En général, nous représentons les locuteurs par des points dans l'espace et nous évaluons la proximité entre eux. L'inconvénient majeur de cette approche géométrique est qu'elle accorde une place symétrique à l'apprentissage et au test, alors qu'en pratique, il existe souvent une asymétrie entre les occurrences d'apprentissage et de test.
Dans ce chapitre, nous proposons une nouvelle représentation des locuteurs basée sur une distribution de distances. L'idée est de modéliser un locuteur par une distribution sur les "distances" mesurées dans l'espace des modèles d'ancrage. Cette idée permet de tenir compte de l'asymétrie entre les données d'apprentissage et de test, l'apprentissage servant à estimer les paramètres de la distribution. En outre, on peut alors introduire des connaissances a priori dans l'estimation des paramètres.
Dans les premiers paragraphes, nous introduisons le principe de la représentation compacte par distribution de distances. Ensuite, nous détaillons la phase d'estimation des modèles. Nous appliquons finalement cette nouvelle représentation en identification et en vérification du locuteur.

9.1 Principe de la représentation compacte par distribution de distances

Pour qu'un système à base de GMM-UBM puisse fonctionner en temps réel, il faut toujours faire un compromis entre les performances, la quantité de données nécessaire à la modélisation et la complexité des modèles, c'est-à-dire le nombre de paramètres nécessaires. A titre d'exemples, si l'espace acoustique est de dimension 42, avec 256 gaussiennes

le nombre de paramètres est de 21760 (42×256 moyennes + 42×256 variances + 256 pondérations) et à 512 gaussiennes, il est de 43520[1].
Les modèles d'ancrage permettent de réduire le nombre de paramètres pour caractériser un locuteur (par un simple vecteur de distances par rapport aux E locuteurs de référence) mais les performances demeurent encore limitées.

Le but de cette nouvelle représentation est de *représenter un locuteur par une densité de probabilité qui modélise ses distances à un ensemble de locuteurs de référence*. En d'autres termes, au lieu de localiser un locuteur par un seul point dans l'espace de représentation, il est localisé par une distribution : c'est une référence statistique au lieu d'être géométrique (voir figure 9.1).

Dans un tel système, nous conservons la représentation compacte des modèles d'ancrage et nous introduisons une densité de probabilité. Ce qui permettra d'une part, d'utiliser des informations a priori pour la modélisation et d'autre part, d'appliquer une mesure statistique entre l'occurrence de test et les modèles des locuteurs à reconnaître (au lieu d'une mesure géométrique).
En pratique, cela consiste à représenter un locuteur λ, pour lequel on dispose d'un ou de plusieurs segments de parole, par :

$$\lambda = \mathscr{N}(\mu^d, \Sigma^d) \tag{9.1}$$

où $\mathscr{N}$ est une distribution gaussienne (dont les paramètres seront expliqués plus loin) de moyenne μ^d et de covariance Σ^d. Ces paramètres sont estimés dans l'espace des coordonnées (ou l'espace des distances) à un ensemble de locuteurs de référence (figure 9.1).

9.2 Espace de représentation

Le calcul de cette représentation est off-line, il n'est donc pas contraint en temps de calcul. Il est conseillé d'utiliser la meilleure distance inter-locuteur possible. Ici, nous utilisons comme distance aux locuteurs de référence le score de vraisemblance entre les vecteurs acoustiques du segment considéré et les modèles GMM des locuteurs de référence :

$$w = \begin{bmatrix} \tilde{p}(X|\overline{\lambda}_1) \\ \tilde{p}(X|\overline{\lambda}_2) \\ \vdots \\ \tilde{p}(X|\overline{\lambda}_E) \end{bmatrix}$$

où E est le nombre des locuteurs de référence sélectionnés et $\tilde{p}(X|\overline{\lambda}_e)$ est un score de vraisemblance normalisé défini dans l'équation 7.7.

[1] Dans le cas où tous les paramètres du modèle GMM-UBM sont appris et les matrices de covariance sont diagonales.

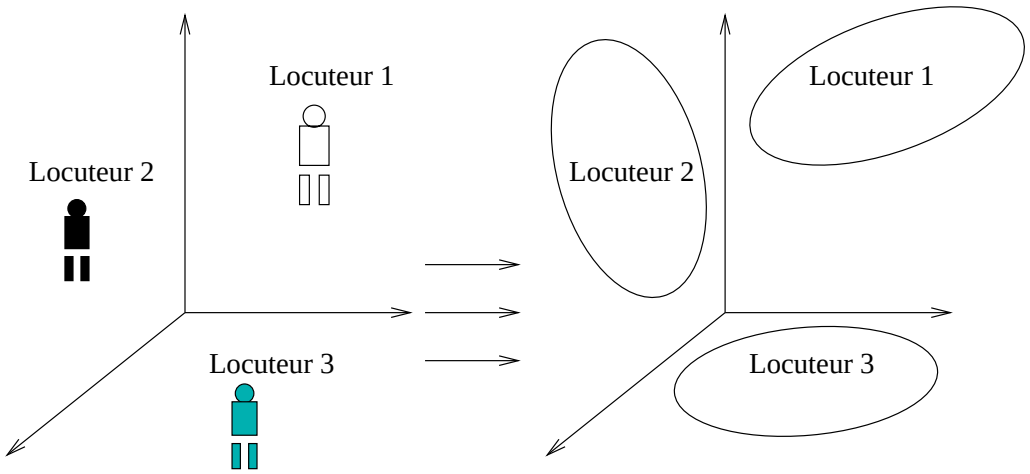


FIG. 9.1 – Représentation relative des locuteurs

Précisons que tout autre calcul de distance inter-locuteur peut également être utilisé.
Ainsi, *dans cette représentation, les données d'apprentissage ou de test ne représentent plus des vecteurs acoustiques mais des vecteurs de coordonnées*[2] *de dimension* E.
Afin de différencier les paramètres estimés dans l'espace acoustique des paramètres estimés dans l'espace des distances, ces derniers seront indicés par l'indice d, soit $\mathscr{X}^d$ où $\mathscr{X}$ représente les vecteurs des coordonnées, le vecteur des moyennes, la matrice de covariance, etc.

9.3 Estimation des paramètres du modèle de locuteur

Soit un locuteur λ pour lequel on dispose de N segments de parole. Ces segments sont représentés dans l'espace des distances par N vecteurs de coordonnées, soit :

$$W^d = \left(\begin{array}{ccc} w_1^\lambda & \cdots & w_N^\lambda \end{array} \right)$$

où W^d représente l'ensemble de N vecteurs de coordonnées du locuteur λ et le vecteur des distances w_j^λ du segment j s'écrit :

$$w_j^\lambda = \left[\begin{array}{c} \widetilde{p}(X_j^\lambda|\overline{\lambda}_1) \\ \vdots \\ \widetilde{p}(X_j^\lambda|\overline{\lambda}_E) \end{array} \right]$$

[2]Les vecteurs des coordonnées sont calculés à partir des vecteurs acoustiques (vraisemblance des vecteurs acoustiques sur les modèles d'ancrage).

9.3.1 Cas mono-gaussien

Une première possibilité pour l'estimation des paramètres de $\mathcal{N}(\mu^d, \Sigma^d)$ (équation 9.1)[3] consiste à les estimer par maximum de vraisemblance, soit pour chaque locuteur :

$$\mu_i^d = \frac{1}{N} \sum_{j=1}^{N} W_{ij}^d \tag{9.2}$$

et

$$\Sigma_{ii'}^d = \frac{1}{N} \sum_{j=1}^{N} (W_{ij}^d - \mu_i^d)(W_{i'j}^d - \mu_{i'}^d) \tag{9.3}$$

où $i, i' = 1, \cdots, E$ et W_{ij}^d est la distance du segment de parole j (du locuteur λ) par rapport au locuteur de référence $\overline{\lambda}_i$.
Cependant, l'estimateur de vraisemblance n'est efficace que lorsque nous disposons de beaucoup de données ($N >>$). Dans le cas contraire (notamment dans des systèmes réels), le nombre de segments de parole disponibles n'est pas suffisant et l'estimation par maximum de vraisemblance n'est pas fiable.
Une possibilité pour remédier à ce problème est d'introduire de l'information a priori. Les paramètres du locuteur seront adaptés à partir des paramètres initiaux par MAP (maximum a posteriori) [Gauvain et Lee, 1994] et la formule de ré-estimation des moyennes est donnée pour chaque locuteur par l'équation suivante :

$$\widetilde{\mu}^d = \frac{N_0 \mu_0^d + N \mu^d}{N_0 + N} \tag{9.4}$$

où N_0 est un paramètre de contrôle qui permet de donner un poids à l'information a priori.

Estimation du modèle a priori

L'information a priori correspond à l'estimation d'une densité de probabilité des distances par rapport aux modèles d'ancrage, indépendamment du locuteur. Elle est estimée à partir d'un corpus de développement de S locuteurs. Ces données sont représentées dans l'espace des modèles d'ancrage : chaque segment de parole d'un locuteur est représenté par un vecteur de coordonnées (figure 9.1). La matrice de toutes les données initiales $\mathscr{W}^d$ correspond à l'expression suivante :

$$\mathscr{W}^d = \left(\begin{array}{cccccccc} w_1^{loc_1} & \cdots & w_{N_1}^{loc_1} & w_1^{loc_2} & \cdots & w_{N_2}^{loc_2} & \cdots \end{array} \right)$$

et

$$N_0 = \sum_{s=1}^{S} N_s$$

[3] Il s'agit de modéliser les N segments (de W^d) du locuteur λ et d'estimer ses paramètres μ^d et Σ^d. L'indice λ a été volontairement supprimé pour alléger les équations.

où les vecteurs de coordonnées w sont de dimension E et N_1, N_2, $\cdots$, N_S sont le nombre de segments du locuteur 1, locuteur 2, etc.
On obtient donc un nuage de points et on estime sa densité de probabilité par maximum de vraisemblance (chaque vecteur de coordonnées devient une réalisation de la gaussienne). Ce nuage est caractérisé par son vecteur moyennes μ_0^d et par sa matrice de covariance Σ_0^d. Cette densité de probabilité $\mathcal{N}(\mu_0^d, \Sigma_0^d)$ est considérée comme le modèle a priori pour maximiser la probabilité a posteriori des nouvelles données. L'estimation par maximum de vraisemblance du vecteur des moyennes donne :

$$\mu_{0\,i}^d = \frac{1}{N_0} \sum_{j=1}^{N_0} \mathscr{W}_{ij}^d \tag{9.5}$$

En ce qui concerne la matrice de covariance Σ_0^d , elle peut être estimée également par maximum de vraisemblance ou par MAP. Dans cette étude, nous utilisons une matrice de covariance intra-classes[4] des données initiales (donc estimée indépendamment des données du locuteur λ) :

$$\Sigma_{0\,ii'}^d = \frac{1}{N_0} \sum_{s=1}^{S} \sum_{j \in I_s^d} (\mathscr{W}_{ij}^d - \overline{\mathscr{W}}_{is}^d)(\mathscr{W}_{i'j}^d - \overline{\mathscr{W}}_{i's}^d) \tag{9.6}$$

où S est le nombre total des locuteurs initiaux. Chaque classe est caractérisée par un ensemble I_s^d de N_s segments de parole provenant d'un même locuteur, de moyenne :

$$\overline{\mathscr{W}}_{is}^d = \frac{1}{N_s} \sum_{j \in I_s^d} \mathscr{W}_{ij}^d \tag{9.7}$$

et $i, i' = 1, \cdots, E$.

Par ailleurs, l'équation 9.4 donne la formule d'estimation du vecteur des moyennes $\widetilde{\mu}^d$ du locuteur λ. En ce qui concerne la matrice de covariance $\widetilde{\Sigma}^d$, elle est commune à tous les locuteurs et correspond à la matrice intra-classes des données initiales c'est-à-dire :

$$\widetilde{\Sigma}^d = \Sigma_0^d \tag{9.8}$$

Il est certes possible d'estimer une nouvelle matrice de covariance pour chaque locuteur λ. Cependant, on choisit de donner à tous les locuteurs la même matrice de covariance et cela pour deux raisons principales :

- D'abord, des matrices de covariance propres à chaque locuteur risquent d'être mal estimées.
- D'autre part, cette matrice (unique pour tous les locuteurs) peut être orthogonalisée en off-line ce qui réduit considérablement le coût de calcul, ensuite utilisée dans la phase de test où l'on calcule des densités de probabilité avec gaussiennes à matrice de covariance diagonale.

[4]Ce choix a été vérifié expérimentalement. En effet, une matrice intra-classe donne des meilleures performances qu'une matrice estimée par MV ou par MAP.

9.3.2 Cas multi-gaussien

Cette démarche peut être étendue au cas multi-gaussien pour la densité initiale. En effet, le nuage de points des données initiales (de l'ensemble S) peut être modélisé par plusieurs gaussiennes M^d. L'obtention d'une densité multi-gaussiennes peut se faire de plusieurs façons :

- En utilisant une connaissance annexe (e.g. sexe des locuteurs, pour apprendre deux gaussiennes, une pour les femmes et une pour les hommes).
- En utilisant directement les données. Par exemple, on peut apprendre les M^d gaussiennes du nuage par éclatement. A partir de la moyenne empirique μ_0^d, on dédouble la gaussienne initiale et on forme deux gaussiennes de moyennes $(\mu_0^d - \epsilon)$ et $(\mu_0^d + \epsilon)$ qu'on estime ensuite par Maximum de Vraisemblance. Ce processus est ré-itéré jusqu'à la convergence de la vraisemblance totale du mélange et pour un nombre M^d de gaussiennes.

Ensuite, la phase d'adaptation consiste à déterminer la meilleure gaussienne de l'ensemble a priori et l'adapter aux nouvelles données du locuteur λ. Ainsi, la formule de ré-estimation des moyennes, pour un locuteur λ, est donnée par :

$$\widetilde{\mu}^d = \frac{N_0 \widehat{\mu}_0^d + N \mu^d}{N_0 + N} \tag{9.9}$$

On définit un paramètre de contrôle normalisé α qui donne un poids à l'information a priori :

$$\alpha = \frac{N_0}{N_0 + N} \tag{9.10}$$

La formule de ré-estimation des moyennes devient :

$$\widetilde{\mu}^d = \alpha \widehat{\mu}_0^d + (1 - \alpha) \mu^d \tag{9.11}$$

où $\widehat{\mu}_0^d$ correspond, cette fois-ci, à la gaussienne qui donne le meilleur score de vraisemblance, soit :

$$\widehat{\mu}_0^d = \arg \max_{\mu_m} p(W^d | \mu^d, \Sigma^d) \tag{9.12}$$

et $m = 1, \cdots, M^d$.
Il s'agit toujours d'une mono-gaussienne pour représenter un locuteur mais adaptée à partir d'une gaussienne initiale choisie dans un ensemble de gaussiennes.

9.4 Application à l'identification et la vérification du locuteur

Quel que soit l'application ou la tâche visée, le module de décision est basé sur les deux processus classiques d'identification et/ou de vérification de locuteur (cf. paragraphe 1.4).

9.4.1 Identification du locuteur

En identification du locuteur, la phase de test consiste à évaluer une mesure de vraisemblance entre les coordonnées du segment de test w_X^d et l'ensemble des nouveaux modèles de locuteurs à identifier. Ainsi le locuteur reconnu correspond à :

$$\widehat{\lambda} = \arg\max_{\lambda} p(w_X^d | \widetilde{\mu}^d, \widetilde{\Sigma}^d) \tag{9.13}$$

Cette mesure de similarité est nettement plus robuste que les distances utilisées dans le chapitre précédent 8 puisqu'on évalue une similarité statistique et non pas géométrique.

9.4.2 Vérification du locuteur

Le score obtenu sur le modèle du locuteur prétendu est normalisé par le score obtenu sur le modèle du monde qui correspond à la distribution indépendante du locuteur (distribution initiale dans le cas mono-gaussien) :

$$score = \frac{p(w_X^d | \widetilde{\mu}^d, \widetilde{\Sigma}^d)}{p(w_X^d | \widetilde{\mu}_0^d, \widetilde{\Sigma}_0^d)} \tag{9.14}$$

où $X = \{\text{abonnés, imposteurs}\}$.

9.5 Evaluation de l'identification et de la vérification du locuteur par distribution sur les modèles d'ancrage

Nous avons évalué cette nouvelle représentation en identification et en vérification du locuteur sur les 50 locuteurs de l'ensemble $\mathscr{E}_1$. Les locuteurs initiaux correspondent aux 57 locuteurs du corpus de développement $\mathscr{E}_2$ (complètement différents des 50 locuteurs)[5]. Les locuteurs sont représentés par des vecteurs de distances de dimension 250.
Sur la figure 9.2, nous avons représenté les variations des taux d'erreur en fonction des valeurs de α (cf. équation 9.10), pour plusieurs quantités d'apprentissage et dans le cas où nous avons une distribution a priori choisie parmi 2 ou 4 gaussiennes.
Pour chaque valeur de quantité d'apprentissage, il existe une valeur optimale de α. En général, ce paramètre est grand quand nous disposons de peu de données et approche la valeur de 0.5 dans les cas contraires. Le tableau 9.1 donne un aperçu sur les valeurs de α en fonction de la quantité d'apprentissage.

Cette figure montre que l'introduction des connaissances a priori permet d'apporter une nette amélioration par rapport à une distribution sans a priori ($\alpha = 0$ ou bien $\widetilde{\mu}^d = \mu^d$).

[5] Il s'agit des locuteurs qui servent à estimer la distribution initiale.

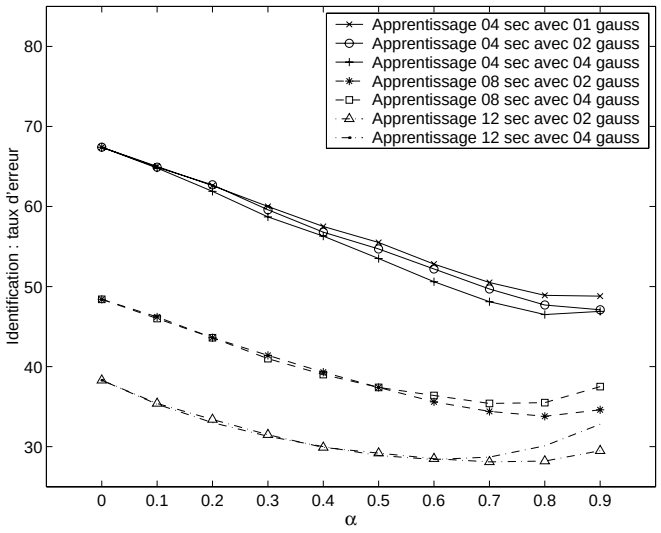


FIG. 9.2 – Performances d'identification : influence du nombre de gaussiennes

Dans le cas où $\alpha = 1$ (ou $\widetilde{\mu}^d = \widehat{\mu}_0$), tous les locuteurs auraient quasiment le même modèle (choisi parmi l'ensemble des distributions initiales).
Le choix de la distribution a priori parmi 4 gaussiennes améliore les résultats uniquement dans le cas où nous disposons de très peu de données (4 secondes de parole, voire figure 9.2).

Apprentissage	04 sec.	08	12	16	20	40	60	80	100
α	0.9	0.8	0.7	0.6	0.6	0.5	0.5	0.5	0.5

TAB. 9.1 – Valeurs optimales de α pour chaque quantité de données d'apprentissage (avec choix de la distribution a priori parmi 04 gaussiennes)

Sur les figures 9.3 et 9.4, nous avons représenté les variations des taux d'erreur ainsi que les taux EER des 50 locuteurs de l'ensemble $\mathscr{E}_1$, en fonction de la quantité d'apprentissage. Ces figures donnent un aperçu des variations des taux d'erreur que l'on peut attendre en fonction de la quantité de données d'apprentissage. Les taux d'identification incorrecte décroissent significativement jusqu'à atteindre la valeur de 16.1% pour 100 secondes de parole.
Quant à l'authentification, l'EER varie entre 13.4% pour 4 secondes de parole et 4.8% pour 100 secondes.
Ces figures montrent aussi que la distribution sur les modèles d'ancrage apporte une amélioration sur l'approche géométrique des modèles d'ancrage notamment dans le cas où nous disposons de peu de données d'apprentissage.

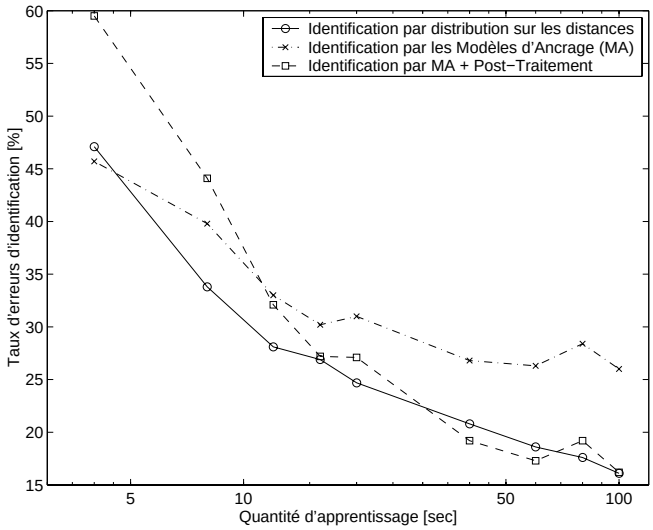


FIG. 9.3 – Performances d'identification par distribution sur les modèles d'ancrage (avec choix de la distribution a priori parmi 04 gaussiennes) et par les modèles d'ancrage (avec et sans post-traitement)

Bien que la phase d'apprentissage de la nouvelle technique puisse s'avérer plus importante en temps de calcul, la phase de test, quant à elle, est nettement moins complexe que celle du GMM-UBM. En effet, plusieurs opérations complexes sont réalisées off-line notamment la représentation par des vecteurs de distances des modèles des locuteurs à reconnaître et des données initiales et l'orthogonalisation de la matrice de covariance commune à tous les locuteurs.

A titre d'exemple, pour une occurrence de test de 300 trames (approximativement 5 secondes de parole), le nombre d'opérations élémentaires durant chaque test GMM-UBM à 256 gaussiennes est[6] :

$$300 \times 256 \times \mathscr{N}_{42} = 3.2\ 10^6 \times \mathscr{N}_1$$

Si le nombre des locuteurs de référence est 250 dans la nouvelle représentation, le nombre d'opérations élémentaires est de :

- $250 \times 300 \times 256 \times \mathscr{N}_{42} = 806\ 10^6 \times \mathscr{N}_1$ pour la localisation de l'occurrence de test.
- $1 \times \mathscr{N}_{250} = 250 \times \mathscr{N}_1$ pour chaque comparaison avec une référence.

Si K est le nombre des occurrences de référence avec lesquelles il faut comparer l'occurrence de test, alors le nombre d'opérations total est :

- GMM-UBM : $3.2\ 10^6 \mathscr{N}_1 \times K$
- Nouvelle représentation : $806\ 10^6 \mathscr{N}_1 + 250 \mathscr{N}_1 \times K$

[6] Ce calcul est approximatif sachant qu'il y a 42 coefficients cepstraux par vecteur acoustique.

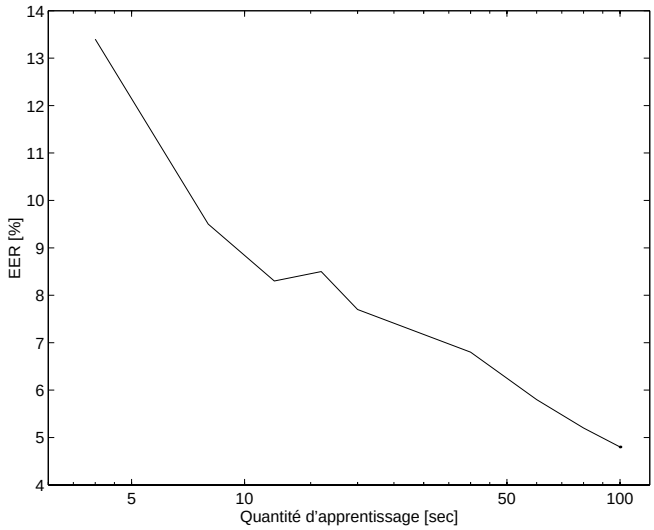


FIG. 9.4 – Performances de vérification par distribution sur les modèles d'ancrage (avec choix de la distribution a priori parmi 04 gaussiennes)

Dans l'approche relative, la localisation de l'occurrence de test coûte cher, mais ensuite, la comparaison avec des occurrences de référence (préalablement localisées) est très peu chère. C'est donc une approche intéressante du point de vue calculatoire lorsqu'on doit comparer l'occurrence de test à des très nombreuses occurrences. Une application intéressante de cette nouvelle représentation est l'indexation des grandes bases de documents audio.

9.6 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle représentation des locuteurs basée sur une distribution de distances. Le but principal était de tenir compte de l'asymétrie entre les données d'apprentissage et de test. Cela a permis également d'introduire des connaissances a priori dans l'estimation des paramètres.
Nous avons appliqué cette représentation en identification et en vérification du locuteur. Les évaluations de cette nouvelle technique ont montré que la distribution sur les modèles d'ancrage apporte une amélioration sur l'approche géométrique notamment dans le cas où nous disposons de peu de données d'apprentissage. L'introduction des connaissances a priori permet d'améliorer significativement les résultats. Il existe des valeurs optimales du poids de l'a priori pour chaque quantité d'apprentissage. Par ailleurs, le choix de la distribution a priori parmi plusieurs gaussiennes améliore les performances mais le gain est peu significatif au-delà de 4 gaussiennes.

Chapitre 10

Ré-estimation des modèles par sélection de voisins

Dans les chapitres précédents, nous avons étudié comment la position d'un locuteur par rapport à un ensemble de locuteurs de référence pouvait être utilisée directement pour modéliser le locuteur. Dans ce chapitre, nous nous intéressons au cas où ce positionnement relatif sert à l'estimation d'un modèle du locuteur, dans la modélisation GMM. Cette approche, qui consiste à sélectionner des locuteurs voisins et à fusionner leurs modèles pour donner un nouveau modèle GMM du locuteur, est originale.

10.1 Principe de la ré-estimation par sélection

Dans le paragraphe 5.1, nous avons présenté quelques techniques d'adaptation rapide utilisant des connaissances a priori obtenues à partir d'un ensemble de locuteurs de référence telles que la RMP, EMAP, RSW et le clustering des locuteurs.
Dans ce chapitre, la position relative d'un locuteur par rapport à un ensemble de locuteurs de référence est exploitée pour l'estimation (ou la ré-estimation) de son modèle GMM [Mami et Charlet, 2003b]. Cette approche fait partie des techniques d'adaptation rapide basées sur le clustering des locuteurs, qu'on peut classer en deux grandes familles :

- La première approche consiste à construire, en off-line, les clusters des locuteurs et à affecter ensuite chaque locuteur au plus proche cluster. Son modèle est adapté à partir de ces clusters [Padmanabhan et Nahamoo, 1998].
- La deuxième approche consiste à créer, on-line, le cluster autour du locuteur qu'on souhaite modéliser.

Dans la présente, c'est la deuxième approche que nous allons étudier et appliquer à la ré-estimation des modèles GMM des locuteurs. En pratique, cela fonctionne en deux étapes :

- Déterminer durant la phase d'apprentissage un ensemble des plus proches voisins.
- Fusionner les modèles des voisins.

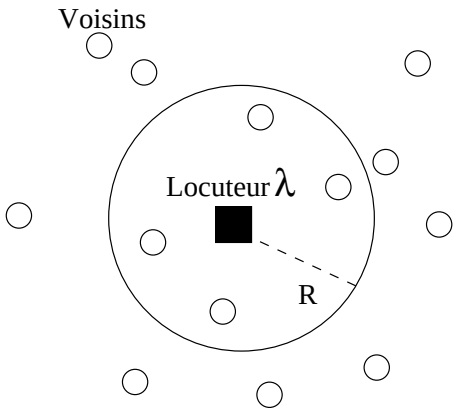


FIG. 10.1 – Détermination des voisins dans un rayon fixe

Par cette démarche, nous espérons hériter quelques connaissances pour la modélisation que nous ne pouvons pas estimer avec le peu de données disponibles (par exemple les variances). Cela suppose que les voisins sont sélectionnés avec fiabilité même avec peu de données.

10.2 Sélection des voisins

Pour chaque locuteur λ, les plus proches voisins sont déterminés durant la phase d'apprentissage en évaluant un score de vraisemblance des données X du locuteur par rapport aux modèles des E locuteurs de référence $\overline{\lambda}_i$:

$$d = -\log p(X|\overline{\lambda}_{i=1,\cdots,E})$$

où d est compatible avec une distance.
Nous pouvons ensuite sélectionner un ensemble des plus proches voisins par deux façons différentes :

- Sélection des V plus proches voisins selon leurs distances.
- Sélection des voisins dans un rayon fixe R (figure 10.1). Les voisins ainsi sélectionnés ont des scores $\leq R$. On normalise au préalable les scores de la façon suivante :

$$d_i^{norm} = \frac{d_i - \min_{j=1,\cdots,E} d_j}{\max_{j=1,\cdots,E} d_j - \min_{j=1,\cdots,E} d_j}$$

avec :

$$d_i = -\log p(X|\overline{\lambda}_i)$$

En pratique, les locuteurs de référence $\overline{\lambda}_i$ correspondent aux locuteurs sélectionnés par knock-out ou par clustering (cf. chapitre 6).

10.3 Fusion des modèles des voisins

Après avoir sélectionné les V voisins les plus proches d'un locuteur $\{\overline{\lambda}_{k=1,\cdots,V}\}$, leurs modèles sont fusionnés pour donner un nouveau modèle du locuteur λ. Le modèle de fusion est estimé à partir des paramètres des modèles des voisins sélectionnés. Dans ce cas, les pondérations des gaussiennes ne sont pas ré-estimées, elles sont héritées du modèle initial GMM-UBM. Pour les autres paramètres, les formules de ré-estimation, pour une gaussienne m, sont les suivantes :

– Les moyennes :

$$\mu^m = \frac{1}{V}\sum_{k=1}^{V}\mu_k^m \tag{10.1}$$

– Les variances :

$$\sigma^{2\,m} = \frac{1}{V}\sum_{k=1}^{V}\left[\sigma_k^{2\,m} + (\mu_k^m)^2\right] - (\mu^m)^2 \tag{10.2}$$

D'évidence, cette fusion n'a de sens que si l'association des composantes du mélange dans l'espace acoustique est correcte. C'est-à-dire que la gaussienne m du modèle voisin A peut être associée à la gaussienne m du modèle voisin B.
Dans nos expériences, ce modèle est appelé NMM (pour *Neighborhood-Merged Model*).

10.4 Adaptation du modèle de fusion

Une fois le nouveau modèle du locuteur λ estimé par fusion des modèles des voisins, il est possible aussi de l'adapter par un apprentissage incrémental sauf que cette fois-ci, le modèle initial correspond au nouveau modèle NMM. Les formules de ré-estimation données dans le paragraphe 2.2.2 restent valables. Dans nos expériences, ce modèle est appelé NMAM (pour *Neighborhood-Merged and Adapted Model*).

10.5 Evaluation de l'identification et de la vérification du locuteur

10.5.1 Protocole expérimental

En identification du locuteur, les étapes de l'évaluation sont les suivantes :

– Apprentissage : génération des modèles de locuteur (de l'ensemble $\mathscr{E}_1$) soit par sélection de voisins ou bien par sélection et fusion des modèles de voisins.

– Test : le locuteur identifié $\widehat{s}$ correspond à

$$\widehat{s} = \arg\max_{1\leq s\leq \mathscr{S}} p(X|\lambda_s)$$

En vérification du locuteur, les étapes de l'évaluation sont les suivantes :

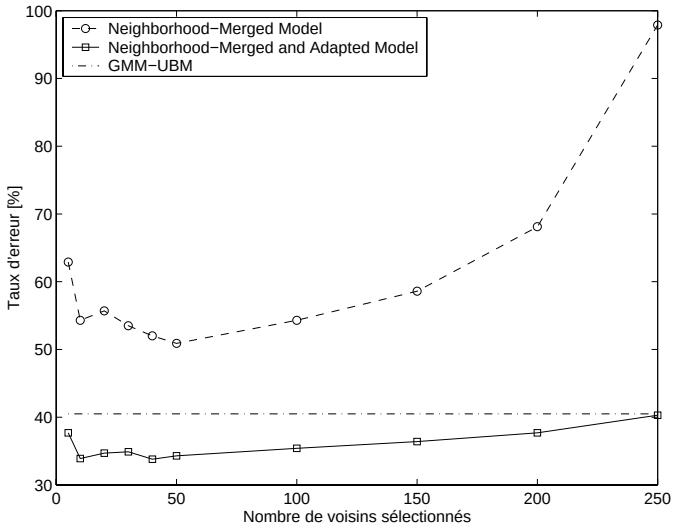


FIG. 10.2 – Performances d'identification du locuteur en fonction des voisins sélectionnés (pour 4 secondes d'apprentissage)

- Apprentissage : génération des modèles de locuteur (de l'ensemble $\mathscr{E}_1$) soit par sélection de voisins ou bien par sélection et fusion des modèles de voisins.
- Test : on évalue pour chaque locuteur le $LLR = \frac{1}{N}\log p(X|\lambda_s) - \log p(X|\lambda_{UBM})$ et on le compare à un seuil commun à tous les locuteurs.

où X est l'occurrence de test (de N trames acoustiques), λ_s est le modèle du locuteur s et $\mathscr{S}$ est le nombre des locuteurs à reconnaître.

10.5.2 Influence du nombre de voisins sélectionnés

Sur les figures 10.2 et 10.3, nous avons représenté les performances d'identification des 50 locuteurs de l'ensemble $\mathscr{E}_1$ en fonction du nombre de voisins sélectionnés et en fonction du rayon de voisinage, dans les deux cas suivants :

- Le modèle de fusion est estimé à partir des paramètres des voisins (*Neighborhood-Merged Model*).
- Le modèle de fusion est estimé à partir des paramètres des voisins et ensuite adapté par un apprentissage incrémental (*Neighborhood-Merged and Adapted Model*).

Les locuteurs sont modélisés par 256 gaussiennes et la quantité de donnée d'apprentissage est de 4 secondes. Les voisins sont déterminés parmi les 250 locuteurs sélectionnés par knock-out.

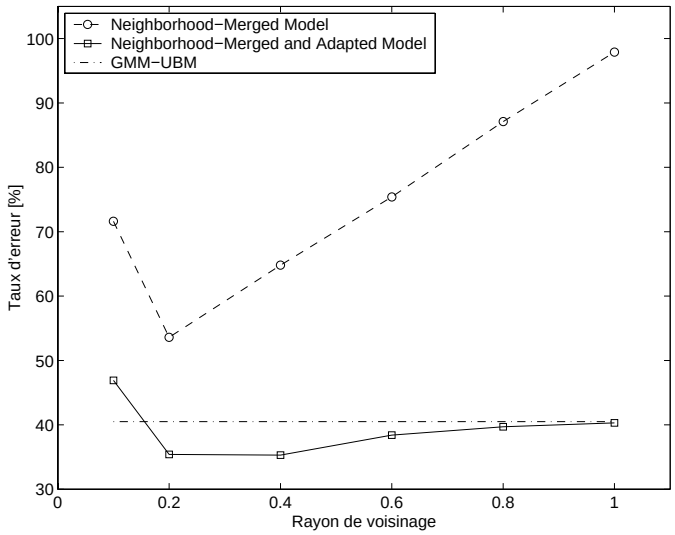


FIG. 10.3 – Performances d'identification du locuteur en fonction du rayon de voisinage (pour 4 secondes d'apprentissage)

Les figures 10.2 et 10.3 montrent que le voisinage caractérise bien le locuteur et cela même si les performances du modèle de fusion sont limitées.

Les deux figures montrent aussi que le modèle fusionné et adapté donne des meilleures performances que le modèle fusionné. Il existe une valeur optimale pour le nombre de voisins sélectionnés. En effet, le taux d'erreur d'identification du *Neighborhood-Merged and Adapted Model* atteint une valeur minimale de 33.8% pour 40 voisins. Le *Neighborhood-Merged and Adapted Model* est meilleur que le GMM-UBM et cela quel que soit le nombre de voisins sélectionnés.
La figure 10.3 montre que le taux d'erreur d'identification du *Neighborhood-Merged and Adapted Model* atteint une valeur minimale de 35.3% pour $R = 0.4$. Le tableau 10.1 donne un aperçu sur le nombre moyen des voisins sélectionnés dans un rayon donné.
Si tous les 250 voisins sont sélectionnés, les NMM auront le même modèle GMM-UBM (ce qui explique le taux d'erreur d'identification de 98% soit l'identification aléatoire d'un locuteur sur 50 !). En ce qui concerne les NMAM, ce modèle est considéré comme le nouveau modèle UBM des locuteurs.

Rayon de voisinage	0.2	0.4	0.6	0.8	1.0
Voisins	14	68	152	220	250

TAB. 10.1 – Nombre moyen des voisins pour un rayon de voisinage donné (pour 4 secondes d'apprentissage)

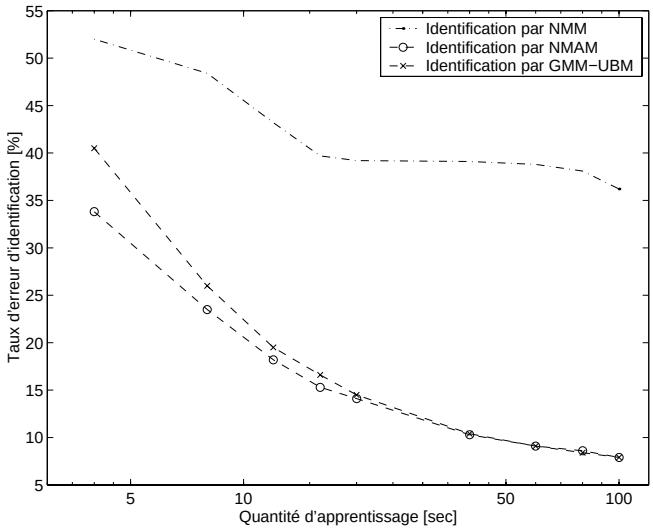


FIG. 10.4 – Neighborhood-merged and adapted model : performances d'identification du locuteur en fonction de la quantité de données d'apprentissage

10.5.3 Influence de la quantité de données d'apprentissage

Les figures 10.4 et 10.5 tracent les performances d'identification et de vérification des 50 locuteurs de l'ensemble $\mathscr{E}_1$ en fonction de la quantité de données d'apprentissage (pour un nombre de voisins de 40). Nous remarquons que nous avons de bonnes performances notamment quand nous disposons de peu de données : à 4 secondes de parole, le taux d'erreur d'identification est de 33.8% et le $EER = 16.4\%$ et à 100 secondes, le taux d'erreur d'identification est de 7.9% et le $EER = 4.2\%$.

Ces figures montrent aussi que le voisinage nécessite au moins 40 secondes de données d'apprentissage pour être fiable (où les performance de la ré-estimation des modèles par NMM (ou *Neighborhood-Merged Model*) atteignent les 40% du taux d'erreur). Cependant, la ré-estimation des modèles par NMAM (ou *Neighborhood-Merged and Adapted Model*) donne des meilleure performances d'identification que l'approche GMM-UBM : nous avons une bonne caractérisation du locuteur et cela malgré les problèmes de correspondances entre gaussiennes. En vérification du locuteur, les performances sont moins bonnes que celles du GMM-UBM. Cela est probablement dû à un problème de normalisation (le modèle du monde n'est plus le modèle initial).

Remarques sur les résultats présentés dans [Mami et Charlet, 2003b]

L'approche présentée dans ce chapitre a été présentée dans [Mami et Charlet, 2003b]. Cependant, les résultats des évaluations ne sont pas les mêmes. Dans cet article, nous

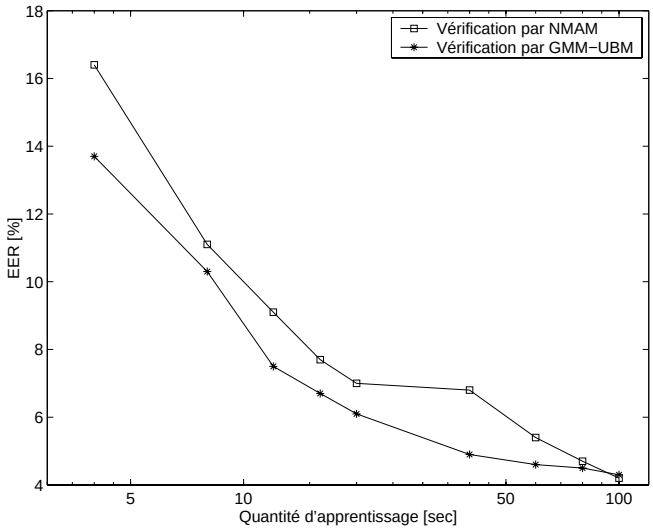


FIG. 10.5 – Neighborhood-merged and adapted model : performances de vérification du locuteur en fonction de la quantité de données d'apprentissage

avions utilisé une analyse acoustique différente : chaque vecteur acoustique était composé de l'énergie temporelle de la trame et des 8 premiers MFCC ainsi que leurs dérivées premières et secondes ; soit 27 coefficients acoustiques (au lieu de 42 dans notre nouvelle analyse acoustique). Dans [Mami et Charlet, 2003b], le voisinage caractérisait bien le locuteur mais dégradait significativement la couverture de l'espace acoustique. Dans ce chapitre, nous avons vu que la ré-estimation donne des résultats aussi bons voir meilleurs que le GMM-UBM. Cette nouvelle analyse acoustique permet d'obtenir des meilleurs modèles GMM des locuteurs et des voisins ce qui rend le voisinage meilleur et plus robuste.

10.6 Conclusion

Dans ce chapitre, nous avons présenté une technique de modélisation basée sur la sélection des plus proches voisins et la fusion de leurs modèles pour obtenir le nouveau modèle GMM du locuteur.
Les évaluations de l'identification et de la vérification du locuteur ont montré que le voisinage caractérise bien le locuteur. Nous avons montré que le modèle fusionné et adapté donne des meilleures performances que le modèle fusionné. En identification du locuteur, le modèle fusionné et adapté donne des meilleures performances que l'approche GMM-UBM. Par contre, en vérification du locuteur, le GMM-UBM est meilleur. Pour améliorer ces performances, plusieurs solutions peuvent être envisagées :

- Etudier la correspondance entre les composantes des gaussiennes (pour améliorer la couverture acoustique).

- Etudier le modèle de normalisation en vérification du locuteur et même introduire une normalisation supplémentaire telle que la *Z-norm*.
- Accorder des poids aux locuteurs lors de la fusion (étant donné que les voisins sont plus au moins proches).

Chapitre 11

Synthèse des résultats

Dans ce chapitre, nous allons évaluer un algorithme de fusion entre l'approche GMM-UBM et l'approche relative. Ensuite, nous présenterons et comparerons les performances de reconnaissance par GMM-UBM et par localisation.

11.1 Fusion de décision

La fusion de décision a pris une importance grandissante au cours de ces dernières années. La fusion de données ou le mélange d'experts est une partie du domaine de l'apprentissage automatique (*machine learning*) qui traite des problèmes complexes, plus particulièrement ceux qui nécessitent une prise de décision à partir de plusieurs sources d'informations distinctes.

11.1.1 Nature des erreurs

L'algorithme de fusion combine les décisions des systèmes d'identification du locuteur par GMM et par l'approche relative. Il peut aussi servir à combiner les décision de deux approches relatives. Dans les deux cas, les deux approches fournissent les N-meilleures solutions ordonnées selon leurs scores, et le locuteur identifié correspond au plus grand score calculé.
Cette démarche de fusion est justifiée par le fait que la nature des erreurs entre les deux approches est différente. A titre d'exemple, sur la figure 11.1, nous avons illustré les erreurs des deux approches GMM-UBM et distribution par les modèles d'ancrage (DMA). Nous avons représenté le nombre d'occurrences où les deux approches concordent à raison (identifient correctement le même locuteur), le nombre d'occurrences où les deux approches font la même erreur et le nombre d'occurrences qui restent.
D'évidence, les erreurs générées par la deuxième catégorie sont irrécupérables mais elles représentent seulement 9% de l'ensemble des occurrences de test. En revanche, les cas de

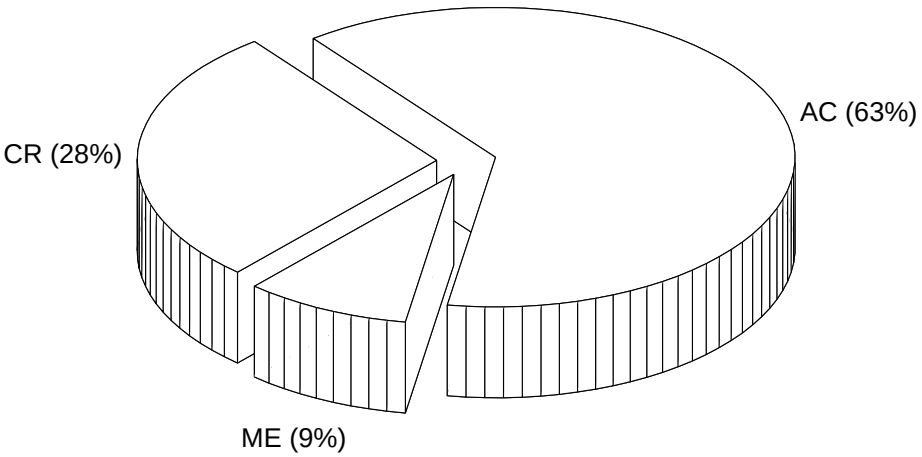


FIG. 11.1 – Répartition des erreurs (CR : cas où les deux approches concordent à raison - ME : cas où les deux approches font la même erreur - AC : autres cas restants)

la dernière catégorie représentent plus de 63% des occurrences de test. C'est sur ce type d'erreurs que nous allons tenter d'identifier le maximum d'occurrences.

11.1.2 Algorithme de fusion

Nous avons deux possibilités pour choisir la solution commune λ_c entre les deux approches (cf. figure 11.2) :

- Critère de rangs : on sélectionne la première solution commune entre les deux approches.
$$\lambda_c = \arg\min_i [rang_1(\lambda_i) + rang_2(\lambda_i)]$$
- Critère de scores : pour chaque test, les scores sont d'abord normalisés et varient entre $[0, 1]$. Pour chaque score S et pour chaque occurrence de test X, on effectue la transformation suivante :
$$S_{norm}(X, \lambda_i) = \frac{S(X, \lambda_i) - \min_i S(X, \lambda_i)}{\max_i S(X, \lambda_i) - \min_i S(X, \lambda_i)} \quad (11.1)$$
Ensuite, on sélectionne la solution commune λ_c telle que la somme des scores d'un locuteur dans les deux approches est maximale (S^1_{norm} et S^2_{norm} sont respectivement les scores normalisés de l'approche 1 et 2), soit :
$$\lambda_c(X) = \arg\max_i \left[S^1_{norm}(X, \lambda_i) + S^2_{norm}(X, \lambda_i)\right] \quad (11.2)$$

Ceci étant, on peut favoriser une approche par rapport à une autre en introduisant un facteur de pondération α. Le score à maximiser devient :
$$\lambda_c(X) = \arg\max_i \left[(1-\alpha)S^1_{norm}(X, \lambda_i) + \alpha S^2_{norm}(X, \lambda_i)\right] \quad (11.3)$$

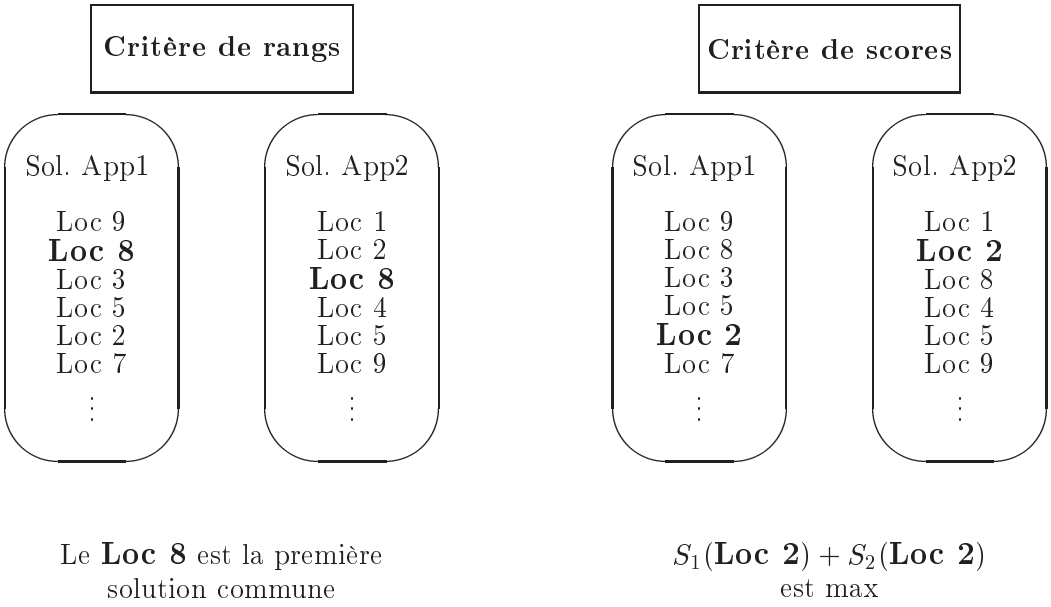


FIG. 11.2 – Algorithmes de fusion

Nous avons appliqué cet algorithme de fusion sur l'identification des 50 locuteurs de l'ensemble $\mathscr{E}_1$.

11.1.3 Evaluation de l'identification du locuteur par fusion

Le but de cette manipulation est d'évaluer les performances d'identification des 50 locuteurs de l'ensemble $\mathscr{E}_1$ et l'impact de différents paramètres tels que le facteur de pondération α (cf. équation 11.3) et la quantité de donnée d'apprentissage. La solution commune entre deux approches est sélectionnée selon le critère des scores.
Dans cette manipulation, nous allons fusionner les décisions des approches suivantes :

- La représentation GMM-UBM avec la représentation des locuteurs par Distribution sur les Modèles d'Ancrage (DMA).
- L'approche de ré-estimation des modèles par sélection de voisins (REM)[1] avec la représentation des locuteurs par Distribution sur les Modèles d'Ancrage (DMA).
- L'approche de ré-estimation des modèles par sélection de voisins (REM) avec l'approche GMM-UBM.

La figure 11.3 trace les variations des taux d'erreur en fonction des valeurs du facteur de pondération α pour les trois systèmes de fusion mentionnés ci-dessus. Le facteur α permet de favoriser une approche par rapport à une autre : lorsque $\alpha = 0$ on retrouve les taux

[1] La REM correspond à l'approche NMAM (*Neighborhood-Merged and Adapted Model*) étudiée dans le chapitre 10.

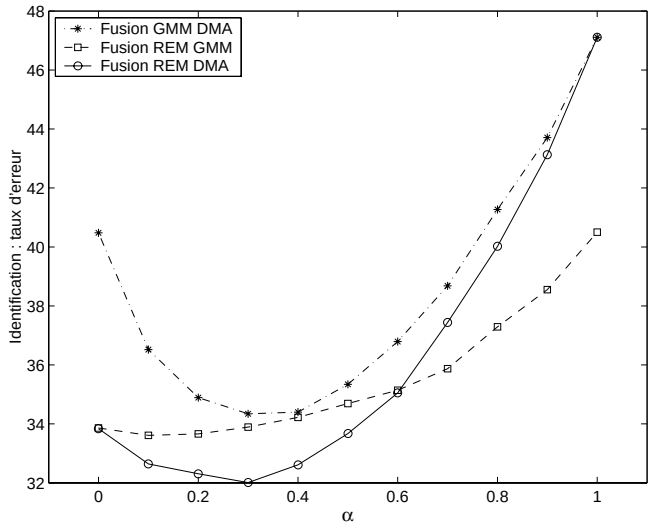


FIG. 11.3 – Fusion de décision : influence du facteur de pondération α (pour 4 secondes d'apprentissage)

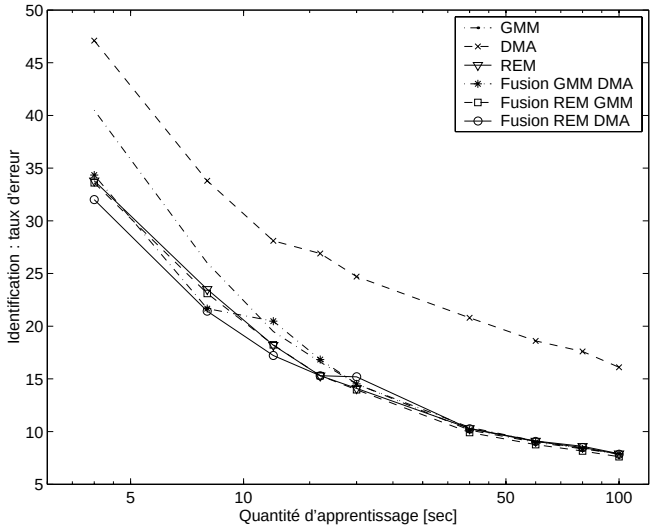


FIG. 11.4 – Influence de la quantité d'apprentissage (DMA : distribution par les modèles d'ancrage, REM : ré-estimation des modèles)

d'erreur de la première approche et lorsque $\alpha = 1$ les taux de la deuxième approche.
Cette expérience est réalisée dans le cas où les locuteurs sont modélisés par 256 gaussiennes et la quantité d'apprentissage est de 4 secondes de parole.
La figure 11.3 montre qu'il existe une valeur optimale de α. En effet, dans le cas où nous fusionnons les décisions de l'approche de ré-estimation des modèles par sélection de voisins (REM) avec la représentation des locuteurs par Distribution sur les Modèles d'Ancrage (DMA), le taux d'erreur atteint la valeur de 32.0% pour $\alpha = 0.3$ au lieu de 33.8% pour la première ($\alpha = 0$) et 47.1% pour la deuxième approche ($\alpha = 1.0$).

La figure 11.4 trace les variations des taux d'erreur en fonction de la quantité de données d'apprentissage pour les approches GMM-UBM, REM, DMA et la fusion de décision entre elles. Sur cette figure, nous constatons que la fusion de décision apporte une amélioration par rapport aux approches GMM-UBM, REM et DMA notamment dans le cas où nous disposons de peu de données. La figure 11.4 montre aussi que la fusion de décision de l'approche de ré-estimation des modèles par sélection de voisins (REM) avec la représentation des locuteurs par Distribution sur les Modèles d'Ancrage (DMA) donne les meilleures performances. Dans le cas où nous disposons de beaucoup de données (> à 20 secondes de parole), les systèmes donnent des performances quasi similaires.

11.2 Synthèse des résultats

Dans cette section, nous allons présenter et comparer les performances de reconnaissance par GMM-UBM et par localisation. Pour cela, nous avons tracé sur la figure 11.5 les variations des taux d'erreur en fonction de la quantité d'apprentissage pour les systèmes de reconnaissance suivants :

- Représentation des locuteurs par GMM-UBM.
- Localisation des locuteurs par les modèles d'ancrage dans un espace construit par sélection (avec un post-traitement ALD). La métrique utilisée pour évaluer la proximité entre locuteurs est l'angle entre leurs vecteurs de coordonnées.
- Représentation des locuteurs par distribution sur les modèles d'ancrage (DMA).
- Ré-estimation des modèles de locuteurs par sélection de voisins (REM)[2].

Les modèles GMM des locuteurs sont tous à 256 gaussiennes.

En ce qui concerne l'approche des modèles d'ancrage et la représentation par distribution sur les modèles d'ancrage, elles s'appuient toutes les deux sur les modèles d'ancrage ce qui permet de réduire la complexité des modèles. Lorsque nous disposons suffisamment de données (> 20 secondes de parole), les performances des deux approches sont similaires. Dans le cas où nous disposons de peu de données d'apprentissage, la figure 11.5 montre

[2]La REM correspond à l'approche NMAM (*Neighborhood-Merged and Adapted Model*) étudiée dans le chapitre 10.

que l'approche par distribution est plus robuste que l'approche géométrique, grâce à l'information a priori qu'elle prend en compte.

La ré-estimation des modèles de locuteurs par sélection de voisins (REM) est une approche proche des GMM-UBM : les deux techniques ont la même complexité et le même processus de reconnaissance. La figure 11.5 montre que les deux approches sont plus performantes que les autres systèmes de reconnaissance. Lorsque nous disposons de peu de données, la ré-estimation des modèles de locuteurs par sélection de voisins est meilleure que les GMM-UBM (jusqu'à 20 secondes de parole pour l'apprentissage). Les voisins apportent des informations significatives pour la modélisation des locuteurs. Les nouveaux modèles ont hérité des connaissances qu'on pouvait pas les avoir avec peu de données d'apprentissage (dans l'approche GMM-UBM) et cela bien que nous n'ayons pas suffisamment étudié la bonne correspondance des gaussiennes. Les travaux futurs s'orienteront certainement, d'une part vers l'amélioration de la couverture acoustique, et d'autre part l'attribution des poids aux voisins selon leur rapprochement ou éloignement du locuteur.

En vérification du locuteur, la figure 11.6 montre que si nous disposons de très peu de données, la représentation des locuteurs par distribution sur les modèles d'ancrage donne les meilleures performances.
La normalisation utilisée dans l'approche de la ré-estimation des modèles est la même que les GMM-UBM. Pourtant dans la REM, les modèles ne sont pas adaptés à partir du modèle UBM mais à partir de la fusion des voisins. Une perspective intéressante pour prolonger ce travail consiste à utiliser une autre technique de normalisation, par exemple la *Z-norm*.
L'approche de la distribution sur les modèles d'ancrage ne présente pas ce type de problèmes. Les scores calculés sont normalisés par un modèle du monde qui correspond à la distribution indépendante du locuteur et qui n'est rien d'autre que distribution initiale dans le cas mono-gaussien.
L'approche par distribution sur les modèles d'ancrage est une modélisation très compacte qui offre de bonnes performances en vérification du locuteur, plus intéressante que la modélisation GMM classique dans le cas où nous disposons de peu de données d'apprentissage.
Quant à l'estimation du modèle du locuteur par adaptation d'un mélange de modèles de locuteurs voisins est significativement meilleure en identification que l'approche GMM-UBM.

L'approche relative permet d'améliorer les performances d'identification et de vérification de locuteur. Elle représente un bon compromis entre la quantité de données disponibles et la complexité des modèles.

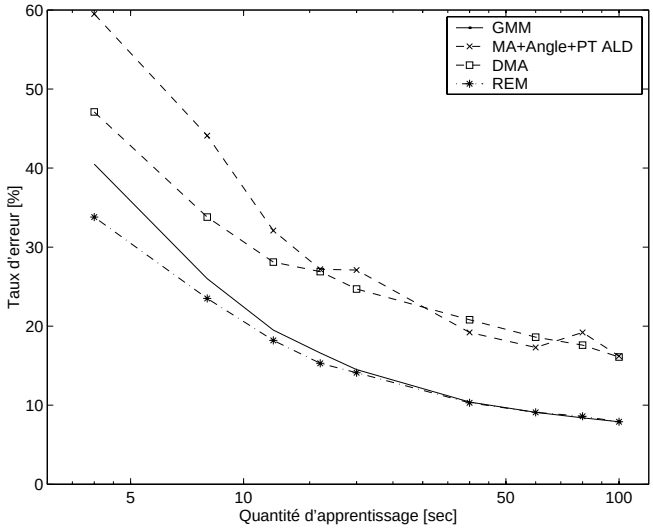


FIG. 11.5 – Performances des systèmes d'identification par GMM-UBM et par localisation

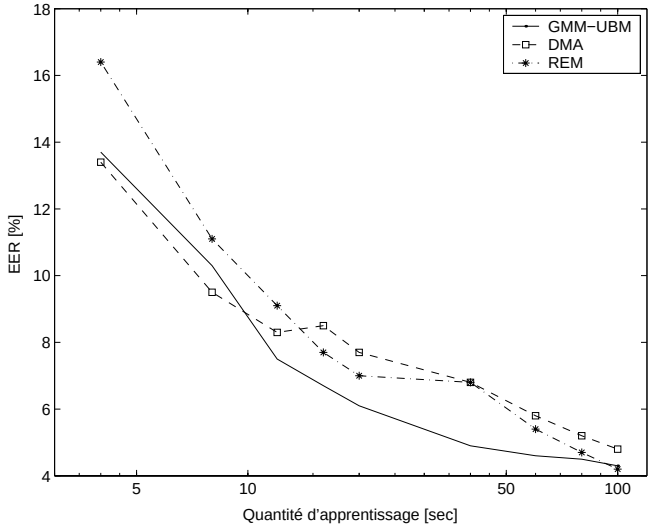


FIG. 11.6 – Performances des systèmes de vérification par GMM-UBM et par localisation

Conclusions et perspectives

Au cours de cette thèse, nous avons traité le problème de la représentation relative des locuteurs. Elle consiste à représenter un nouveau locuteur, non plus de façon absolue, mais relativement à un ensemble de locuteurs dont les modèles sont bien appris. Chaque locuteur est représenté par sa localisation dans un espace de référence. L'objectif était d'hériter des connaissances pour la modélisation que nous ne pouvions pas estimer avec peu de données.

Nous avons commencé par rappeler le principe de la reconnaissance automatique du locuteur et nous avons présenté les différentes étapes du système de reconnaissance. Cette introduction a permis de présenter le contexte général de la reconnaissance du locuteur et de comprendre la terminologie de l'identification et de la vérification du locuteur.

Dans la première partie de cette thèse, nous nous sommes intéressés à la modélisation par mélange de gaussiennes où les locuteurs sont modélisés par une somme pondérée de gaussiennes. Nous avons évalué notre système de reconnaissance de locuteur par GMM sur une base de données de France Télécom et sur une base publique (NIST 2000). Ces résultats d'évaluation constituent les résultats de référence avec lesquels nous allons comparer les performances de la reconnaissance par localisation.

La deuxième partie de ce document est consacrée à la modélisation relative des locuteurs où nous avons plusieurs contributions originales. Nous avons tout d'abord dressé un état de l'art sur les techniques de modélisation relative. Cette démarche a permis d'introduire la notion d'espace représentatif de locuteurs où un modèle de locuteur est estimé relativement à des modèles de référence.
Au cours de cette partie, nous avons appliqué le principe de représentation relative à la reconnaissance automatique du locuteur. Deux étapes du système ont été particulièrement étudiées : la construction de l'espace de locuteurs de référence et la localisation.
L'espace des locuteurs de référence peut être construit par différentes méthodes. La construction par les méthodes d'analyse de données permet d'une part, de bien illustrer la notion d'espace de locuteurs, et d'autre part, garantit l'orthogonalité de l'espace construit. En revanche, elles se basent généralement sur une simple approche géométrique. Nous avons proposé deux approches originales pour la construction de l'espace de référence : le regroupement hiérarchique et la sélection. Ainsi dans cette thèse, nous nous sommes proposés

d'étudier un système de reconnaissance où l'espace représentatif est généré par regroupement hiérarchique ou par sélection et les locuteurs sont placés par les modèles d'ancrage. Les évaluations de ces méthodes ont montré que la sélection fournit un meilleur espace de représentation et qu'il existe un espace "optimal" dans les deux approches (il est plus avantageux de sélectionner un sous-ensemble de locuteurs les plus dispersés que de conserver tous les locuteurs initiaux). Pour évaluer la proximité spatiale entre les locuteurs, l'angle semble la métrique la plus discriminante entre eux. De plus, nous avons proposé d'appliquer une orthogonalisation ACP/ALD sur les coordonnées des locuteurs. Ce post-traitement améliore significativement les performances d'identification de locuteur.

Dans cette thèse, les questions de recherche d'espace de locuteurs et de localisation dans cet espace ont été traitées indépendamment l'une de l'autre ; la recherche de l'espace ayant été faite dans un souci de couverture maximale de l'espace. Une perspective intéressante pourrait être de traiter conjointement les questions de localisation et d'espace de représentation (ce qui avait été abordé par [Sturim et al., 2001] dans une approche de sélection des modèles d'ancrage minimisant les distances intra-locuteur et maximisant les distances inter-locuteur).

Ces expériences ont permis de se rendre compte que les modèles d'ancrage permettent de réduire le nombre de paramètres mais les performances demeurent insuffisantes. Ceci est dû essentiellement au fait que ces approches géométriques accordent une place symétrique à l'apprentissage et au test, alors qu'en pratique, il existe souvent une asymétrie entre les occurrences d'apprentissage et de test. Ainsi, nous avons présenté une nouvelle représentation des locuteurs basée sur une distribution de distances. Par cette démarche, nous avons conservé la représentation compacte des modèles d'ancrage et nous avons introduit une densité de probabilité. Ce qui nous a permis d'une part, d'introduire des informations a priori pour la modélisation et d'autre part, d'appliquer une mesure statistique entre l'occurrence de test et les modèles des locuteurs à reconnaître (au lieu d'une mesure géométrique). Les évaluations ont montré que les performances de vérification de locuteur dépassent celles des GMM lorsque nous disposons de peu de données d'apprentissage.
De nombreuses perspectives permettent de prolonger ce travail notamment l'utilisation de cette représentation compacte dans les tâches particulières requises par un système d'indexation, comme la segmentation en locuteurs ou le regroupement en locuteurs. Cette modélisation compacte pourra être utilisée seule ou comme première passe d'élagage avant d'utiliser une modélisation plus riche et plus coûteuse telle que les mixtures de gaussiennes.

Par ailleurs, nous avons montré que la position relative d'un locuteur peut être exploitée pour l'estimation de son modèle GMM. Nous avons étudié une méthode de modélisation qui consiste à sélectionner les plus proches voisins et à fusionner leurs modèles pour générer le nouveau modèle du locuteur. Des expériences ont montré que le modèle fusionné capture des informations significatives du locuteur. Ainsi, la ré-estimation des modèles par sélection de voisins donne des bons résultats notamment en identification du locuteur où

les performances dépassent celles des GMM.
Cependant, cette fusion n'a de sens que si l'association des composantes du mélange dans l'espace acoustique est correcte. Les études futures de cette approche doivent porter sur la correspondance entre gaussiennes en introduisant, par exemple, des GMM de phonèmes. Les futurs travaux doivent étudier aussi le modèle de normalisation en vérification du locuteur et l'attribution des poids aux locuteurs lors de la fusion.

Ainsi, ces travaux de thèse ont permis de se rendre compte que l'approche relative offre des perspectives intéressantes de modélisation du locuteur avec peu de paramètres et peu de données d'apprentissage. La position relative d'un locuteur peut être utilisée directement dans la modélisation des locuteurs ou bien elle peut aider à l'estimation des paramètres de la modélisation classique GMM-UBM. Cette approche représente un bon compromis entre la quantité de données disponibles et la complexité des modèles. Elle mérite des études supplémentaires, tant dans son application aux tâches d'indexation en locuteur que dans son utilisation comme aide à la modélisation classique. Plusieurs points peuvent faire l'objet d'améliorations notables notamment dans l'approche de la ré-estimation des modèles par sélection de voisins et la distribution sur les modèles d'ancrage.

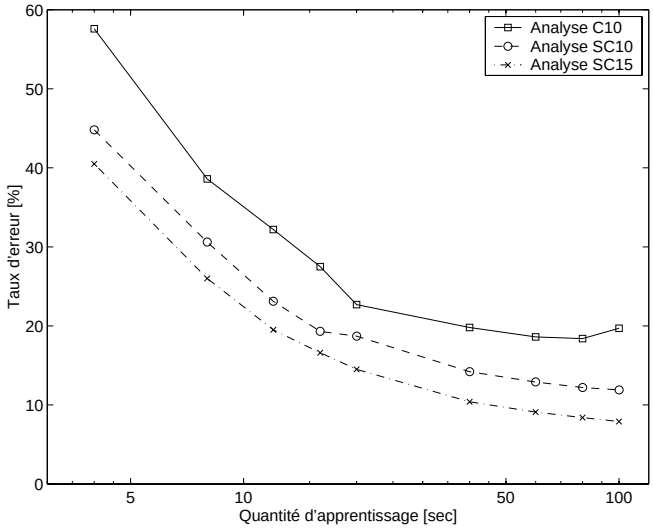
Annexe A

Compléments d'évaluations de la reconnaissance par GMM : influence de l'analyse acoustique

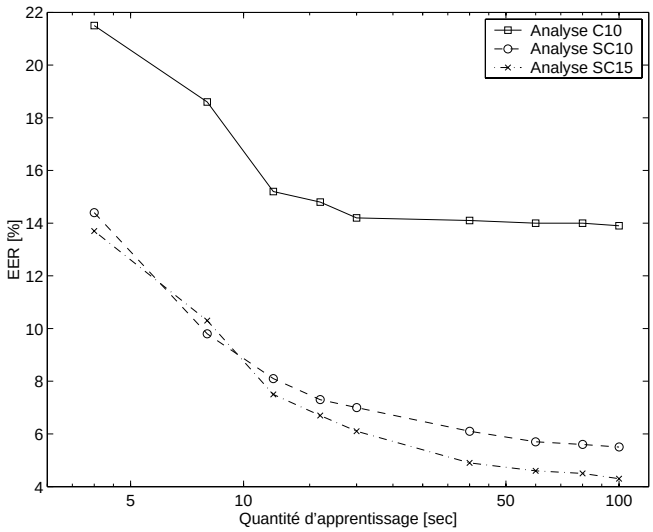
Dans cette partie, nous présenterons les performances de reconnaissance des 50 locuteurs de l'ensemble $\mathscr{E}_1$ en fonction de l'analyse acoustique. Nous appliquerons les trois analyses suivantes :

- Chaque vecteur acoustique est composé de l'énergie temporelle de la trame et des 8 premiers MFCC. A cela on rajoute leurs dérivées premières et secondes ; ce qui donne 27 coefficients acoustiques (notée analyse C10).
- Chaque vecteur acoustique est composé de l'énergie temporelle de la trame et des 8 premiers MFCC. A cela on rajoute leurs dérivées premières et secondes ; ce qui donne 27 coefficients acoustiques. On applique ensuite un module de soustraction cepstrale et un module de détection d'activité vocale (notée analyse SC10).
- Chaque vecteur acoustique est composé de l'énergie temporelle de la trame et des 13 premiers MFCC. A cela on rajoute leurs dérivées premières et secondes ; ce qui donne 42 coefficients acoustiques. On applique ensuite un module de soustraction cepstrale et un module de détection d'activité vocale (notée analyse SC15).

Les figures suivantes tracent les variations des taux d'erreur d'identification et de vérification en fonction de la quantité de données d'apprentissage. Ces figures montrent que l'introduction d'un module de détection d'activité vocale améliore significativement les performances d'identification et de vérification du locuteur. L'augmentation du nombre de vecteurs acoustiques (de 27 à 42 coefficients acoustiques) apporte aussi une nette amélioration (du SC10 au SC15).



(a) Performances de vérification du locuteur



(b) Performances d'identification du locuteur

FIG. A.1 – Influence de l'analyse acoustique sur les performances de reconnaissance du locuteur

Annexe B

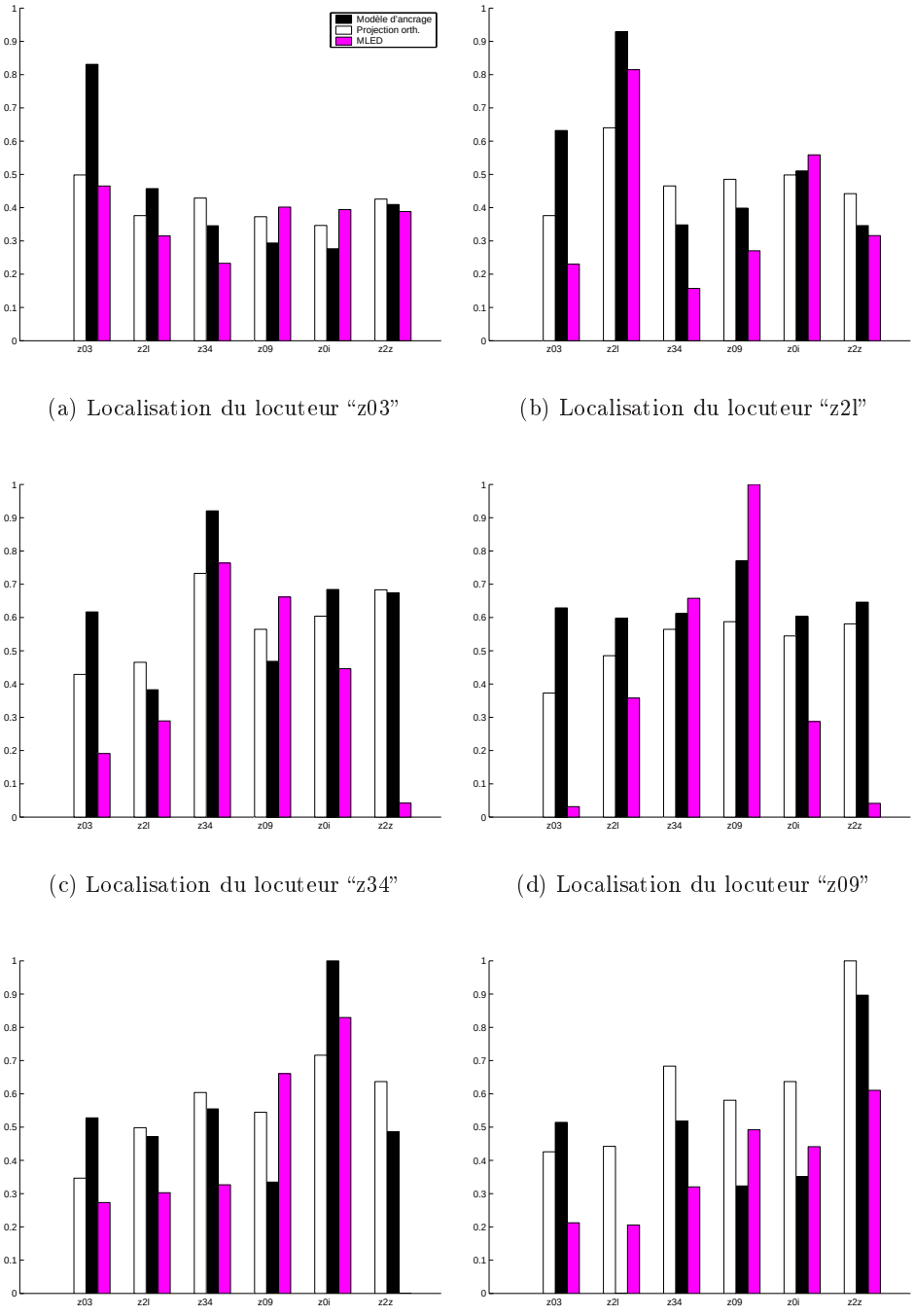
Qualité de localisation

Afin d'illustrer la qualité de localisation des locuteurs, nous allons localiser six locuteurs par rapport à eux même. Nous nous attendons à ce que la projection d'un locuteur par rapport à lui-même tende vers 1 tandis que sa projection par rapport aux autres tende vers 0. Nous avons effectué les manipulations suivantes :

- localisation des six locuteurs par projection orthogonale dans un espace constitué de leurs moyennes GMM[1] ;
- localisation des six locuteurs par MLED dans un espace constitué de leurs moyennes GMM ;
- localisation des six locuteurs par les modèles d'ancrage et par rapport à eux même.

Nous observons sur la figure B.1, qu'en général, la localisation par les modèles d'ancrage est meilleure que la projection orthogonale ou la MLED. Cependant, il est difficile de mesurer ou de quantifier la qualité de localisation.

[1] La localisation est réalisée sur un espace qui n'est pas orthogonal ; il s'agit d'une simple approximation.



(a) Localisation du locuteur “z03”

(b) Localisation du locuteur “z2l”

(c) Localisation du locuteur “z34”

(d) Localisation du locuteur “z09”

(e) Localisation du locuteur “z0i”

(f) Localisation du locuteur “z2z”

FIG. B.1 – Localisation des locuteurs “z03”, “z2l”, “z34”, “z09”, “z0i” et “z2z” par les modèles d’ancrage, projection orthogonale et MLED

Annexe C

Arbres de classifications de locuteurs

La construction d'un espace représentatif de locuteurs par regroupement hiérarchique ascendant présente l'avantage de pouvoir associer à chaque voix virtuelle des trames de parole. L'intérêt est double :

- d'abord, nous pouvons travailler directement sur les trames de parole en appliquant des métriques sur les coefficients acoustiques ou bien en les projetant dans un autre espace acoustique.
- Ensuite, nous avons la possibilité de faire un clustering sur un ensemble de locuteurs avec différentes valeurs du nombre de gaussiennes. Ainsi, nous pouvons construire des classes de locuteurs avec plusieurs niveaux de précision[1].

A titre d'exemple, les figures C.1 et C.2 tracent les dendrogrammes (ou arbres de classification) des 500 locuteurs de l'ensemble $\mathscr{E}_3$ à 16 et à 256 gaussiennes. Ces arbres montrent que pour deux niveaux de précision, nous n'avons pas les même classes de locuteurs.

[1] Par ailleurs, nous avons aussi la possibilité d'obtenir des classes de locuteurs (et leurs trames de parole) et de ré-estimer leurs modèles avec plus de gaussiennes.

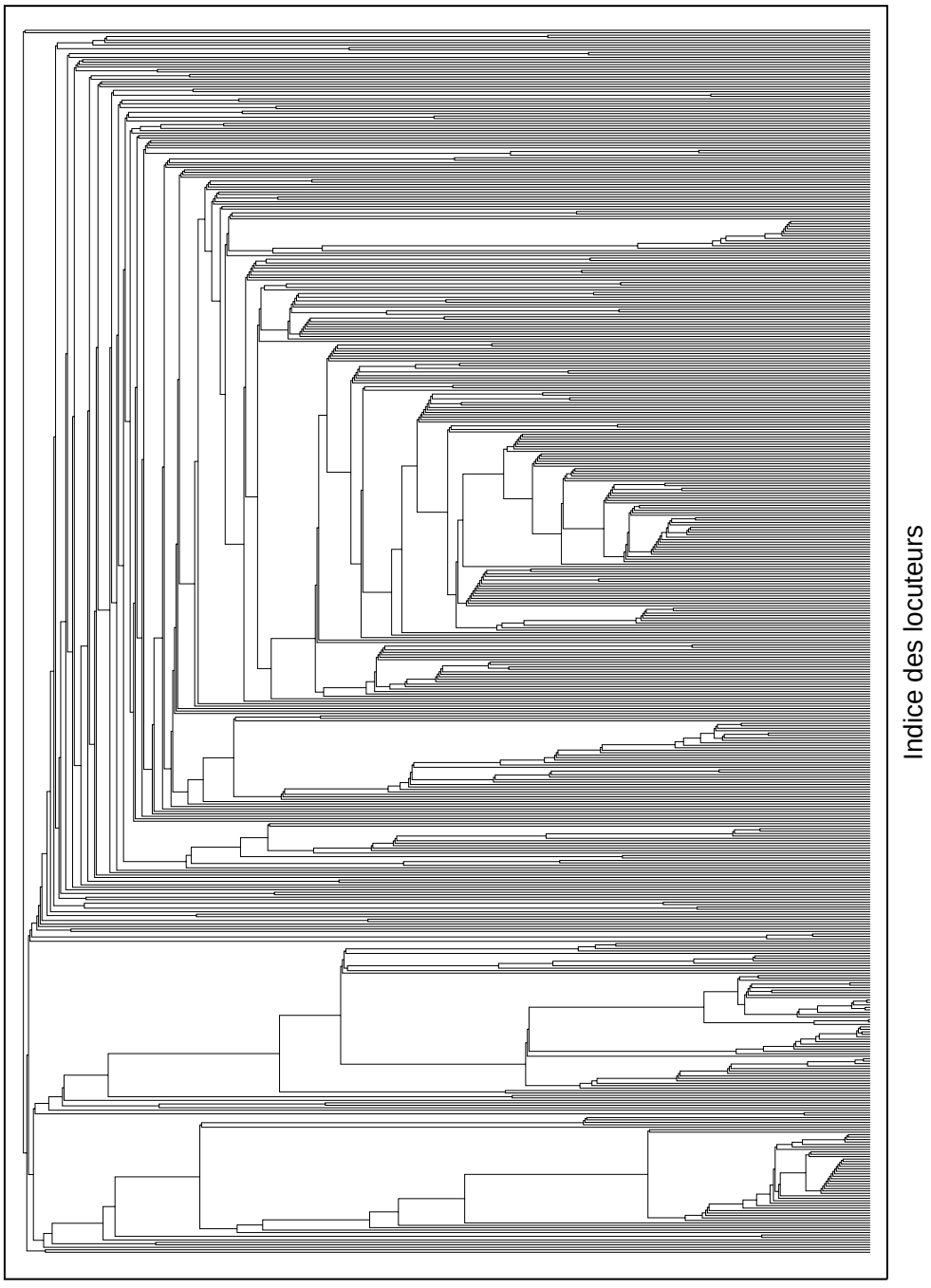


FIG. C.1 – Arbre des 500 locuteurs de l'ensemble $\mathscr{E}_3$ obtenu par regroupement hiérarchique avec 16 gaussiennes

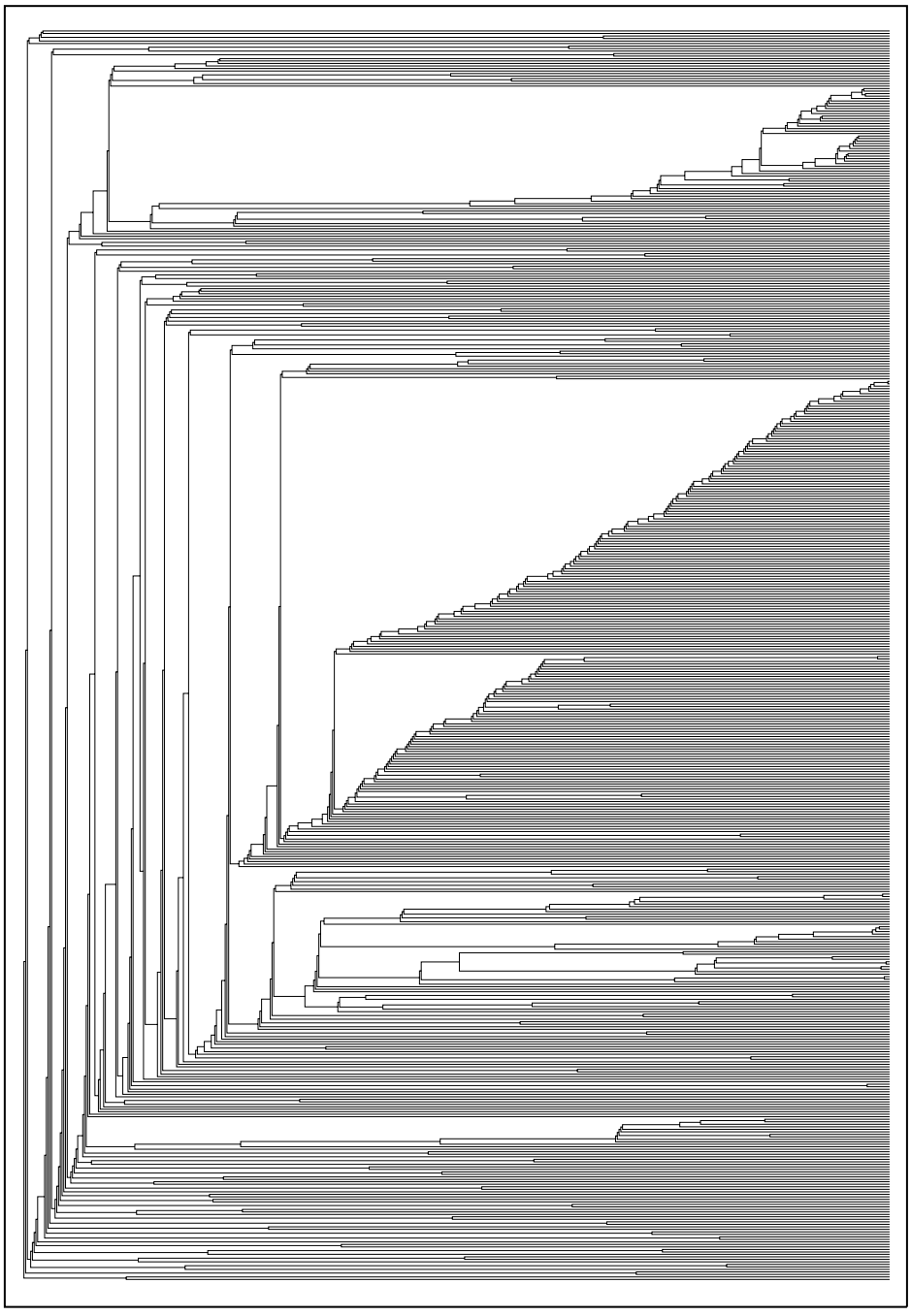


FIG. C.2 – Arbre des 500 locuteurs de l'ensemble $\mathscr{E}_3$ obtenu par regroupement hiérarchique avec 256 gaussiennes

Bibliographie

[Ahadi-Sarkani, 1996] Ahadi-Sarkani, S. M. *Bayesian and predictive techniques for speaker adaptation.* Thèse de Doctorat, University of Cambridge (1996).

[Artières et Gallinari, 1994] Artières, T. et Gallinari, P. Approches prédictives neuronales pour l'identification. Dans *XXèmes Journées d'Études sur la Parole (JEP)*, pages 275–280, Trégastel, France (1994).

[Auckenthaler et al., 2000] Auckenthaler, R., Carey, M., et Lloyd-Thomas, H. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10 :42–54 (2000).

[Bartkova, 2002] Bartkova, K. (2002). Production, description et perception du signal vocal. Rapport technique, France Telecom R&D Lannion.

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

[Boite et al., 1999] Boite, R., Boulard, H., Dutoit, T., Hancq, J., et Leich, H. *Traitement de la parole.* Collection Electricité - Presse Polytechniques et Universitaires Romandes (1999).

[Booth et al., 1993] Booth, I., Barlow, M., et Watson, B. Enhancements to DTW and VQ decision algorithms for speaker recognition. *Speech Communication*, 13(3-4) :427–433 (1993).

[Carrey et al., 1991] Carrey, M., Parris, E., et Bridle, J. A speaker verification system using alpha-nets. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 397–400 (1991).

[Charlet, 1997] Charlet, D. *Authentification vocale par téléphone en mode dépendant du texte.* Thèse de Doctorat, Telecom Paris (1997).

[Charlet, 2002] Charlet, D. Speaker indexing for retrieval of voicemail messages. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–124 (2002).

[Chen et Gopalakrishnan, 1998] Chen, S. S. et Gopalakrishnan, P. Clustering via the Bayesian information criterion with applications in speech recognition. Dans *International*

Conference on Acoustics, Speech and Signal Processing (ICASSP), volume 2, pages 645–648 (1998).

[Demirekler et Haydar, 1999] Demirekler, M. et Haydar, A. Feature selection using genetics-based algorithm and its application to speaker identification. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 329–332 (1999).

[Dempster et al., 1977] Dempster, A., Laird, N., et Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39 :1–38 (1977).

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

[Fukunaga, 1990] Fukunaga, K. *Introduction to statistical pattern recognition.* Academic Press (1990).

[Furui, 1981] Furui, S. Cepstral analysis technique for automatic speaker verification. Dans *IEEE Transactions Acoustics, Speech, and Signal Processing*, volume 29, pages 254–272 (1981).

[Gales, 1998] Gales, M. Cluster adaptive training for speech recognition. Dans *International Conference on Spoken Language Processing (ICSLP)* (1998).

[Gauvain et Lee, 1994] Gauvain, J.-L. et Lee, C.-H. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2) :291–298 (1994).

[Gish et al., 1991] Gish, H., Siu, M.-H., et Rohlicek, R. Segregation of speakers for speech recognition and speaker identification. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 873–876 (1991).

[Hazen, 1998] Hazen, T. J. *The use of speaker correlation information for automatique speech recognition.* Thèse de Doctorat, Massachusetts Institute of Technology (1998).

[Homayounpour et Chollet, 1995] Homayounpour, M. et Chollet, G. Neural net approaches to speaker verification : Comparison with second order statistic measures. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 353–356 (1995).

[Jollife, 1986] Jollife, I. T. *Principal Component Analysis.* Springer-Verlag (1986).

[Jom et al., 2001] Jom, E., Kim, D. K., et Kim, N. S. EMAP-based speaker adaptation with robust correlation estimation. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 321–324 (2001).

[Kuhn et al., 2000] Kuhn, R., Junqua, J.-C., Nguyen, P., et Niedzielski, N. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8(6) :695–707 (2000).

[Kuhn et al., 1998] Kuhn, R., Nguyen, P., Junqua, J.-C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K., et Contolini, M. Eigenvoices for speaker adaptation. Dans *International Conference on Spoken Language Processing (ICSLP)* (1998).

[L. Rabiner, 1993] L. Rabiner, B. H. J. *Fundamentals of speech recognition.* Prentice Hall Signal Processing Series (1993).

[Lebart et al., 2000] Lebart, L., Morineau, A., et Piron, M. *Statistique exploratoire multidimensionnelle.* Dunod (2000).

[Magrin-Chagnolleau et al., 1995] Magrin-Chagnolleau, I., Bonastre, J.-F., et Bimbot, F. Effect of utterance duration and phonetic content on speaker identification using second-order statistical methods. Dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 1, pages 337–340 (1995).

[Mami, 2000] Mami, Y. Segmentation en locuteurs d'un flux sonore. Rapport de stage, Ecole Supérieure d'Electricité Supelec, France Telecom R&D (2000).

[Mami et Charlet, 2002a] Mami, Y. et Charlet, D. Identification des locuteurs par regroupement hiérarchique ascendant et modèles d'ancrage. Dans *XXIVèmes Journées d'Études sur la Parole (JEP)*, pages 225–228, Nancy (France) (2002a).

[Mami et Charlet, 2002b] Mami, Y. et Charlet, D. Speaker identification by location in an optimal space of anchor models. Dans *International Conference on Spoken Language Processing (ICSLP)*, volume 2, pages 1333–1336, Denver (USA) (2002b).

[Mami et Charlet, 2003a] Mami, Y. et Charlet, D. Speaker identification by anchor models with PCA/LDA post-processing. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 180–183, Hong Kong (Chine) (2003a).

[Mami et Charlet, 2003b] Mami, Y. et Charlet, D. Speaker modeling from selected neighbors applied to speaker recognition. Dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 2629–2632, Genève (Suisse) (2003b).

[Matsui et Furui, 1992] Matsui, T. et Furui, S. Comparison of text-independent speaker recognition methods using VQ-Distortion and Discrete/Continuous HMMs. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 157–160 (1992).

[Matsui et Furui, 1994] Matsui, T. et Furui, S. Comparison of text-independent speaker recognition methods using VQ-Distortion and Discrete/Continuous HMMs. *IEEE Transactions on Speech and Audio Processing*, 2(3) :456–459 (1994).

[Merlin et al., 1999] Merlin, T., Bonastre, J.-F., et Fredouille, C. Non directly acoustic process for costless speaker recognition and indexation. Dans *COST-254 International Workshop on Intelligent Communication Technologies and Applications, with emphasis on Mobile Communications* (1999).

[Mokbel et Collin, 1999] Mokbel, C. et Collin, O. Incremental enrollment of speech recognizers. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 453–456 (1999).

[Nguyen et al., 1999] Nguyen, P., Wellekens, C., et Junqua, J.-C. Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments. Dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, volume 6, pages 2519–2522 (1999).

[Oglesby et Mason, 1989] Oglesby, J. et Mason, J. S. Speaker recognition with a neural classifier. Dans *First IEE International Conference*, pages 306–309 (1989).

[Padmanabhan et Nahamoo, 1998] Padmanabhan, M. et Nahamoo, D. Speaker clustering and transformation for speaker adaptation in speech recognition systems. *IEEE Transactions Speech Audio Processing*, 6(1) :71–77 (1998).

[Reynolds, 1995] Reynolds, D. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1) :91–108 (1995).

[Reynolds, 1997] Reynolds, D. Comparison of background normalization methods for text-independent speaker verification systems. Dans *European Conference on Speech Communication and Technology (EUROSPEECH)*, pages 963–966 (1997).

[Reynolds et al., 1998] Reynolds, D., Singer, E., Carlson, B., O'Leary, G., McLaughlin, J., et Zissman, M. Blind clustering of speech utterances based on speaker and language characteristics. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1998).

[Reynolds et al., 2000] Reynolds, D. A., Quatieri, T. F., et Dunn, R. B. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3) :19–41 (2000).

[Rissanen et Webb, 1993] Rissanen, E. L. et Webb, J. J. Speaker identification experiments using HMMs. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 387–390 (1993).

[Rosenberg et al., 1992] Rosenberg, A., Delong, J., Lee, C., Juang, B., et Soong, F. The use of cohort normalized scores for speaker recognition. Dans *ICSLP*, pages 599–602 (1992).

[Rosenberg et al., 1991] Rosenberg, E., Lee, C.-H., et Gokcen, S. Connected word talker verification using whole word hidden Markov models. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 381–384 (1991).

[Rozzi, 1991] Rozzi, W. A. M. *Speaker adaptation in continuous speech recognition via estimation of correlated mean vectors.* Thèse de Doctorat, Carnegie Mellon University, Pittsburg Pennsylvania (1991).

[Sambur, 1975] Sambur, M. Selection of acoustic features for speaker identification. *IEEE Transactions on Speech and Audio Processing*, 23(2) :176–182 (1975).

[Savic et Gupta, 1990] Savic, M. et Gupta, S. K. Variable parameter speaker verification system based on hidden Markov modeling. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 281–284 (1990).

[Seck et al., 2000] Seck, M., Blouet, R., et Bimbot, F. The IRISA/ELISA speaker detection and tracking systems for the NIST99 evaluation campaign. *Digital Signal Processing*, 13(10) :154–171 (2000).

[Solomonoff et al., 1998] Solomonoff, A., Mielke, A., Schmidt, M., et Gish, H. Clustering speakers by their voices. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 757–760 (1998).

[Sturim et al., 2001] Sturim, D., Reynolds, D., Singer, E., et Campbell, J. Speaker indexing in large audio databases using anchor models. Dans *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 429–432 (2001).

[Thyes et al., 2000] Thyes, O., Kuhn, R., Nguyen, P., et Junqua, J.-C. Speaker identification and verification using eigenvoices. Dans *International Conference on Spoken Language Processing (ICSLP)* (2000).

[Tipping et Bishop, 1997] Tipping, M. E. et Bishop, C. M. Mixtures of principal component analyzers. Dans *Fifth International Conference on Artificial Neural Networks*, pages 13–18 (1997).

[Yang et Honavar, 1998] Yang, J. et Honavar, V. Feature subset selection using a genetic algorithm. Dans *Intelligent Systems*, volume 13, pages 44–49 (1998).

[Yu et al., 1995] Yu, K., Mason, J., et Oglesby, J. Speaker recognition using hidden Markov models, dynamic time warping and vector quantization. *Vision, Image and Signal Processing*, 142(5) :313–316 (1995).

[Zwicker et Feldtkeller, 1981] Zwicker, E. et Feldtkeller, R. *Psychoacoustique*. CNET/ENST, Collection technique et scientifique des télécommunications, Masson Paris (1981).