REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole Nationale Polytechnique

Département d'Electronique

PROJET DE FIN D'ETUDES

THEME

IDENTIFICATION DU LOCUTEUR EN MODE INDEPENDANT DU TEXTE

Proposé et dirigé par :

B.BOUSSEKSOU

Etudié par :

TAKHEDMIT HAKIM

AIT SAADI NACER

Promotion Juin 2005

E.N.P 10 Avenues Hassen Badi EL HARRACH - ALGER

Dédicaces

A nos parents qui nous ont soutenu, orienté et encouragé le long de nos études

A nos frères et sœurs

A nos proches

A tous nos amis

Remerciements

Nous commençons par remercier notre promoteur Monsieur B.Bousseksou qui a accepté de nous proposer ce sujet, et de nous encadrer le long de cette thèse. Pour tous ses conseils et critiques sur le plan scientifique qui nous ont permet de bien orienter nos recherches.

Nous exprimons notre profonde reconnaissance à Monsieur M.Bouchamekh qui nous a beaucoup aidé et encouragé.

Nous tenons à remercier nos parents, frères et sœurs ainsi que tous nos proches qui nous ont encouragé, soutenu et aidé sur tous les plans, le long de nos études.

Nos remerciements vont également à tous les enseignants de l'Ecole Nationale Polytechnique qui ont contribué à notre formation.

Nous remercions tous ceux, qui de prés ou de loin, nous ont soutenu et aidé dans la réalisation de ce travail.

ملخص

تعتبر أنظمة تشخيص المتكلم (identification du locuteur) اختصاص من بين اختصاصات التعيين الصوتي (reconnaissance vocale), إذ تعرف هذه الأنظمة تطورا كبيرا في العشرية الأخيرة, وهذا يرجع إلي ازدياد الحاجة لاستعمال الإلكترونيك في مختلف التطبيقات الاحترافية.
هذا التطور فرض توسيع الإمكانيات الأمنية من أجل حماية حقوق العبور وكدا استعمال الخصائص الذاتية للأشخاص المتوفرة في أصواتهم.

بعد تعريف نظام تشخيص المتكلم وتقديم مختلف المراحل الإجبارية والمخططات من أجل دراسة الصوت : قولبة, تحليل وتصنيف سوف نتطرق إلي مختلف القوالب الإحصائية (GMM).

الكلمات المفتاحية : التشخيص, معاملات التنبؤ الخطي, التكميم الشعاعي, الغوصية المتعددة.

Résumé

L'identification du locuteur peut être considérée comme une tâche particulière de la reconnaissance de formes. Elle est en pleine expansion depuis dix ans et cela est dû à la généralisation de l'utilisation de l'électronique dans le domaine domestique et également professionnel.

Cette expansion nécessite la protection des droits d'accès, et la reconnaissance automatique du locuteur constitue le moyen le plus sûr et le plus ergonomique.

Dans cette thèse, on a étudié principalement la modélisation par mélange de gaussiennes (GMM) qui constitue l'état de l'art en la matière.

Les mots clés : identification automatique du locuteur, MFCC, LPC, QV, modélisation statistique, GMM, OGMM, HMM, DTW, ACP, FDR, ALD, orthogonalisation, espace acoustique.

Abstract

The speaker identification can be considered as a particular task of form recognition. It is into full expansion since ten years and that because of widening with the use of electronics in the domestic and professional fields. This widening required the protection of the rights of access, and the speaker recognition techniques are the surest means to resolve this problem.

In this thesis, we have study the Gaussian Mixture Models who constitute the state of art in this task.

Keywords: speaker identification, MFCC, LPC, VQ, statistical modelling, GMM, OGMM, HMM, DTW, PCA, FDR, LDA, orthogonalisation, acoustic space.

TABLE DES MATIERES

Chapitre 3

Chapitre 4

Liste des Figures

Liste des Tableaux

Acronymes

ACP: **A**nalyse en **C**omposantes **P**rincipales.

ALD : **A**nalyse **L**inéaire **D**iscriminante.

AR : **A**uto **R**égressif.

DTW: **D**ynamic **T**ime **W**arping (Alignement Temporel Dynamique).

EM : **E**xpectation **M**aximisation.

FDR : **F**isher **D**iscriminant **R**atio (Rapport Discriminant de Fisher).

FFT: **F**ast **F**ourier **T**ransform (Transformée de Fourier Rapide).

GMM: **G**aussian **M**ixture **M**odels.

HMM: **H**idden **M**arkov **M**odel (Modèles de Markov Cachés).

IAL : **I**dentification **A**utomatique du **L**ocuteur.

LBG: **L**inde **B**uzo **G**ray.

LFCC: **L**inear **F**requency **C**epstral **C**oefficients.

LPC: **L**inear **P**rediction **C**oefficients.

LPCC: **L**inear **P**rediction **C**epstral **C**oefficients.

MFCC: **M**el **F**requency **C**epstral **C**oefficients.

OGMM: **O**rthogonal **G**aussian **M**ixture **M**odels.

RAL : **R**econnaissance **A**utomatique du **L**ocuteur.

RTC : **R**éseau **T**éléphonique **C**ommuté.

SNR: **S**ignal to **N**oise **R**atio (Rapport Signal sur Bruit).

TFD: **T**ransformée de **F**ourier **D**iscrète.

VAL : **V**érification **A**utomatique du **L**ocuteur.

Introduction

La reconnaissance automatique du locuteur s'inscrit dans le domaine plus général du traitement de la parole, et est interprétée comme une tâche particulière de reconnaissance de formes. Ce domaine regroupe les problèmes relatifs à l'identification et à la vérification du locuteur sur la base de l'information contenue dans le signal vocal et relative à l'identité du locuteur. Le champ d'application de ces techniques est très vaste, allant des applications domestiques aux applications militaires, en passant par des applications judiciaires.

Un système de reconnaissance automatique du locuteur est constitué généralement de trois modules : un module pour l'extraction des coefficients acoustiques, un autre module pour la modélisation des locuteurs et enfin un module de classification et de décision.

Au cours de ce PFE, qui consiste à l'identification du locuteur en mode indépendant du texte, nous nous intéressons essentiellement à l'information extralinguistique contenue dans le signal vocal. Pour extraire du signal vocal l'information relative à l'identité du locuteur, on utilise les coefficients cepstraux qui permettent une bonne séparation de la contribution du conduit vocal et celle de la source d'excitation glottique.

Pour la modélisation des locuteurs, plusieurs approches existent : approche vectorielle, connexionniste, statistique et relative. De cette large gamme d'approches, l'approche statistique demeure au premier plan.
En effet, la modélisation par un mélange de gaussiennes (GMM : Gaussian Mixture Models) fournit de 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 gaussiennes.

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 une classe acoustique. L'autre raison poussant à utiliser la 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.

Pour tenter de remédier à ce problème, une intéressante méthode de modélisation consiste tout d'abord à orthogonaliser l'espace acoustique propre à chaque locuteur et ensuite d'appliquer la GMM sur les vecteurs acoustiques résultants. Cette nouvelle technique s'appelle OGMM (Orthogonal Gaussian Mixture Models).
Cette méthode permet de diminuer de manière considérable le nombre de gaussiennes à utiliser, ce qui réduit beaucoup le temps de calcul requis pour l'apprentissage et le test.

L'apprentissage des modèles GMM par maximum de vraisemblance en utilisant l'algorithme EM, qui garantit la convergence vers un maximum local, pose le problème de temps de calcul et de complexité. Avec les deux algorithmes de la quantification vectorielle : LBG et K-moyennes, on tente d'apporter une solution au problème posé par l'algorithme EM. Ces deux algorithmes permettent une réduction considérable de la complexité et du temps de calcul, tout en fournissant des performances comparables à celles obtenues par l'algorithme EM, ce qui les rend le choix idéal pour les applications grand public qui ne nécessitent pas un niveau de sécurité élevé.

Cet ouvrage s'articule autour de quatre chapitres. Le premier chapitre constitue une introduction aux systèmes de reconnaissance automatique du locuteur.

Le deuxième chapitre traite les problèmes relatifs à la production et à la perception de la parole, la modélisation autorégressive de la parole ainsi qu'à l'extraction des coefficients acoustiques.

Le troisième chapitre expose les différentes approches de modélisation des locuteurs, et détaille les deux approches GMM et OGMM.

Le quatrième et dernier chapitre décrit le contexte expérimental et expose les résultats des différents tests effectués. Pour cette dernière section, on a essayé d'examiner et de voir l'influence d'un certain nombre de paramètres (la qualité des données d'apprentissage et de test, le nombre de coefficients acoustiques, l'algorithme d'apprentissage, le nombre de locuteurs et la quantité des données de test) sur le taux d'identification correcte et sélectionner par la suite l'ensemble des paramètres qui donne les meilleures performances, pour une éventuelle conception d'un système d'identification du locuteur.

Enfin, un ensemble de conclusions et de perspectives conclue le travail de ce projet de fin d'études.

Chapitre 1

Introduction à la reconnaissance automatique du locuteur : RAL

La reconnaissance automatique du locuteur (*RAL*) regroupe l'identification et la vérification du locuteur sur la base de l'information contenue dans le signal vocal, en d'autres termes, il s'agit de reconnaître une personne à partir de sa voix.
Un système de reconnaissance du locuteur procède en trois étapes : l'analyse acoustique du signal de parole, la modélisation du locuteur et en dernier lieu vient l'étape de décision.

1.1 Introduction à la reconnaissance automatique du locuteur

Comme l'illustre 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 extralinguistiques du signal vocal.
Les variations individuelles entre locuteurs ont deux origines essentielles. D'abord, les caractéristiques morphologiques de l'appareil de phonation qui diffèrent d'un locuteur à un autre, et ensuite les différences dans les débits d'élocution, l'étendue des variations du pitch ou encore les différences liées au milieu socioculturel. Cette variabilité inter-locuteurs 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. Cependant, la voix ne peut être considérée comme une caractéristique biométrique d'une personne compte tenu de la variabilité intra-locuteur. Ainsi, on préfère la qualifier comme signature vocale plutôt qu'une empreinte vocale.

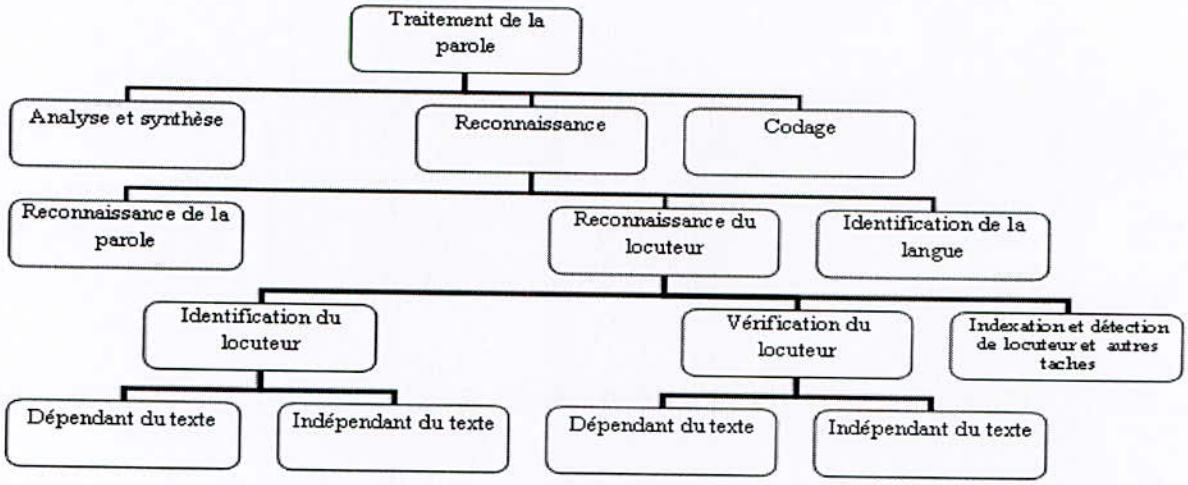


FIG. 1.1 Traitement de la parole

1.2 Systèmes de reconnaissance automatique du locuteur

Les systèmes de reconnaissance automatique du locuteur comportent plusieurs modules. Tout d'abord, un module d'acquisition qui capte le signal vocal et le convertie en un signal numérique. Ensuite vient le module d'analyse acoustique, à l'issue duquel des vecteurs de coefficients pertinents, servants pour la modélisation du locuteur, sont extraits.

Dans l'étape d'apprentissage, un modèle est créé pour chaque locuteur. Dans l'étape de reconnaissance, un module de classification va mesurer la similarité entre les données de test et un ou tous 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.

1.3 Les différentes tâches en RAL

A l'identification automatique du locuteur (**IAL**) et la vérification automatique du locuteur (**VAL**) vient s'ajouter, récemment et pour des applications spécifiques, d'autres tâches comme l'indexation du locuteur qui consiste à indiquer à quel moment chaque locuteur intervenant dans une conversation a pris la parole. Une application connexe est la détection d'un locuteur lors d'une conversation.

1.3.1 Identification Automatique du Locuteur (IAL)

L'identification automatique du locuteur consiste à reconnaître une personne parmi un ensemble de locuteurs en comparant ses paramètres de test aux différents modèles de locuteurs présents dans la base.

Dans le cas où le système doit fournir un ensemble d'au moins un locuteur, on parle d'une identification dans un ensemble fermé. Mais dans certaines applications, où le système peut être amené à fournir un ensemble vide, on parle d'une identification en ensemble ouvert.

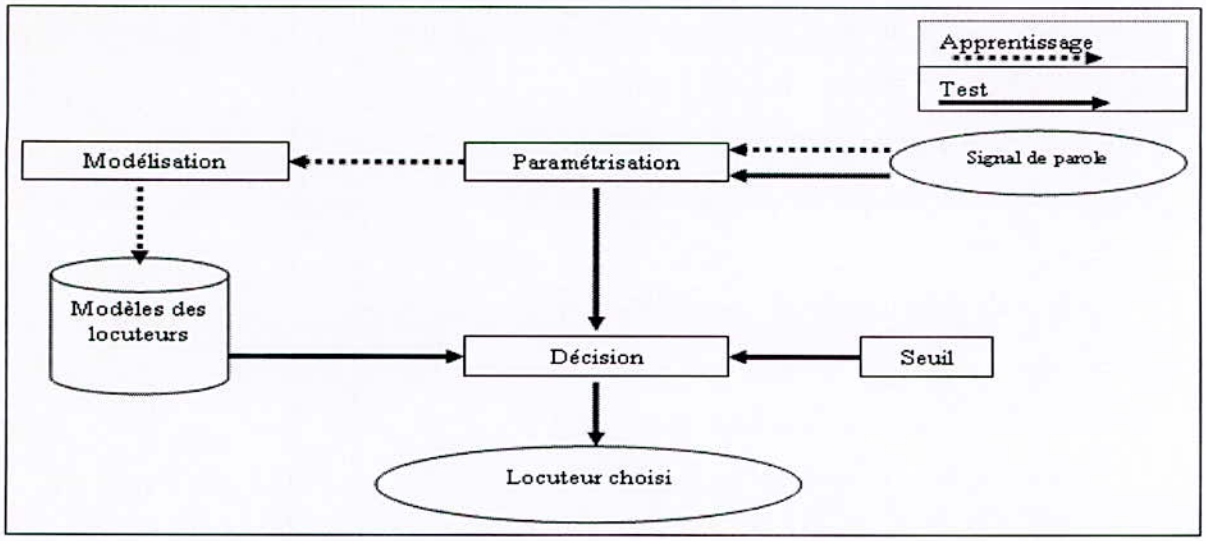


FIG. 1.2 Schéma modulaire d'un système d'IAL

1.3.2 Vérification Automatique du Locuteur (VAL)

La vérification du locuteur consiste, après que le locuteur ait décliné son identité, à vérifier l'adéquation de son message vocal avec la référence acoustique du locuteur qu'il prétend être.

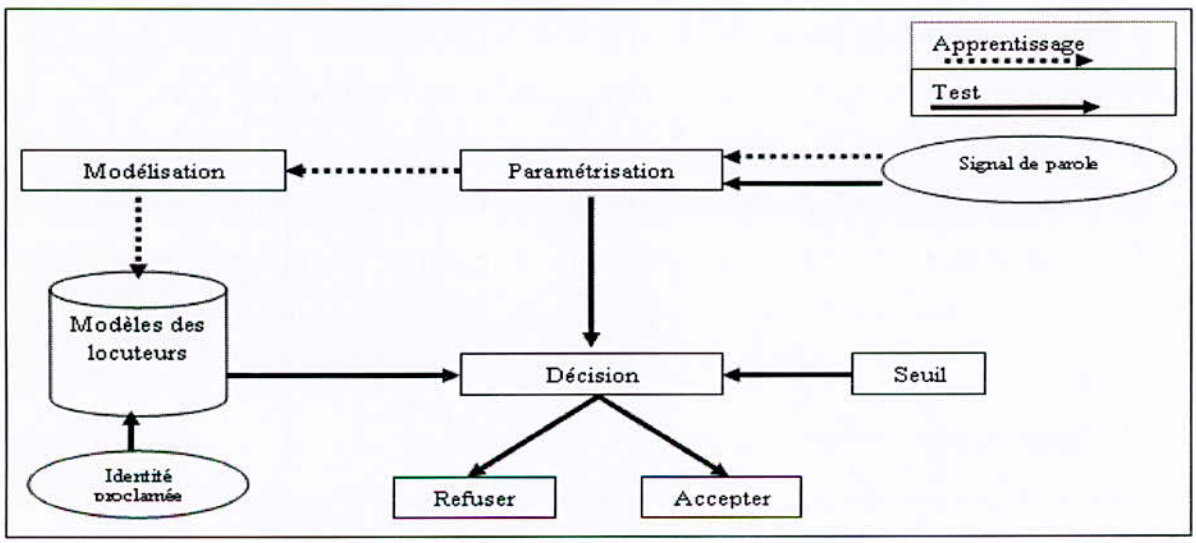


FIG. 1.3 Schéma modulaire d'un système de VAL

1.4 Modes dépendant et indépendant du texte

On peut classer les systèmes de reconnaissance automatique du locuteur en deux catégories, qui correspondent aux deux modes dépendant ou indépendant du texte.
Les niveaux de dépendance au texte sont classés suivant les applications :

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

1.5 Domaines d'applications

Les applications des systèmes de reconnaissance automatique du locuteur sont nombreuses et diversifiées. Néanmoins, elles peuvent être regroupées en trois catégories : application sur sites géographiques, applications juridiques et applications téléphoniques.

1.5.1 Applications sur sites géographiques

Cette catégorie concerne les applications qui se trouvent sur un site géographique particulier, elles sont utilisées principalement pour limiter l'accès à des lieux privés. On peut citer :

- Le verrouillage automatique pour la protection de domiciles, garages, bâtiments, etc.
- La sécurisation accrue des cartes d'accès et le contrôle d'accès à des zones protégées.
- Les validations des transactions sur site (au niveau des distributeurs bancaires).

1.5.2 Applications téléphoniques

C'est la catégorie la plus importante car elle permet de vérifier ou d'identifier un locuteur à longue distance. Parmi ces applications on cite :

- Validation des transactions bancaires par téléphone.
- Accès à des bases de données pour plus de sécurité et plus de protection.
- Accès à des services téléphoniques.
- Le commerce électronique.

1.5.3 Applications juridiques

Dans cette catégorie, la reconnaissance automatique du locuteur est utilisée par exemple pour :

- L'orientation des enquêtes.
- La constitution des éléments de preuves au cours d'un procès.

Dans cette catégorie d'applications, on trouve beaucoup d'inconvénients :

- La quantité de parole à disposition est généralement très limitée.
- Les conditions d'environnement sont très mauvaises.
- Les locuteurs impliqués sont rarement coopératifs.

1.6 Les sources d'erreurs

Il existe plusieurs facteurs qui peuvent augmenter la variabilité intra-locuteur et qui, par conséquent, influencent sur la décision du système de reconnaissance :

- L'état pathologique du locuteur (maladie, émotions, …).
- Vieillissement (la voix d'une personne change avec l'âge).
- Facteurs socioculturels (le locuteur peut changer d'accent).
- Locuteurs non coopératifs (notamment dans les applications judiciaires).
- Conditions de prise de son, bruit ambiant,…

Chapitre 2
Modélisation et paramétrisation de la parole

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage, cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications.

L'importance particulière du traitement de la parole dans ce cadre plus général s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine.

L'extraordinaire singularité de cette science, qui la différencie fondamentalement des autres composantes du traitement de l'information, tient sans aucun doute au rôle fascinant que joue le cerveau humain à la fois dans la production et dans la compréhension de la parole et à l'étendue des fonctions qu'il met, inconsciemment, en œuvre pour y parvenir de façon pratiquement instantanée.

2.1 Modélisation de la parole

L'étude des mécanismes de phonation permettra donc de déterminer, dans une certaine mesure, ce qui est parole et ce qui n'en est pas. De même, l'étude des mécanismes d'audition et des propriétés perceptuelles qui s'y rattachent permettra de déterminer ce qui, dans le signal de parole, est réellement perçu.

2.1.1 Production et perception de la parole

2.1.1.1 La parole

La parole est la faculté de communiquer la pensée par un système de sons articulé ; c'est le moyen de communication privilégié entre les humains qui sont les seuls êtres vivants à utiliser un tel système structuré.

L'information d'un message parlé réside dans les fluctuations de la pression de l'air, engendrées, puis émises, par l'appareil phonatoire. Ces fluctuations constituent le signal vocal ; elles sont détectées par l'oreille, laquelle procède à une certaine analyse dont les résultats sont interprétés par le cerveau.

L'information portée par le signal de parole peut être analysée de plusieurs façons. On en distingue généralement plusieurs niveaux de description : *acoustique, phonétique, phonologique, morphologique, syntaxique, sémantique, et pragmatique.*

Dans le cadre de cette thèse, on s'intéresse principalement au niveau phonétique.

2.1.1.2 Le niveau phonétique

Dans l'analyse phonétique, on s'intéresse à la façon dont le signal de parole est produit par le système articulatoire et perçu par le système auditif.

2.1.1.3 Appareil phonatoire et mécanisme de la phonation

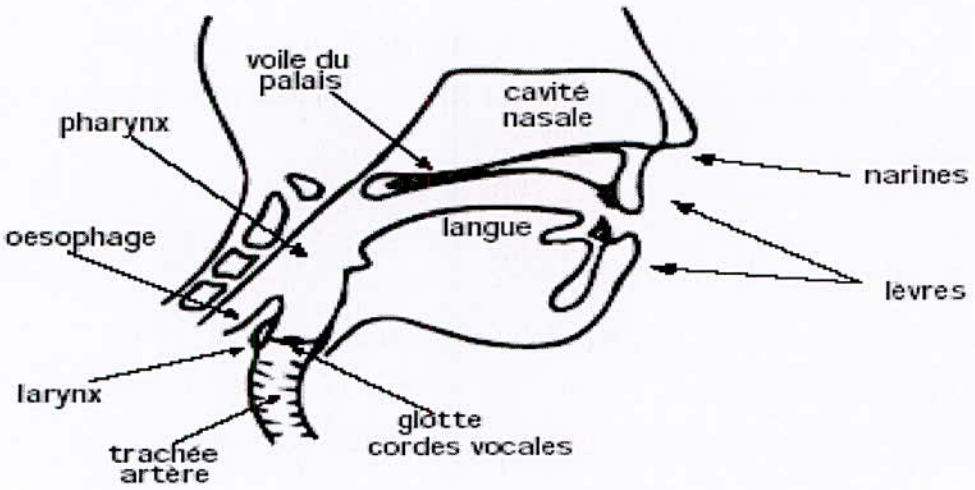


FIG. 2.1 L'appareil phonatoire

La contribution coordonnée et volontaire des deux appareils respiratoire et masticatoire, sous le contrôle du cerveau, permet la production de la parole.

Le processus de production de la parole passe par les étapes suivantes :

- L'appareil respiratoire fournit l'énergie nécessaire.
- La trachée artère expire l'air.
- Le larynx, et grâce aux cordes vocales, module la pression de l'air et l'applique au conduit vocal.
- La langue a un rôle prépondérant dans le processus phonatoire. Sa hauteur détermine : la hauteur du pharynx, le lieu d'articulation qui est la région de rétrécissement maximal du canal buccal, ainsi que l'aperture qui est l'écartement des organes au point d'articulation.

Le conduit vocal, qui s'étend du pharynx jusqu'aux lèvres, est l'ensemble des cavités acoustiques suivantes : la cavité pharyngienne, la cavité buccale et la cavité nasale en dérivation.

L'ouverture complète des cordes vocales engendre des sons non voisés, par contre leurs vibrations périodiques engendrent des sons voisés.

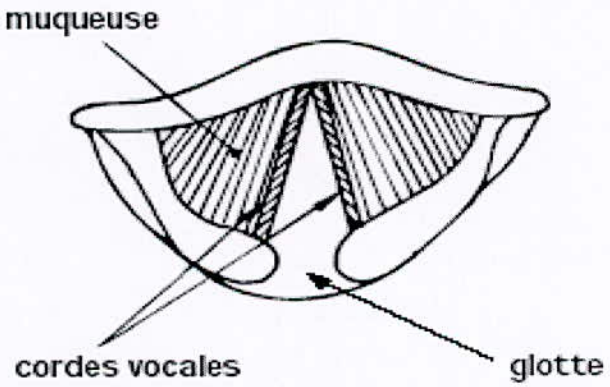


FIG. 2.2 Section du larynx vue du haut

Avant d'entamer la suite de ce chapitre, nous avons jugé indispensable la définition de quelques concepts utilisés dans le décodage acoustico-phonétique.

- **Les sons voisés**

Les sons voisés ont une structure quasi-périodique, ils résultent de l'excitation du conduit vocal par un train périodique d'impulsions de pression liées aux oscillations des cordes

vocales ; l'ouverture brusque de la glotte libère la pression accumulée en amont, elle se referme ensuite plus graduellement.

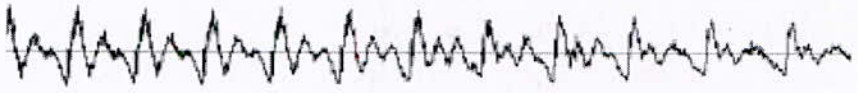


FIG. 2.3 Son voisé

➢ **Les sons non voisés**

Un son non voisé ne présente pas de structure périodique, il peut être approximé par la réponse du conduit vocal à un bruit blanc gaussien.

FIG. 2.4 Son non voisé

➢ **La fréquence du fondamentale**

La fréquence du fondamentale, appelée aussi pitch, est la fréquence de vibration des cordes vocales. Elle peut varier :

- de 80 à 200 Hz pour une voix masculine.
- de 150 à 450 Hz pour une voix féminine.
- de 200 à 600 Hz pour une voix d'enfant.

➢ **Le timbre**

Le timbre est déterminé par les amplitudes relatives des harmoniques du pitch.

➢ **Les formants**

Les formants sont les maximums de l'enveloppe des raies, correspondants aux harmoniques du fondamentale. Ils correspondent aux fréquences propres du conduit vocal.

➢ **La mélodie**

La mélodie de la voix est liée aux fluctuations du pitch au cours du temps.

➢ **La prosodie**

La prosodie introduit des nuances dans la prononciation d'une phrase. Les caractéristiques prosodiques permettent à un auditeur de suivre une conversation même en milieu bruité. Les principaux paramètres prosodiques sont :

- **L'intonation**

  L'intonation correspond à la hauteur ou à l'amplitude du pitch.

- **L'intensité**

  L'intensité donne des informations sur l'amplitude de la voix. Celle-ci peut être normale, chuchotée ou criée.

- **La durée**

  La durée fixe le rythme de la phrase.

➢ **L'articulation**

Les phénomènes d'articulation sont à l'origine des traits distinctifs entre locuteurs, et concernent l'activité musculaire du locuteur.

- **La co-articulation**

  L'absence de pauses dans la parole a pour conséquence le phénomène de la co-articulation, qui consiste en une prononciation fonction des unités adjacentes.

- **Occlusives**

  C'est la durée du silence précédent l'explosion dans les plosives.

- **Enveloppe énergétique**

  L'enveloppe énergétique qui est par définition la distribution de l'énergie du signal dans le domaine fréquentiel, est liée à l'identité du locuteur.

2.1.1.4 Appareil auditif et mécanisme d'audition

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. En effet, tout ce qui peut être mesuré acoustiquement ou observé par la phonétique articulatoire n'est pas nécessairement perçu. Par ailleurs, nous avons déjà souligné le rôle fondamental que joue l'audition dans le processus même de production de la parole.

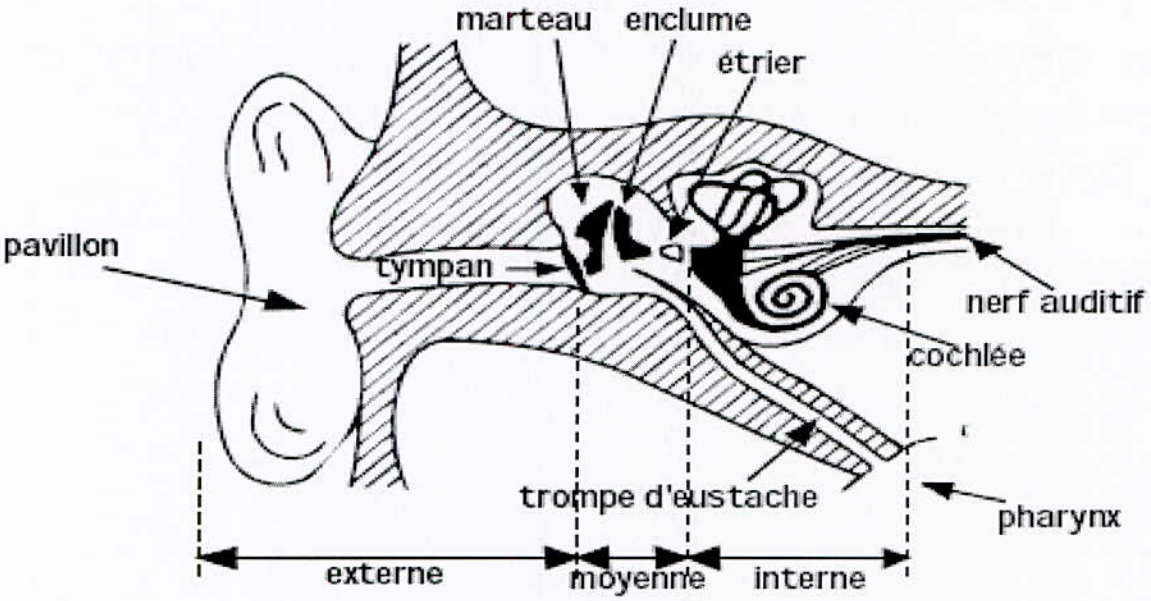


FIG. 2.5 Le système auditif

Comme l'illustre la figure 2.5, l'appareil auditif comprend l'oreille externe, l'oreille moyenne et l'oreille interne.

Les ondes sonores sont recueillies par l'appareil auditif. Ces ondes de pression sont ensuite analysées dans l'oreille interne, puis les résultats sont interprétés par le cerveau.

Le conduit auditif relie le pavillon au tympan : c'est un tube acoustique de section uniforme fermé à une extrémité, son premier mode de résonance est situé vers 3 KHz, ce qui accroît la sensibilité du système auditif humain dans cette gamme de fréquences.

Le mécanisme de l'oreille interne (*marteau, étrier, enclume*) permet une adaptation d'impédance entre l'air et le milieu liquide de l'oreille interne. Les vibrations de l'étrier sont transmises au liquide de la *cochlée*. Celle-ci contient la *membrane basilaire* qui transforme

les vibrations mécaniques en impulsions nerveuses. La membrane s'élargit et s'épaissit au fur et à mesure que l'on se rapproche de l'apex de la cochlée; elle est le support de l'*organe de Corti* qui est constitué par environ 25000 *cellules ciliées* raccordées au nerf auditif. Chacune de ces cellules présente une réponse en fréquence telle que celle illustrée à la figure 2.6.

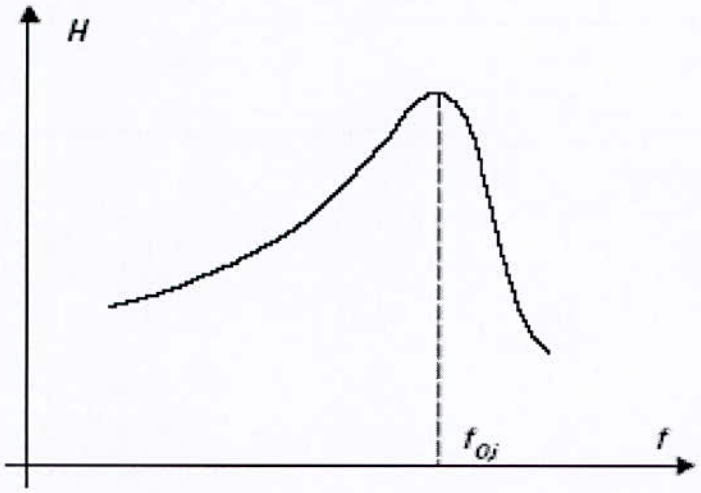


FIG. 2.6 Réponse en fréquence d'une cellule ciliée

La fréquence de résonance dépend de la position occupée par la cellule sur la membrane; au-delà de cette fréquence, la fonction de réponse s'atténue très vite.
Le système auditif humain est surtout sensible dans une gamme de fréquences situées entre 800 Hz à 8000 Hz : les limites extrêmes sont respectivement 20 Hz et 20 KHz.

2.1.1.5 Caractéristiques phonétiques

2.1.1.5.1 Le phonème

Un phonème est la plus petite unité présente dans la parole et susceptible par sa présence de changer la signification d'un mot.
Le nombre de phonèmes est toujours très limité, il est de 36 pour la langue française et de 42 pour la langue anglaise.

2.1.1.5.2 Classification des phonèmes

La figure 2.7 donne la répartition des phonèmes en classes et en sous-classes ; chaque classe correspond à un mode articulatoire donné de l'appareil de phonation.

On distingue trois classes principales : les voyelles, les semi-consonnes et les consonnes.

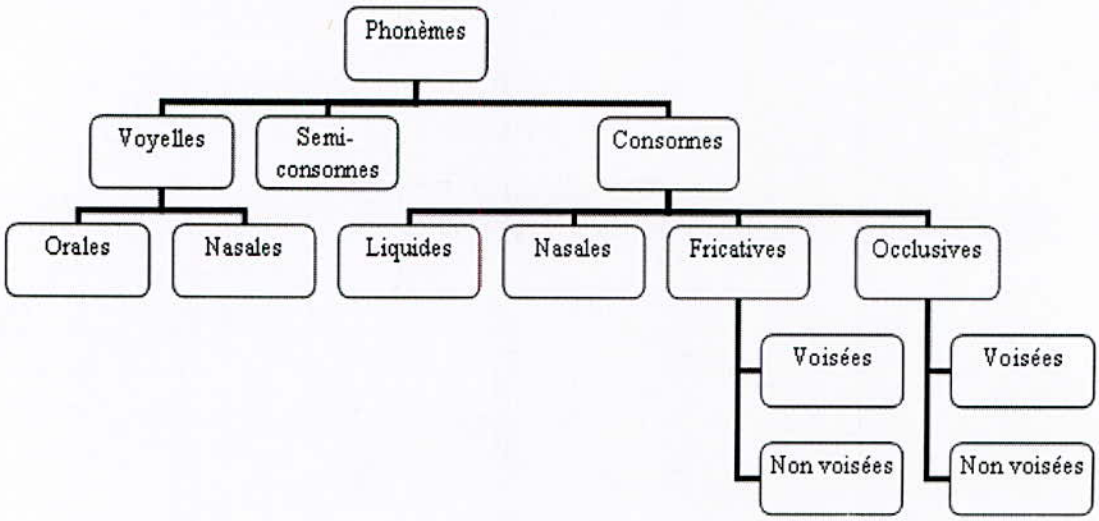


FIG. 2.7 Classification des phonèmes

Les voyelles diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal. Si le conduit vocal est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, alors une voyelle est produite. Le rôle de la bouche se réduit alors à une modification du timbre vocalique.

Si , par contre, le passage se rétrécit par endroit, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit, et une consonne est produite. La bouche est, dans ce cas, un organe de production à part entière.

Les voyelles sont classées en voyelles orales et nasales.

- Les voyelles orales sont des sons voisés ; chacune d'elles correspond à une configuration particulière du conduit vocal, sans intervention de la cavité nasale.
- Les voyelles nasales font intervenir le conduit nasal et la cavité buccale, et l'émission se produit à la fois par les narines et par la bouche.

On classe principalement les consonnes en fonction de leur mode d'articulation, de leur lieu d'articulation, et de leur nasalisation.

➢ Les fricatives non voisées résultent d'une turbulence créée par le passage de l'air dans une constriction du conduit vocal.

➢ Les fricatives voisées font intervenir une source périodique liée à la vibration des cordes vocales qui vient s'ajouter à la source de bruit.

➢ Les consonnes occlusives correspondent à des sons essentiellement dynamiques. Une forte pression est créée en amont d'une occlusion maintenue en un certain point du conduit vocal puis relâchée brusquement. La période d'occlusion est appelée la *phase de tenue*.
Pour les occlusives voisées, un son basse fréquence est émis par vibration des cordes vocales pendant la phase de tenue et pour les occlusives non voisées, la phase de tenue est un silence.

La figure 2.8 représente les voyelles dans le plan des deux premiers formants. On observe un certain recouvrement dans les zones de dispersion ; les trois premiers formants caractérisent beaucoup mieux toutes les voyelles.

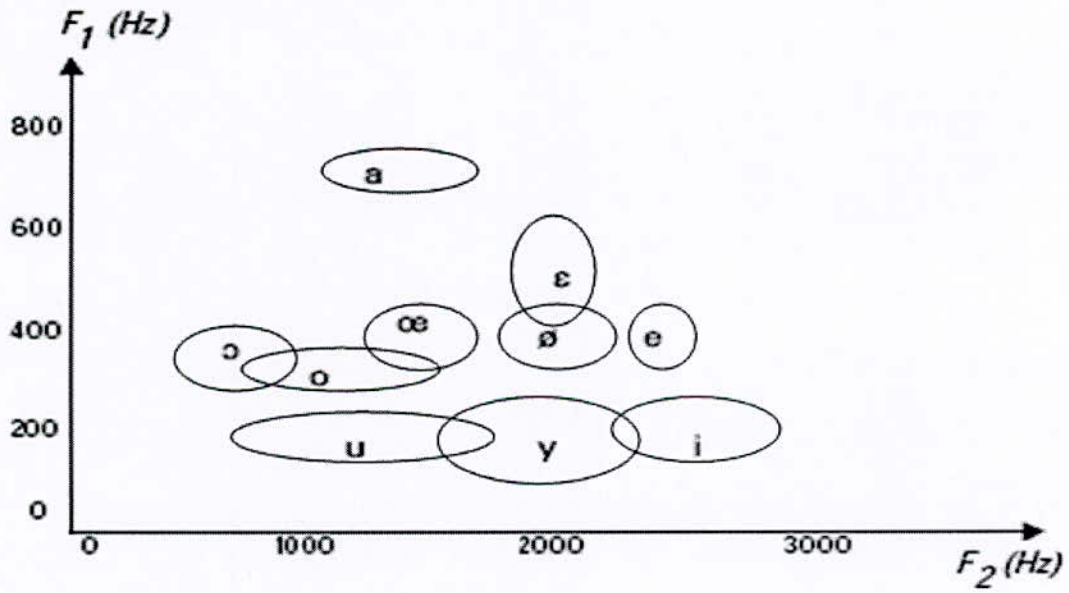


FIG. 2.8 Représentation des voyelles dans le plan F_1-F_2

2.1.2 Modélisation autorégressive du signal vocal

La modélisation du signal vocal $x(n)$ consiste en l'estimation des paramètres d'un filtre linéaire $H(z)$ qui, soumis à une excitation particulière $u(n)$, reproduit ce signal le plus fidèlement possible.

L'objectif de cette modélisation étant la réduction du nombre de paramètres décrivant le signal $x(n)$, simplifiant ainsi son enregistrement, transmission ou sa reproduction.

Le modèle AR est une modélisation mathématique basée sur la mise en équation simplifiée du modèle physique et aboutissant à une transmitance $H(z)$, dite *tous-pôles,* du système.

L'excitation du conduit vocal, idéalisée, est soit un bruit blanc (sons non voisés), soit un train périodique d'impulsion (sons voisés).

Le conduit vocal lui est modélisé par une succession de tubes acoustiques, c'est à dire une cascade de résonateurs.

Au final, le modèle AR consiste à dire que le son X est le résultat du filtrage par un filtre *tous-pôles* H d'une source U qui est soit une bruit blanc gaussien centré, soit un train périodique d'impulsion ayant pour fréquence le pitch.

En terme de transmitance, on obtient :

$$X(z) = U(z)\ H(z) \tag{2.1}$$

Avec :

$$H(z) = \frac{\sigma}{A(z)} \tag{2.2}$$

$U(z)$: L'excitation (bruit blanc ou train périodique d'impulsions).

σ : Le gain du modèle.

$$A(z) = \sum_{i=0}^{p} a_p(i)\ z^{-i}\ ,\ a_p(0) = 1 \tag{2.3}$$

$a_p(i)$: Coefficients de prédiction linéaire.

p : Ordre du modèle.

Ce modèle de production d'un signal est appelé *modèle autorégressif*; en effet à l'équation (2.1) correspond dans le domaine temporel la récurrence linéaire suivante :

$$x(n) + \sum_{i=1}^{p} a_p(i)\, x(n-i) = \sigma\, u(n) \qquad (2.4)$$

qui exprime qu'un échantillon quelconque x(n) est une combinaison linéaire des p échantillons qui le précèdent plus le terme d'excitation.

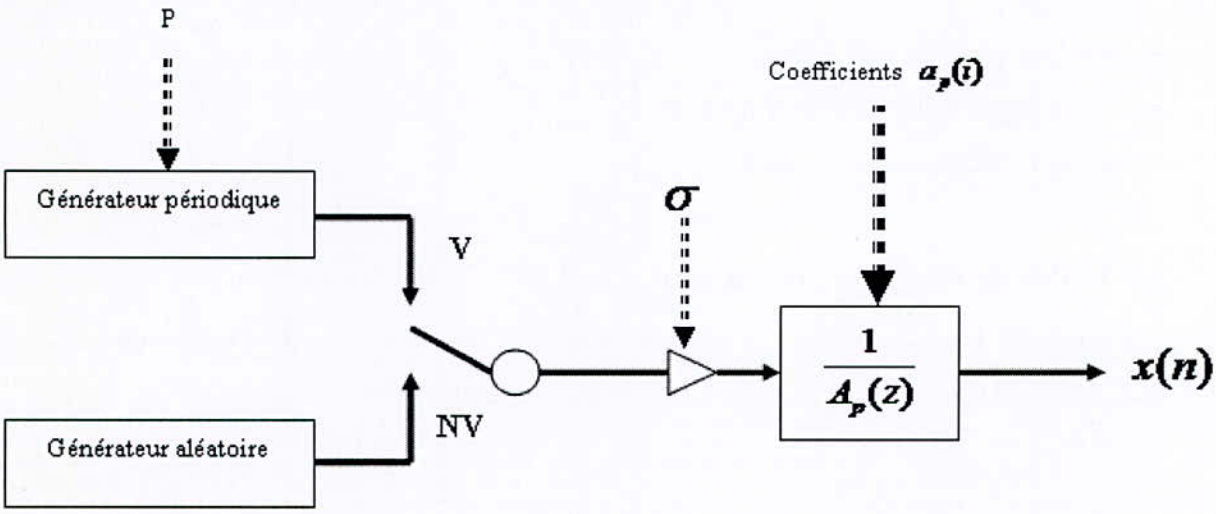


FIG. 2.9 Modèle autorégressif de production de la parole

Comme l'illustre la figure 2.9, la définition du modèle AR décrit plus haut revient à chercher les paramètres suivants : le pitch, la décision V/NV, le gain et les coefficients de prédiction.

La modélisation autorégressive du signal vocal n'est valable que dans la mesure où la condition de stationnarité est vérifiée.

En raison que le signal vocal ne peut être considéré comme quasi stationnaire que sur des intervalles de temps de durée limitée, on est amené à considérer des tranches successives et à estimer un modèle AR pour chacune d'elles ; une procédure usuelle consiste à effectuer l'analyse sur des tranches de 20 *ms* avec décalage de 10 *ms* d'une tranche à la suivante.

Dans l'annexe B, nous avons exposé une méthode d'estimation du modèle AR basée sur l'autocorrélation, et on a optimisé le calcul du modèle par l'algorithme de Levinson-Durbin qui profite de la structure particulière de la matrice d'autocorrélation pour réduire la complexité des calculs.

2.2 Analyse et paramétrisation du signal vocal

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.

2.2.1 Pré-traitement acoustique

Le calcul des paramètres acoustiques passe par une phase de pré-traitement contenant deux étapes, la pré-accentuation acoustique et le fenêtrage.

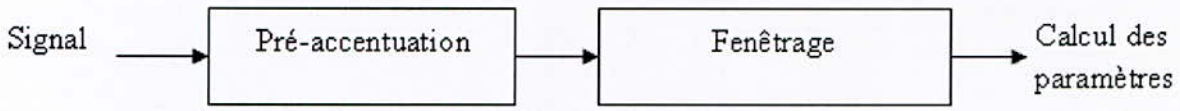


FIG. 2.10 Pré-traitement et extraction des paramètres

2.2.1.1 La pré-accentuation

L'onde acoustique sortante des lèvres subit, à cause de la désadaptation entre les deux milieux intérieur et extérieur, une distorsion assimilable à une désaccentuation de 6 *dB* par octave sur tout le spectre. Pour pouvoir compenser cette distorsion, et accentuer les hautes fréquences, on applique un filtre de pré-accentuation passe haut de transmitance :

$$H(z) = 1 - \alpha z^{-1} \tag{2.5}$$

avec $0.9 \leq \alpha \leq 1$.

2.2.1.2 Le fenêtrage

L'étape de fenêtrage consiste à appliquer au signal vocale une fenêtre glissante de durée limitée, et ce afin de limiter le nombre d'échantillons et de réduire les effets de bords (phénomène de Gibbs).

Parmi les différentes fenêtres de pondération, les plus utilisées sont : la fenêtre rectangulaire, la fenêtre de Hamming, la fenêtre de Hanning et la fenêtre de Blackmann. En traitement de la parole, la fenêtre de Hamming est la plus utilisée.

La fenêtre de Hamming est donnée par l'expression :

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad , 0 \le n \le N-1 \tag{2.6}$$

N : Le nombre d'échantillons dans une fenêtre.

2.2.2 Les paramètres acoustiques

2.2.2.1 L'énergie du signal

L'énergie du signal est un indice qui peut contribuer à augmenter les performances d'un système de reconnaissance, elle est calculée directement dans le domaine temporel et sur chaque trame du signal par :

$$E = \sum_{n=0}^{N-1} s^2(n) \tag{2.7}$$

Comme paramètre acoustique, on peut aussi utiliser l'énergie logarithmique qui est définie comme suit :

$$E = \ln\left(\sum_{n=0}^{N-1} s^2(n)\right) \tag{2.8}$$

où N est le nombre d'échantillons du signal, et les $s(n)$ sont les échantillons du signal.

L'énergie ainsi obtenue est sensible au niveau d'enregistrement ; on choisit en général de la normaliser, et d'exprimer sa valeur en décibels par rapport à un niveau de référence.

2.2.2.2 Les coefficients de prédiction linéaire LPC

Les coefficients LPC découlent directement du modèle de production de la parole. Pour rappel, le signal de parole $s(n)$ peut être modélisé comme étant la réponse d'un filtre tous pôles $H(z)$ à une excitation $u(n)$ qui peut être soit un train périodique d'impulsions, soit un bruit blanc gaussien.

On a vu que $H(z)$ peut se mettre sous la forme :

$$H(z) = \frac{\sigma}{A(z)} \tag{2.9}$$

avec

$$A(z) = 1 + \sum_{i=1}^{P} a_i \; z^{-i} \tag{2.10}$$

La modélisation étant faite, il convient à présent d'estimer les coefficients de prédiction a_i ainsi que le gain σ du système. L'estimation est fondée soit sur le calcul de la matrice de covariance, soit sur le calcul de la matrice d'autocorrélation (Annexe B).

2.2.2.3 Les coefficients cepstraux de prédiction linéaire LPCC

Les coefficients cepstraux peuvent être calculés à partir de la sortie d'un banc de filtres ou à partir des coefficients de prédiction linéaire, ainsi les coefficients LPCC (*Linear Prediction Cepstral Coefficients*) sont dérivés directement des coefficients LPC.

Les coefficients cepstraux c_k sont obtenus par :

$$c_k = - a_k - \sum_{i=1}^{k-1} (1 - \frac{i}{k}) a_i \; c_{k-i} \qquad , k > 0 \tag{2.11}$$

2.2.2.4 Les coefficients MFCC (Mel Frequency Cepstral Coefficients)

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 glottique de celle du conduit vocal.

Le cepstre du signal de parole est défini comme étant 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.

$$s(t) = e(t) * h(t) \tag{2.12}$$

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 (2.12) du logarithme du module de la transformée de Fourier donne :

$$\log|S(f)| = \log|E(f)| + \log|H(f)| \tag{2.13}$$

Par une transformée de Fourier inverse, on obtient :

$$s'(cef) = e'(cef) + h'(cef) \tag{2.14}$$

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*.

Ce domaine est intéressant pour faire la séparation des contributions du conduit vocal et de la source d'excitation dans le signal de parole. En effet, si les contributions relevant du conduit vocal et les contributions de la source d'excitation évoluent avec des vitesses différentes dans le temps, alors il est possible de les séparer par application d'une simple fenêtre dans le domaine quéfrentiel (liftrage passe-bas) pour le conduit vocal.

Les coefficients cepstraux les plus répandus sont les MFCC (Mel Frequency Cepstral Coefficients). Il présentent l'avantage d'être faiblement corrélés entre eux, et qu'on peut donc approximer leur matrice de covariance par une matrice diagonale.

Pour simuler le fonctionnement du système auditif humain, les fréquences centrales du banc de filtres sont réparties uniformément sur une échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'information utile dans le signal de parole. Les échelles perceptives les plus utilisées sont l'échelle Mel et l'échelle Bark.

- Echelle Mel

$$Mel(f) = 2595\ log\left(1 + \frac{f}{700}\right) \tag{2.15}$$

- Echelle Bark

$$Bark(f) = 6\ Arc\sinh\left(\frac{f}{1000}\right) \tag{2.16}$$

f représente la fréquence [Hz].

La procédure de calcul des coefficients MFCC est illustrée sur la figure 2.11

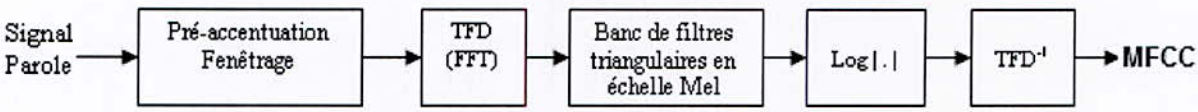


FIG. 2.11 Calcul des coefficients MFCC

Soit un signal discret $s(n)$ avec $0 \le n \le N-1$, N est le nombre d'échantillons d'une fenêtre d'analyse, F_s est la fréquence d'échantillonnage, la transformée de Fourier discrète court terme $S(k)$ est obtenue avec la formule :

$$S(k) = \sum_{n=0}^{N-1} s(n)\ exp\left(\frac{-j\,2\,\pi\,n\,k}{N}\right),\ 0 \le k \le N-1 \tag{2.17}$$

Le spectre du signal est filtré par un banc de filtres triangulaires, dont les bandes passantes sont de même largeur dans le domaine des fréquences Mel. Les points de frontières B_m des filtres en échelle de fréquence Mel sont calculés à partir de la formule :

$$B_m = B_b + m \frac{B_h - B_b}{M + 1} \quad , \quad 0 \le m \le M + 1 \tag{2.18}$$

M : Le nombre de filtres.

B_h : La fréquence la plus haute du signal.

B_b : La fréquence la plus basse du signal.

Dans le domaine fréquentiel, et d'après (2.15), les points f_m discrets correspondants sont calculés d'après :

$$f_m = B^{-1}\left(B_b + m \frac{B_h - B_b}{M + 1} \right) \tag{2.19}$$

Où $B^{-1}(x)$ désigne la fréquence correspondante à la fréquence x sur l'échelle Mel,

$$B^{-1}(x) = 700 \left(10^{\frac{x}{2595}} - 1 \right) \tag{2.20}$$

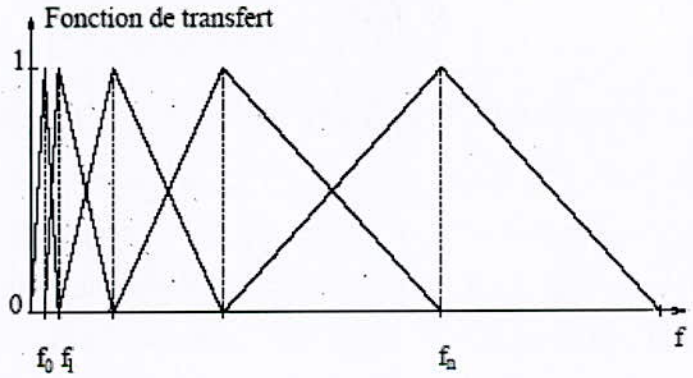


FIG. 2.12 Banc de filtres sur l'échelle linéaire

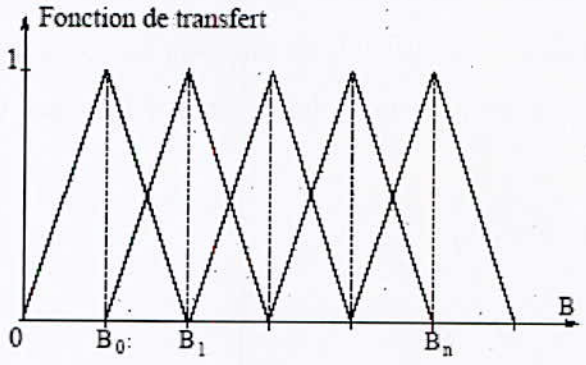


FIG. 2.13 Banc de filtres sur l'échelle Mel

Les coefficients cepstraux de fréquence en échelle Mel *(MFCC)* peuvent être obtenus par une transformée de Fourier inverse à partir des énergies d'un banc de filtres. Les d premiers coefficients cepstraux peuvent être calculés directement à partir du logarithme des énergies E_i issues d'un banc de M filtres par la transformée en cosinus discrète définie comme :

$$c_k = \sum_{i=1}^{M} log\, E_i \; cos\left[\frac{\pi k}{M}\left(i-\frac{1}{2}\right)\right] \quad , 1 \le k \le d \tag{2.21}$$

et qui permet d'obtenir des coefficients peu corrélés.

Le coefficient c_0 qui est la somme des énergies n'est pas utilisé ; il est éventuellement remplacé par le logarithme de l'énergie totale E calculée dans le domaine temporel et normalisée.

2.2.2.5 Les coefficients LFCC (Linear Frequency Cepstral Coefficients)

Aux coefficients MFCC s'ajoute un autre type de paramètres, les LFCC (*Linear Frequency Cepstral Coefficients*) qui sont calculés de la même manière que les MFCC, mais avec la seule différence que les fréquences des filtres sont uniformément réparties sur l'échelle linéaire des fréquences, et non pas sur une échelle perceptive de type Mel.

2.2.2.6 Les coefficients différentiels

Pour prendre en compte la dynamique temporelle du signal de parole, on utilise en plus des paramètres cités précédemment, des coefficients différentiels du premier ordre et du second ordre issus des coefficients cepstraux ou de l'énergie. Soit $c_k(t)$ le coefficient cepstral d'indice k de la trame t, alors le coefficient différentiel $\Delta c_k(t)$ correspondant est calculé sur $2n_\Delta + 1$ trames par :

$$\Delta c_k(t) = \frac{\sum_{i=-n_\Delta}^{n_\Delta} i\, c_k(t+i)}{\sum_{i=-n_\Delta}^{n_\Delta} i^2} \tag{2.22}$$

La dérivée première de l'énergie ΔE est calculée de la même façon :

$$\Delta E(t) = \frac{\sum_{i=-n_\Delta}^{n_\Delta} i\, E(t+i)}{\sum_{i=-n_\Delta}^{n_\Delta} i^2} \tag{2.23}$$

Les coefficients différentiels du second ordre peuvent aussi contribuer à l'amélioration des systèmes de reconnaissance. Les coefficients $\Delta\Delta c_k$ et $\Delta\Delta E$ sont calculés par régression linéaire des coefficients Δc_k et ΔE respectivement, et sur $n_{\Delta\Delta}$ (typiquement $n_\Delta = n_{\Delta\Delta} = 2$).

2.2.3 Réduction du nombre de coefficients

L'utilisation de la totalité des composantes des vecteurs acoustiques est coûteuse en temps de calcul et des ressources CPU et mémoire. En classant les coefficients acoustiques selon un critère particulier, il est possible de ne considérer que certains coefficients.

Des analyses sont proposées pour réduire la dimension de l'espace des paramètres, comme le critère de Fisher (*FDR*), l'analyse en composantes principales (*ACP*) , ou l'analyse linéaire discriminante (*ALD*).

2.2.3.1 Le rapport discriminant de Fisher (FDR)

Ce rapport estime la capacité discriminante de chaque paramètre, en mesurant le chevauchement de leurs fonctions de densité de probabilité.

Le rapport discriminant de Fisher pour des fonctions de densités de probabilité gaussiennes, peut être calculé pour chaque paramètre comme suit :

$$FDR = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} (\bar{c}[i] - \bar{c}[j])^2}{\sum_{i=1}^{k} Var(\bar{c})[i]} \tag{2.24}$$

où $\bar{c}[i]$ désigne la moyenne du paramètre c pour le locuteur i et $Var(c)[i]$ la variance du paramètre c pour le locuteur i.

Ce paramètre peut être interprété comme étant le rapport de la variabilité inter-locuteurs du paramètre par la variabilité intra-locuteur du même paramètre.

Avec ce critère, les paramètres pertinents peuvent être sélectionnés. Le désavantage de ce critère est qu'il n'intègre pas les relations de corrélation entre les paramètres.

2.2.3.2 Analyse Linéaire Discriminante (ALD)

Elle consiste à appliquer une transformation linéaire sur chaque vecteur acoustique. Cette transformation peut décorréler les paramètres et augmenter leurs capacités discriminantes.

La transformation du vecteur acoustique est effectuée à l'aide d'une matrice A comme suit :

$$P_{tr} = A\,P \tag{2.25}$$

où P_{tr} est le vecteur acoustique transformé et P est le vecteur acoustique initial.

La matrice A prend la forme suivante :

$$A = [u_1, u_2, ..., u_D]^t \tag{2.26}$$

La matrice A est appelée « la matrice de covariance de Fisher ». Cette matrice est égale au rapport entre la matrice de dispersion inter-locuteurs S_b et la matrice de dispersion intra-locuteur S_w.

$$A = S_w^{-1} S_b \tag{2.27}$$

Cette réduction de la dimension de l'espace des paramètres peut entraîner une dégradation du taux d'identification, mais elle permet un gain considérable en temps de calcul et en espace mémoire.

2.2.3.3 Analyse en composantes principales (ACP)

On peut aussi effectuer une analyse ACP pour décorréler les coefficients issus d'un banc de filtres, ce qui permet de représenter ensuite la dispersion des coefficients avec des matrices de covariance diagonales.

Elle consiste à appliquer une transformation linéaire sur chaque vecteur acoustique. Cette transformation peut décorréler les paramètres et augmenter leurs capacités discriminantes.

La transformation du vecteur acoustique est effectuée à l'aide d'une matrice A comme suit :

$$P_{tr} = A\,P \tag{2.28}$$

où P_{tr} est le vecteur acoustique transformé et P est le vecteur acoustique initial.

La matrice A prend la forme suivante :

$$A = [u_1, u_2, ..., u_D]^t \tag{2.29}$$

Où les u_i désignent les vecteurs propres de la matrice de covariance ordonnés de manière décroissante.

2.2.4 Distances et mesures de dissemblance dans l'espace acoustique

2.2.4.1 Définitions et propriétés

Dans toute approche de reconnaissance, le choix de la distance associée à l'espace des paramètres est important. Il est possible d'utiliser toutes les distances classiques, en particulier les distances de Minkovski, parmi lesquelles la distance euclidienne, et la distance de Mahalanobis qui normalise les coefficients par leur matrice de covariance.
Des distances spécifiques aux espaces de représentation de la parole existent aussi, comme les distances cepstrales pondérées et la mesure d'Itakura pour les coefficients LPC.
Dans un espace métrique, la distance entre deux vecteurs X et Y doit satisfaire les conditions suivantes :

1. $d(X,Y) \geq 0$
2. $d(X,Y) = d(Y,X)$
3. $d(X,Y) \leq d(X,U) + d(U,Y)$

En traitement de la parole, ces conditions ne sont pas toujours respectées par les distances utilisées, et c'est pour cette raison qu'on préfère parler de mesures de dissemblance ou de mesures de distorsion.

La condition 2 peut être assurée en posant

$$d_S(X,Y) = \frac{1}{2}[d(X,Y) + d(Y,X)] \tag{2.30}$$

La condition 3 est rarement utile en traitement de la parole.

2.2.4.2 Distances usuelles

1. Distances de Minkovski

Les distances de Minkovski entre deux vecteurs $X = (x_1,....,x_D)^t$ et $Y = (y_1,....,y_D)^t$ sont données par :

$$L_r(X,Y) = \left(\sum_{k=1}^{D} |x_k - y_k|^r \right)^{1/r} \tag{2.31}$$

Les distances les plus courantes sont la distance de Manhattan pour $r = 1$, la distance du max pour $r = \infty$, et la distance euclidienne d_E pour $r = 2$.

a. Distance euclidienne

La distance euclidienne donne la même importance à chacun des coefficients, et elle est définie par :

$$d_E^2(X,Y) = (X-Y)^t (X-Y) = \sum_{k=1}^{D} (x_k - y_k)^2 \tag{2.32}$$

b. Distance de Mahalanobis

Si l'on dispose d'un ensemble de n paramètres $\{X_i\}_{1 \le i \le n}$, il est possible d'estimer leurs vecteurs moyen $\hat{\mu}$ comme suit :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{2.33}$$

et leur matrice de covariance se calcule par :

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^t - \hat{\mu}\hat{\mu}^t \tag{2.34}$$

La distance de Mahalanobis entre deux vecteurs de paramètres est définie comme :

$$d_M^2(X,Y) = (X-Y)^t \hat{\Sigma}^{-1} (X-Y) \tag{2.35}$$

et permet de décorréler linéairement les coefficients.

Si les coefficients ne sont pas corrélés, $\hat{\Sigma}$ devient une matrice diagonale, et la distance de Mahalanobis devient une distance euclidienne pondérée par l'inverse des variances des coefficients :

$$d_p^2(X,Y) = \sum_{k=1}^{D} [w_k(x_k - y_k)]^2 \tag{2.36}$$

avec $w_k = \frac{1}{\sigma_k}$

2.2.4.3 Distances adaptées à une représentation

1. Mesure d'Itakura

Les distances classiques ne sont pas adaptées à la comparaison des modèles autorégressifs. Si $A = (1, a_1, ..., a_p)^t$ et $B = (1, b_1, ..., b_p)^t$ sont des vecteurs de coefficients LPC d'ordre p, la mesure d'Itakura est définie comme :

$$d_I(A,B) = log\left[\frac{A^t R_b A}{B^t R_b B}\right] \tag{2.37}$$

R_b : matrice d'autocorrélation du signal produit par le modèle B.

2. Distance cepstrale

Soit la fonction $f(\theta)$ représentant une densité spectrale d'énergie $P_x(\theta)$ ou le spectre du modèle $P_M(\theta)$; la différence logarithmique entre deux spectres vaut :

$$V(\theta) = \ln f(\theta) - \ln f'(\theta) \tag{2.38}$$

et la distance spectrale logarithmique est la norme :

$$d_p = \left[\frac{1}{2\pi} \int_{-\pi}^{\pi} |V(\theta)|^p \, d\theta \right]^{1/p} \tag{2.39}$$

La norme d_2 est la plus utilisée. Toutefois, la distance spectrale logarithmique donnée par (2.39) est coûteuse en temps de calcul, alors on lui substitue la distance cepstrale.
Les coefficients du cepstre réel sont donnés par :

$$\ln f(\theta) = \sum_n c(n) \exp(-jn\theta) \tag{2.40}$$

En remplaçant p par 2 dans (2.39), on obtient :

$$d_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \sum_n \left(c(n) - c'(n) \right) \exp\left(- jn\theta \right) \right|^2 d\theta$$

$$d_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[\sum_n \left(c(n) - c'(n) \right) \exp\left(- jn\theta \right) \sum_m \left(c(m) - c'(m) \right) \exp\left(- jm\theta \right) \right]^2 d\theta$$

$$d_2^2 = \sum_l \left(c(l) - c'(l) \right)^2 = \left(c(0) - c'(0) \right)^2 + 2 \sum_{l=1}^{\infty} \left(c(l) - c'(l) \right)^2 \tag{2.41}$$

La distance cepstrale est basée sur un nombre fini L de termes :

$$d_{CEP} = \left(c(0) - c'(0) \right)^2 + 2 \sum_{l=1}^{L} \left(c(l) - c'(l) \right)^2 \tag{2.42}$$

2.3 Conclusion

Dans ce chapitre, on a, en premier lieu, abordé la modélisation autorégressive de la production du signal vocal. Ensuite, nous avons cité les différentes représentations du signal vocal ainsi que quelques techniques de réduction de la dimension de l'espace acoustique. Enfin, nous avons cité les distances et mesures de dissemblances les plus utilisées dans le domaine du traitement de la parole.

Après études des différents types de représentations acoustiques, on constate que les coefficients MFCC sont les plus adaptés pour caractériser l'identité du locuteur. Pour cela, ils sont les coefficients les plus utilisés en reconnaissance du locuteur en mode indépendant du texte.

Chapitre 3

Identification du locuteur par mélanges de gaussiennes

Au cours de ce chapitre nous nous intéressons essentiellement à l'identification du locuteur en mode indépendant du texte en utilisant la modélisation par mélanges de gaussiennes qui fournie de bonnes performances et qui constitue l'état de l'art en la matière.

Premièrement, nous présentons les différentes approches de modélisation en reconnaissance automatique du locuteur : l'approche vectorielle, l'approche statistique, l'approche connexionniste et enfin l'approche relative, sans pour autant entrer dans les détails.

Dans la seconde partie de ce chapitre, nous allons présenter en détail : la technique GMM, les différents algorithmes d'apprentissage utilisés, la stratégie de décision adoptée ainsi que le protocole d'évaluation des performances des systèmes d'identification du locuteur.

Dans la dernière partie, nous allons introduire une nouvelle technique de modélisation : OGMM, qui fournit les meilleures performances en contre partie d'une orthogonalisation de l'espace des paramètres acoustiques.
Les algorithmes d'apprentissage et le protocole d'évaluation utilisés par cette technique sont les mêmes que ceux utilisés pour la GMM et pour cela nous nous contentons seulement de présenter le principe de cette méthode.

3.1 Modélisation des locuteurs

Les différentes approches de modélisation des locuteurs sont classées en quatre grandes familles.

- L'approche vectorielle : le locuteur est représenté par un ensemble de vecteurs issus directement de la phase de paramétrisation. Ses principales techniques sont la reconnaissance à base de l'alignement temporel dynamique (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 (HMM), par les mélanges de gaussiennes (GMM) 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 non pas de façon absolue mais relativement par rapport à d'autres locuteurs de référence, dont les modèles sont bien déterminés.

3.1.1 L'approche vectorielle

3.1.1.1 L'Alignement Temporel Dynamique (DTW : Dynamic Time Warping)

Utilisée en mode dépendant du texte, cette technique effectue la comparaison entre la forme d'entrée à reconnaître et une ou plusieurs formes de référence en calculant la distance entre les paramètres des deux formes. Elle détermine le meilleur chemin reliant le début et la fin des deux blocs de paramètres. Ainsi cet algorithme permet de trouver un alignement temporel optimal entre la forme d'entrée et la forme de référence. Cet alignement est réalisé par une technique de programmation dynamique.

Malgré les bonnes performances obtenues par cette technique, elle reste très sensible à la qualité de l'alignement et notamment le choix du point de départ des deux formes à comparer.

3.1.1.2 La Quantification Vectorielle

Cette technique permet une compression considérable des données, elle consiste à représenter l'espace acoustique par un nombre fini de vecteurs acoustiques formant un dictionnaire, et ce en faisant un partitionnement de cet espace en régions ou classes, qui seront représentées par leurs vecteurs centroïdes.

En reconnaissance de locuteur ce dictionnaire est réalisé à partir des vecteurs de paramètres issus de la phase de paramétrisation. Les performances et la rapidité de cette technique dépendent fortement de la taille du dictionnaire. En effet, plus la taille du dictionnaire est grande meilleures sont les performances, mais le processus de test devient trop lent.

3.1.2 L'approche statistique

3.1.2.1 Les Modèles de Markov Cachés HMM : Hidden Markov Models

Le modèle HMM est introduit dans un cadre purement statistique, il s'est ensuite imposé en reconnaissance de la parole avant d'être appliqué en reconnaissance automatique du locuteur.

Le modèle HMM présente différents avantages : clarté, rigueur, efficacité et généralité.

Un modèle HMM se caractérise par un système à états comportant deux processus.

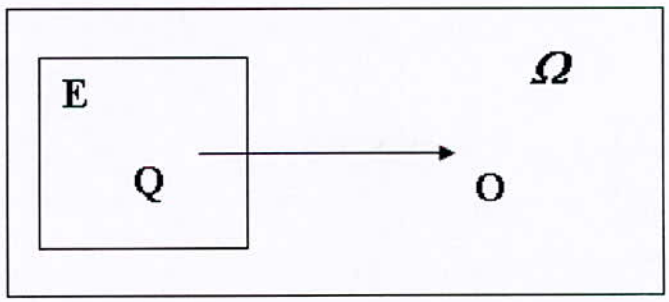


FIG. 3.1 Constituants d'un HMM

Les réalisations du premier processus sont des chaînes cachées $Q = q_1 q_2 .....q_T$ des états du système avec un état initial q_1 et un état final q_T.

Les réalisations du second processus sont des chaînes externes ou observations $O = o_1 o_2 .....o_T$ où chaque o_t est un élément d'un espace d'observation Ω.

Dans une modélisation par HMM, on suppose que la suite des vecteurs acoustiques d'observation est stationnaire par blocs. Ainsi, les vecteurs acoustiques d'un bloc suivent la même loi de probabilité. La modélisation d'un bloc de vecteurs acoustiques représente un état du modèle HMM. Dans cette approche, chaque entité est modélisée par une machine d'états (automate), appelée machine Markovienne et qui est composée d'un ensemble d'états et de transitions qui permettent de passer d'un état à un autre. Un modèle HMM est un modèle statistique séquentiel qui suppose que les caractéristiques observées forment une succession d'états distincts.

Soit λ un modèle Markovien de N états et $Q=(q_1,q_2,...,q_T)$ une séquence d'états correspondant à l'observation $O=(o_1,o_2,.....,o_T)$ où q_t est le numéro de l'état atteint par le processus à l'instant t. L'état du modèle de Markov λ qui correspond à o_t n'étant pas directement observable, on dit qu'il est caché. D'où le nom de modèle de Markov caché. La figure 3.2 représente un exemple de modèle de Markov. Un tel modèle est définit par :

- Un ensemble d'états cachés $\{S_1, S_2, \ldots, S_N\}$.
- Un ensemble d'observations $\{v_1, v_2, \ldots, v_M\}$.
- Probabilités de transition $a_{ij} = P(q_{t+1} = S_j / q_t = S_i)$.
- Probabilités d'observation $b_j(k) = P(o_t = v_k / q_t = S_j)$, qui sont en général des mélanges de gaussiennes.
- Un ensemble de probabilités initiales de se trouver dans chaque état $\pi = \{\pi_i / \pi_i = P(q_1 = S_i)\ i = 1, .., N\}$.

Un modèle de Markov caché est donc spécifié par un triplet $\lambda = \{A, B, \pi\}$ où A est la matrice des probabilités de transition, B la matrice des probabilités d'observation et π les probabilités initiales.

Problèmes des modèles HMM

Trois problèmes se posent avec les modèles de Markov cachés :

1. L'évaluation

Étant donné une séquence d'observations $O = o_1\ o_2\ ....\ o_T$ et un modèle $\lambda = \{A, B, \pi\}$, déterminer la probabilité que l'observation ait été engendrée par le modèle, $P(O/\lambda)$.

Il existe deux méthodes pour résoudre ce problème. La méthode dite directe et qui consiste à calculer cette probabilité en énumérant toutes les séquences d'états possibles de même longueur que la séquence d'observation. Cette technique demande beaucoup de temps de calcul. Un moyen plus rapide pour calculer cette probabilité est l'utilisation des algorithmes de programmation dynamique.

2. Estimation des états cachés

Le deuxième problème posé avec les HMM est le décodage qui consiste à chercher la séquence $Q = q_1 q_2 .... q_T$ d'état qui maximise la probabilité $P(O,Q/\lambda)$, étant donné une séquence d'observations $O = o_1 o_2 .... o_T$ et un modèle $\lambda = \{A, B, \pi\}$. Pour cela, l'algorithme de Viterbi est le plus utilisé. Il permet de chercher la séquence d'états cachés la plus probable en ne gardant que les états S_i qui maximisent la probabilité à chaque instant t.

3. Apprentissage

C'est le problème principal d'un modèle HMM. En effet, la qualité d'un système utilisant une modélisation HMM dépend principalement de la qualité de ses modèles. C'est pourquoi l'étape d'apprentissage qui consiste à estimer les paramètres des modèles HMM est très importante. Il existe plusieurs méthodes pour résoudre ce problème, les plus utilisées sont :

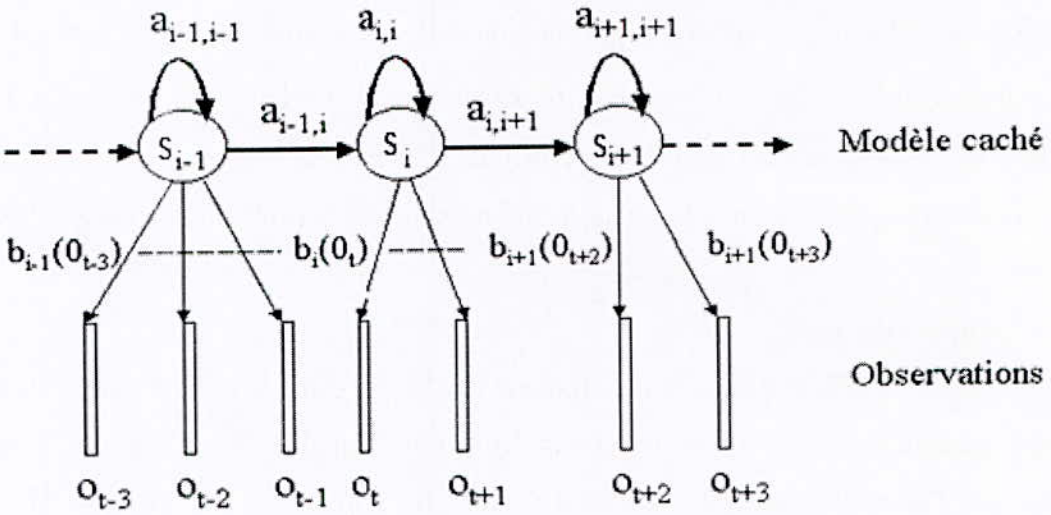
- L'algorithme de Viterbi associé à des estimateurs empiriques : l'algorithme de Viterbi sert à déterminer la séquence d'états cachés la plus vraisemblable, correspondant aux données d'apprentissage. Les paramètres des densités de probabilité de chaque état peuvent être alors ré-estimés en utilisant des estimateurs empiriques et les observations associées à chaque état le long du chemin de Viterbi.
- L'algorithme EM (Expectation-Maximisation) : Cet algorithme permet de résoudre le problème d'apprentissage en estimant de manière itérative les paramètres d'un modèle au sens du maximum de vraisemblance.

La phase de reconnaissance

La phase de reconnaissance consiste, étant donné une observation, à évaluer la probabilité qu'elle soit engendrée par chacun des modèles et sélectionner celui qui est le plus probable.

Le principal avantage de l'approche HMM est sa grande capacité d'apprendre les propriétés statistiques. En reconnaissance de locuteur le choix le plus fréquent consiste à utiliser un modèle dont la distribution conditionnelle dans chaque état est un mélange de gaussiennes.

L'utilisation de ces modèles est plus importante dans le mode dépendant du texte parce qu'en mode indépendant du texte l'information supplémentaire apportée par les transitions entre états n'améliore pas les performances de la reconnaissance du locuteur.



$$Q = (\ldots, q_{t-3} = e_{i-1}, q_{t-2} = e_{i-1}, q_{t-1} = e_{i-1}, q_t = e_i, q_{t+1} = e_i, q_{t+2} = e_{i+1}, q_{t+3} = e_{i+1}, \ldots)$$

FIG. 3.2 Exemple d'une machine Markovienne

3.1.2.2 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. Chaque composante gaussienne est supposée modéliser un ensemble de classes acoustiques.

Les GMM sont considérés comme un cas particulier des HMM et une extension de la quantification vectorielle.

3.1.2.3 Mesures statistiques du second ordre

Cette partie présente une famille de mesures de similarité entre locuteurs reposant sur les statistiques du second ordre (vecteur moyen et matrice de covariance) d'une séquence de vecteurs.

3.1.3 L'approche connexionniste

Les systèmes connexionnistes ou Réseaux de Neurones (***RN***), qui furent redécouverts et développés dans la fin des années 80, ont suscité beaucoup d'intérêt dans plusieurs domaines. Cette approche comprend une grande famille de méthodes très différentes. Chaque méthode est représentée par un réseau qui implémente une fonction de transfert globale spécifiée par l'architecture et les fonctions élémentaires du réseau.

Dans cette approche, un locuteur est représenté par un ou plusieurs réseaux de neurones appris directement des trames obtenues en phase de paramétrisation et permettant de le discriminer par rapport à un ensemble de locuteurs.

Les réseaux de neurones sont capables d'implanter des techniques discriminantes très efficaces et offrent des outils de classification qui permettent la séparation des classes de façon non linéaire. Néanmoins, ils restent incapables de résoudre leur principal problème qui est la durée d'apprentissage importante et nécessaire pour une grande population.

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.

3.1.4 L'approche relative

C'est une nouvelle technique, qui consiste à modéliser un locuteur non plus de façon absolue, mais relativement à un ensemble de locuteurs bien appris, en fait, chaque locuteur est représenté par sa localisation dans un espace de référence.

Cette technique trouve sont application lorsqu'on dispose de peu de données d'apprentissage. Il faut donc estimer avec très peu de données un modèle robuste du locuteur, qui permettra sa reconnaissance.

Cette approche a donné naissance à la notion d'espace de locuteurs, où un locuteur est représenté par une combinaison linéaire des modèles de référence, ce qui réduit considérablement le nombre de paramètres. Cette approche repose sur le principe d'utiliser des connaissances a priori obtenues a partir de l'ensemble des locuteurs de référence.

3.2 Identification du locuteur par mélanges de gaussiennes standards (GMM)

3.2.1 Les mélanges de gaussiennes

Les mélanges de gaussiennes sont utilisés pour modéliser un locuteur donné par une somme pondérée de composantes gaussiennes. Cette méthode est la plus utilisée en ce qui concerne la reconnaissance du locuteur en mode indépendant du texte.

L'utilisation d'un modèle GMM se justifie essentiellement en faisant appel à l'interprétation des classes du mélange. En effet, 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 ième composante pourra être représentée par sa moyenne et sa matrice de covariance. 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 composantes gaussiennes, on peut représenter une large gamme de distributions.

3.2.2 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 :

$$p(x/\lambda_s) = \sum_{m=1}^{M} \pi_m^s \, b_m^s(x) \tag{3.1}$$

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

$$b_m^s(x) = \frac{1}{(2\pi)^{D/2}.\left|\Sigma_m^s\right|^{1/2}} \exp\left[-\frac{(x-\mu_m^s)'(\Sigma_m^s)^{-1}(x-\mu_m^s)}{2}\right] \tag{3.2}$$

et les π_m^s représentent les poids du mélange, avec :

$$\sum_{m=1}^{M} \pi_m^s = 1 \tag{3.3}$$

Un locuteur est donc modélisé par un ensemble de paramètres noté λ_s :

$$\lambda_s = \left\{ \pi_m^s \ , \ \mu_m^s \ , \ \Sigma_m^s \right\}_{m=1,\ldots,M}$$

Ce modèle peut prendre plusieurs formes, notamment en ce qui concerne les matrices de covariance. On peut utiliser une matrice de covariance pour chaque gaussienne, ou bien une matrice de covariance globale, commune à toutes les gaussiennes.

3.2.3 Apprentissage du modèle

La phase d'apprentissage consiste à estimer l'ensemble λ des paramètres d'un modèle GMM pour chaque locuteur, et ce à partir des vecteurs acoustiques issus de la phase de paramétrisation.

Dans cette partie nous présentons trois algorithmes utilisés pour l'apprentissage. En premier lieu, nous introduisons des modifications sur deux algorithmes de quantification vectorielle : K-moyennes et LBG, et ce pour pouvoir les adapter à l'apprentissage des modèles GMM. En second lieu, nous présentons l'algorithme EM (Expectation-Maximisation) qui maximise la vraisemblance du modèle de façon itérative et garantie la convergence vers un maximum local.

Avant d'aborder la partie apprentissage, nous rappelons tout d'abord ce qui est la quantification vectorielle.

3.2.3.1 Quantification Vectorielle

La quantification vectorielle est une généralisation de la quantification scalaire. Elle consiste à substituer à un vecteur x, dont les composantes sont à valeurs réelles continues ($x \in R^k$), un vecteur voisin appartenant à l'ensemble fini $\left\{ y_i \in R^k \ , \ i = 1, 2, \ldots, M \right\}$.

Les vecteurs y_i sont dits vecteurs quantifiés, et constituent un dictionnaire (code-book) de points dans R^k.

Le dictionnaire est organisé de manière à minimiser la distorsion moyenne (moyenne des erreurs de quantification).

Construire un système de quantification vectorielle consiste à opérer une partition de l'espace R^k en classes C_i, représentées par leurs vecteurs centroïdes y_i.Chaque vecteur $x \in C_i$ sera représenté par le centroïde associé y_i.

3.2.3.2 Algorithme K-moyennes

Soit une séquence de vecteurs d'apprentissage $X = \{ x_1, x_2, \ldots\ldots, x_T \}$ que l'on désire répartir en M classes. On désignera par :

- $x_t^{(i)}$: les vecteurs appartenant à la classe i.
- μ_i : le centroïde de la classe i.
- T_i : le nombre de vecteurs appartenant à la classe i.
- $d (x_t^{(i)} , \mu_i)$: distance ou mesure de distorsion entre $x_t^{(i)}$ et μ_i.
- D_i : la distorsion totale de la classe i.

$$D_i = \sum_t d (x_t^{(i)} , \mu_i)$$

- D : la distorsion totale.

$$D = \sum_{i=1}^{M} D_i$$

Un nombre M de classes étant imposé à priori, le problème consiste à trouver la partition et les centroïdes de façon à minimiser la distorsion totale D.

Une procédure itérative peut être basée sur les deux observations suivantes :

- Pour un ensemble donné de centroïdes, la partition qui minimise D est celle pour laquelle chaque vecteur x_t est affecté à la classe dont le centroïde est le plus proche.
- Pour une partition donnée, il existe pour chaque classe i un vecteur qui minimise la distorsion totale D_i de la classe.

L'algorithme qui en résulte est celui de LLOYD généralisé, appelé aussi K-moyennes.

a. On fait le choix de M centroïdes répartis en principe d'une façon aléatoire dans R^k.

b. On affecte l'ensemble des vecteurs aux diverses classes sur la base de calcul de distances. Chaque vecteur x_t, $1 \leq t \leq T$, est affecté à la classe i si et seulement si $\| x_t - \mu_i \| < \| x_t - \mu_k \| \quad \forall\ k \neq i\ ,\ 1 \leq i, k \leq M$.

c. On recalcule la nouvelle position de chaque centroïde μ_i pour minimiser chaque distorsion D_i.

d. On calcule la distorsion totale D.

e. On répète les étapes (b), (c) et (d) jusqu'à ce que D varie de moins de $\varepsilon\%$ d'une itération à la suivante (par exemple $\varepsilon = 1$).

f. Les poids du mélange sont déterminés en calculant les proportions des vecteurs associés à chaque classe, $\pi_i = \frac{T_i}{T}$, $i = 1, 2, ..., M$ et la matrice de covariance d'une classe est égale à la matrice de covariance des vecteurs appartenant à cette classe.

3.2.3.3 Algorithme LBG

L'algorithme LBG (Linde Buzo Gray) garde beaucoup de caractéristiques de l'algorithme K-moyennes. Comme pour la K-moyennes, l'algorithme LBG est développé pour la quantification vectorielle. Son objectif est de minimiser la distorsion totale donnée par :

$$D = \sum_{i=1}^{M} \sum_{t=1}^{T} \| x_t - \mu_i \|$$

L'algorithme LBG part d'une seule classe pour atteindre M classes par éclatement binaire. A chaque itération, le nombre de classes double, donc le nombre de sous ensembles est la taille M du dictionnaire et ce sera une puissance de 2, $M = 2^p$.

Pour l'apprentissage des modèles GMM avec l'algorithme LBG, on garde les mêmes notations utilisées dans le cas de l'algorithme K-moyennes. Les étapes de cet algorithme sont les suivantes :

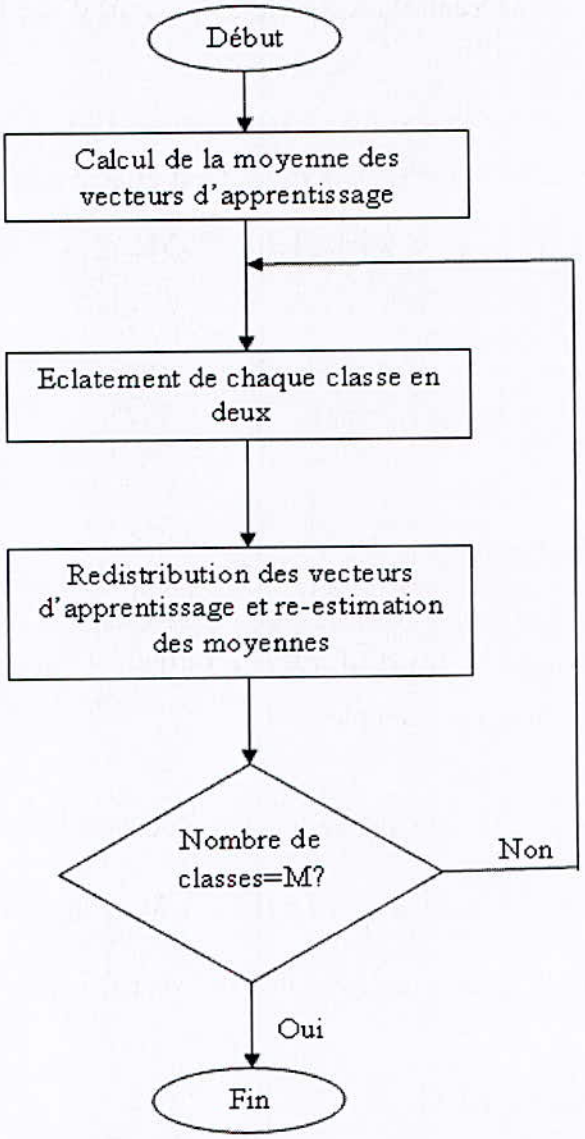


FIG. 3.3 L'algorithme LBG

1. Initialisation

Les vecteurs d'apprentissage $\{x_1, x_2, \ldots\ldots, x_T\}$ sont supposés appartenir à une seule classe dont le centroïde est la moyenne de cette classe, et est donné par : $\mu_1 = \frac{1}{T}\sum_{t=1}^{T} x_t$.

2. Eclatement du dictionnaire

Chaque sous-ensemble du dictionnaire va être éclaté. Soit ε un vecteur de petite amplitude. Le nombre de classes est doublé par éclatement du vecteur moyen μ_i en deux vecteurs : $\mu_i + \varepsilon$ et $\mu_i - \varepsilon$.

3. Optimisation du dictionnaire

Après éclatement du dictionnaire et affectation des vecteurs à l'ensemble des classes, les centroïdes sont recalculés.

4. Teste d'arrêt

Tant que $M < 2^p$ le dictionnaire est à nouveau éclaté et optimisé on réitérant les étapes 2 et 3. Pour le calcul des poids et matrices de covariance du modèle, on utilise la même procédure développée dans le cas de la K-moyennes. La figure 3.3 illustre l'organigramme de l'algorithme LBG.

3.2.3.4 Apprentissage par Maximum de vraisemblance

Le but de la méthode du Maximum de Vraisemblance est de déterminer les paramètres du modèle qui maximisent la vraisemblance des données d'apprentissage.

Pour une séquence de N vecteurs d'apprentissage $X = \{x_1, x_2, ..., x_N\}$, la vraisemblance du modèle GMM est :

$$p(X/\lambda) = \prod_{n=1}^{N} p(x_n/\lambda) = \prod_{n=1}^{N} \sum_{m=1}^{M} p(x_n/\pi_m, \mu_m, \Sigma_m) \quad (3.4)$$

L'apprentissage, dans ce cas, se décompose en deux étapes :

- Une étape d'initialisation qui permet l'obtention des valeurs approximatives des paramètres du modèle par l'algorithme K-moyennes.
- Une étape d'optimisation des valeurs de ces paramètres par un algorithme de type EM (Expectation-Maximisation).

Algorithme EM (Expectation-Maximisation)

L'algorithme EM fait intervenir des variables latentes que l'on ne peut observer directement. Dans notre cas, chaque vecteur x_j est défini non seulement par les D paramètres acoustiques mais aussi par le sous-ensemble S_i (défini par un centroïde) auquel il se rattache. Dans le cas des algorithmes K-moyennes et LBG, on a vu que chaque vecteur x se rattache réellement à un sous-ensemble. Dans le cas de l'algorithme EM ce ne sera plus le cas. Celui-ci va maximiser la vraisemblance de façon itérative, mais le vecteur x

sera maintenant rattaché aux M sous-ensembles S_i avec une probabilité particulière, sans que l'on puisse déterminer à quel sous-ensemble S_i il appartient. C'est ce paramètre que l'on qualifie de donnée cachée ou latente.

La maximisation de la fonction de vraisemblance 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 observée et les variables cachées) sur l'ensemble complet des variables d'apprentissage.

$$Q(\theta,\theta^{(t)}) = \sum_{m=1}^{M} \sum_{n=1}^{N} p(m / x_n, \theta^{(t)}) \log p(x_n, m / \theta) \tag{3.5}$$

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

$$\begin{aligned} Q(\theta,\theta^{(t)}) = & \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\log \pi_m - \frac{D}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma_m| \right] \\ & - \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\frac{1}{2} (x_n - \mu_m)' \Sigma_m^{-1} (x_n - \mu_m) \right] \end{aligned} \tag{3.6}$$

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

$$\gamma_{n,m}^{(t)} = \frac{\pi_m^{(t)} \, p(x_n / \mu_m^{(t)}, \Sigma_m^{(t)})}{\sum_{k=1}^{M} \pi_k^{(t)} \, p(x_n / \mu_k^{(t)}, \Sigma_k^{(t)})} \tag{3.7}$$

La ré-estimation des paramètres $(\pi_m^{(t+1)}, \mu_m^{(t+1)}, \Sigma_m^{(t+1)})$ à partir des paramètres estimés à l'itération t constitue la deuxième étape de l'algorithme EM.

Les formules de calcul des paramètres $(\pi_m^{(t+1)}, \mu_m^{(t+1)}, \Sigma_m^{(t+1)})$ sont données par :

$$\pi_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \tag{3.8}$$

$$\mu_m^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \, x_n}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{3.9}$$

$$\Sigma_m^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \, (x_n - \mu_m^{(t)})(x_n - \mu_m^{(t)})'}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{3.10}$$

3.2.4 Décision d'un système d'identification

Nous présentons dans cette partie la phase de décision d'un système d'identification du locuteur par GMM.

Soit un groupe de ℓ locuteurs, représentés par les modèles GMM : $\lambda_1, \lambda_2, \ldots, \lambda_\ell$. L'objectif de la phase de décision consiste à trouver, à partir d'une séquence observée X, le modèle qui a la probabilité à posteriori maximale, c'est-à-dire :

$$\hat{s} = arg \max_{1 \le s \le \ell} p(\lambda_s / X) \tag{3.11}$$

Ce qui donne, d'après la loi de bayes :

$$\hat{s} = arg \max_{1 \le s \le \ell} \frac{p(X/\lambda_s)}{p(X)} \, p(\lambda_s) \tag{3.12}$$

En supposant tous les locuteurs équiprobables, la loi de classification devient :

$$\hat{s} = arg \max_{1 \le s \le \ell} p(X / \lambda_s) \tag{3.13}$$

En utilisant le logarithme et l'indépendance entre les observations, le système d'identification se base sur l'équation :

$$\hat{s} = arg \max_{1 \le s \le \ell} \sum_{n=1}^{N} log\, p(x_n / \lambda_s) \quad (3.14)$$

3.2.5 Mesure des performances d'un système d'identification

Les performances d'un système d'identification sont données en termes de taux d'identification correcte I_c ou incorrecte I_i.

$$I_c = \frac{\text{Nombre de segments de test correctement identifiés}}{\text{Nombre total de segments de tests}} \times 100 \quad (3.15)$$

$$I_i = \frac{\text{Nombre de fausses identifications}}{\text{Nombre total de segments de tests}} \times 100 \quad (3.16)$$

Avec :

$$I_c + I_i = 100\% \quad (3.17)$$

3.3 Identification par mélanges de gaussiennes orthogonales (OGMM)

La modélisation par mélanges de gaussiennes est largement utilisée en reconnaissance du locuteur (identification et vérification).

Dans la théorie, pour chaque composante gaussienne, on doit calculer une matrice de covariance pleine. Néanmoins, dans la pratique, les matrices de covariance diagonales sont les plus utilisées, ce qui réduit considérablement la complexité des calculs, surtout en ce qui concerne l'inversion des matrices.

Généralement, les éléments des vecteurs de paramètres extraits à partir du signal parole sont corrélés. Une combinaison linéaire de fonctions gaussiennes diagonales (c'est-à-dire à matrice de covariance diagonale) est capable de modéliser cette corrélation. Néanmoins, pour fournir une bonne approximation, un grand nombre de gaussiennes doit être utilisé, ce qui entraîne une augmentation du temps de réponse du système.

Pour remédier à ce problème, nous introduisons dans cette partie une modification sur la GMM standard, pour obtenir un autre modèle appelé *OGMM* (Orthogonal GMM), et qui réduit considérablement les temps de calcul et l'espace mémoire requis.

L'idée de base est d'effectuer une analyse en composantes principales (*ACP*) que nous avons déjà brièvement décrit à la fin du chapitre 2.

Pour réduire la corrélation entre les coefficients acoustiques, une transformation linéaire est opérée sur les vecteurs acoustiques. La matrice de transformation dans ce cas est propre à chaque locuteur, et est composée des vecteurs propres de la matrice de covariance initiale du même locuteur.

Cette étape terminée, on applique sur les vecteurs acoustiques résultants la modélisation GMM standard.

3.3.1 Les mélanges de gaussiennes orthogonales (OGMM)

Un modèle GMM est une somme pondérée de M fonctions gaussiennes multidimensionnelles.

$$P(x/\lambda) = \sum_{i=1}^{M} \pi_i b_i(x) \tag{3.18}$$

avec

$$b_i(x) = \frac{1}{(2\pi)^{(D/2)} \left|\Sigma_{x_i}\right|^{1/2}} exp\left[-\frac{1}{2}(x-\mu_{x_i})^T \Sigma_{x_i}^{-1}(x-\mu_{x_i})\right] \tag{3.19}$$

Soient Σ_x la matrice de covariance initiale du locuteur courant et Ω une matrice de transformation composée par les vecteurs propres de Σ_x.

En appliquant la transformation linéaire Ω sur les vecteurs acoustiques issus de la phase de paramétrisation :

$$y = \Omega^T x \tag{3.20}$$

on obtient une matrice de covariance diagonale Σ_y. Σ_x et Σ_y sont reliées par l'équation suivante :

$$\Sigma_y = \Omega^T \Sigma_x \Omega \tag{3.21}$$

En remplaçant dans l'équation (3.19), on obtient :

$$b_i(y) = \frac{1}{(2\pi)^{(D/2)}\left|\Sigma_{y_i}\right|^{1/2}} exp\left[-\frac{1}{2}(y-\mu_{y_i})^T \Sigma_{y_i}^{-1}(y-\mu_{y_i})\right] \tag{3.22}$$

avec Σ_{y_i} et μ_{y_i} sont définis par :

$$\begin{cases} \Sigma_{y_i} = \Omega^T \Sigma_{x_i} \Omega \\ \mu_{y_i} = \Omega^T \mu_{x_i} \end{cases} \tag{3.23}$$

La figure 3.4 illustre le diagramme bloc de l'OGMM. Le modèle est composé de deux blocs, le premier bloc est une transformation linéaire dépendant du locuteur et qui sert à décorréler les coefficients acoustiques, et le second bloc est le modèle GMM standard avec matrices de covariance diagonales.

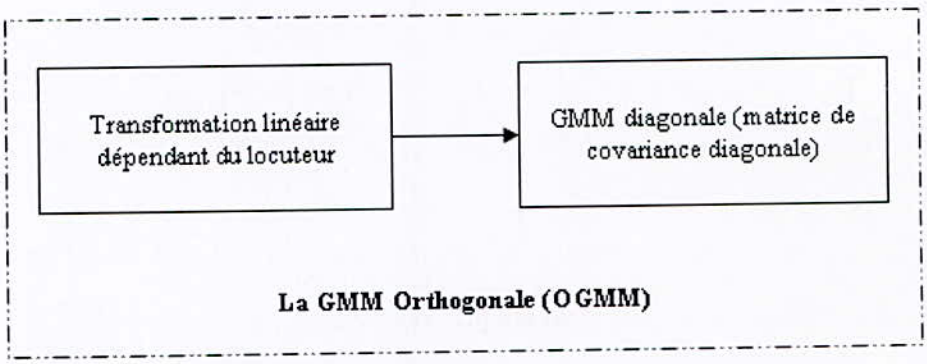


FIG. 3.4 Diagramme bloc de l'OGMM

Soient D la dimension du vecteur de paramètres et M le nombre de composantes gaussiennes dans le mélange. Dans la modélisation GMM standard, chaque modèle aura $M(2D+1)$ paramètres à calculer. L'OGMM aura besoin de $D \times D$ paramètres en plus, et ce pour la transformation linéaire.

Cette différence du nombre de paramètres n'engendre pas une différence notable dans le temps de réponse du système.

3.4 Conclusion

L'algorithme EM a l'avantage de garantir la convergence vers un maximum local, néanmoins, il présente les inconvénients suivants :

- une complexité de calcul très élevée. Le nombre d'opérations requis augmente exponentiellement avec le nombre de vecteurs d'apprentissage et linéairement avec le nombre d'itérations.
- l'algorithme EM est de nature itérative, il prend plusieurs itérations pour converger, ce qui engendre un temps d'apprentissage très long.
- la phase d'initialisation de l'algorithme EM nécessite un algorithme séparé (K-moyennes en général), ce qui augmente le coût du module d'apprentissage.

L'introduction des deux algorithmes : LBG et K-moyennes réduit considérablement les temps de calcul, et remédie ainsi aux problèmes posés par l'algorithme EM.

Pour gagner en temps de calculs, on a considéré dans la modélisation GMM des matrices de covariance diagonales. Néanmoins, cette approximation ne permet pas de prendre en compte la corrélation qui existe entre les coefficients acoustiques, impliquant ainsi une diminution des performances d'identification.

L'orthogonalisation de l'espace acoustique, permet de réduire la corrélation entre les coefficients acoustiques. Considérer des matrices de covariance diagonales, dans ce cas, serait plus justifié.

Chapitre 4

Evaluations expérimentales

Ce chapitre est décomposé en deux grandes parties, la première partie décrit le contexte expérimental de toutes les expériences effectuées le long de cette étude, et la deuxième partie expose et commente les différents résultats obtenus et donne quelques conclusions.

4.1 Contexte expérimental

Cette partie présente le contexte expérimental des évaluations des deux techniques d'identification du locuteurs en mode indépendant du texte : GMM et OGMM. En premier lieu, nous décrivons la base de données utilisée. Ensuite, nous rappelons l'analyse acoustique appliquée, les algorithmes d'apprentissage utilisés ainsi que le protocole d'évaluation utilisé. Enfin, nous décrivons l'interface graphique réalisée.

Cette partie décrit les conditions expérimentales de toutes les évaluations d'identification tant par GMM que par OGMM.

4.1.1 Description de la base de données utilisée

Dans le cadre de ce projet de fin d'études, on a utilisé une base de données composée de 48 locuteurs (32 hommes et 16 femmes) extraite exclusivement de la base de données TIMIT.

Pour chaque locuteur, on dispose de 10 phrases, chacune de 3 secondes en moyenne. On a concaténé 7 phrases pour l'apprentissage et 3 phrases pour le test. Les fichiers d'apprentissage sont étiquetés de « train_locuteur01 » à « train_locuteur48 » et les fichiers de test de « test_locuteur01 » à « test_locuteur48 ».

4.1.2 Description de la base de données TIMIT

La base de données TIMIT est une base de données acoustiques et phonétiques dédiée à la reconnaissance automatique de la parole, ainsi qu'au développement et à l'évaluation des systèmes de reconnaissance automatique de la parole. Elle contient les enregistrements de 630 locuteurs américains, prononçant chacun 10 phrases. Le texte est lu dans de bonnes conditions d'enregistrement et les données sont échantillonnées avec 16 KHz sur 16 bits.

Nombre de locuteurs	630
Nombre de session par locuteur	1
Intervalle entre sessions	Pas d'intervalle
Type de la parole	Lecture de phrases
Type de microphones utilisés	Large bande
Canal	Large bande / signal propre

TAB. 4.1 Description de la base de données TIMIT

4.1.3 Analyse acoustique et paramétrisation du signal vocal

L'analyse de la parole consiste à extraire l'information pertinente et à réduire au maximum la redondance.

On s'intéresse essentiellement à l'information relative à l'identité du locuteur, et pour cela on a choisi d'utiliser les coefficients MFCC (Mel Frequency Cepstral Coefficients) qui permettent une parfaite déconvolution de la contribution du conduit vocal et celle de la source d'excitation.

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 des 16 premiers coefficients MFCC.

La figure 4.1 illustre les étapes suivies afin d'extraire les coefficients MFCC.

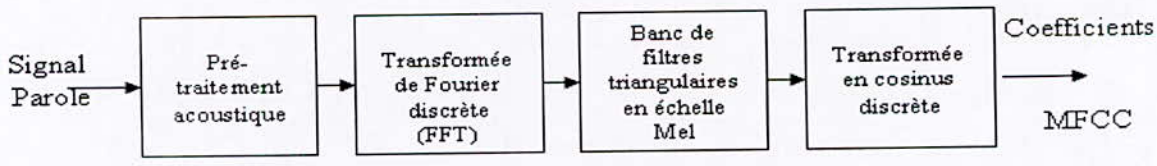


FIG. 4.1 Extraction des coefficients MFCC

La phase de pré-traitement acoustique contient deux étapes :

1. L'étape de pré-accentuation acoustique qui consiste à filtrer le signal vocal par un filtre passe haut de transmitance $H(z) = 1 - 0.95\,z^{-1}$.

2. L'étape de fenêtrage qui consiste à multiplier le signal vocal par une fenêtre de pondération glissante. Dans notre travail, on a utilisé une fenêtre de Hamming de durée de 32 *ms* avec déplacement de 16 *ms*.

La figure 4.2 illustre une fenêtre de pondération de Hamming sur 512 échantillons, et qui est définie par :

$$w(n) = 0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right] \quad \text{et} \quad 0 \le n \le N-1$$

N : nombre d'échantillons dans la fenêtre d'analyse.

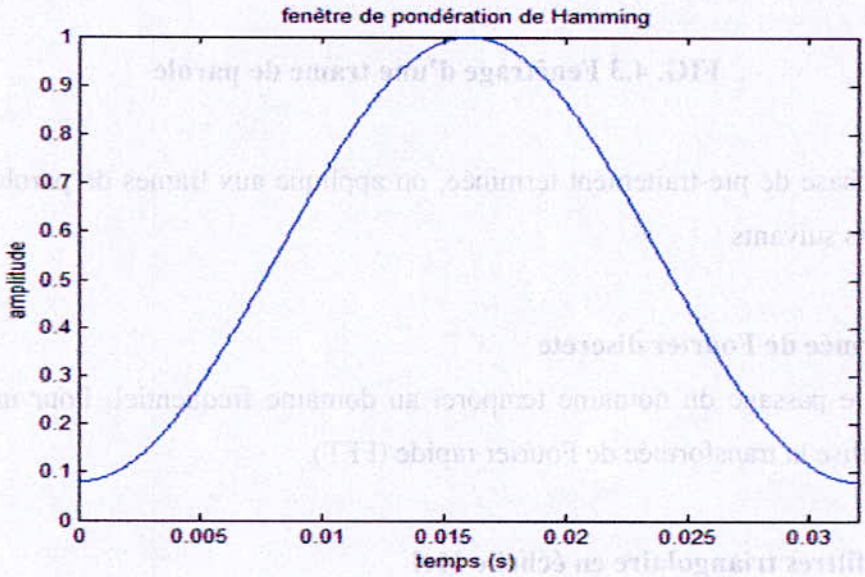


FIG. 4.2 Fenêtre de pondération de Hamming

La figure 4.3 illustre les effets du fenêtrage sur une trame de parole de 32 *ms*.

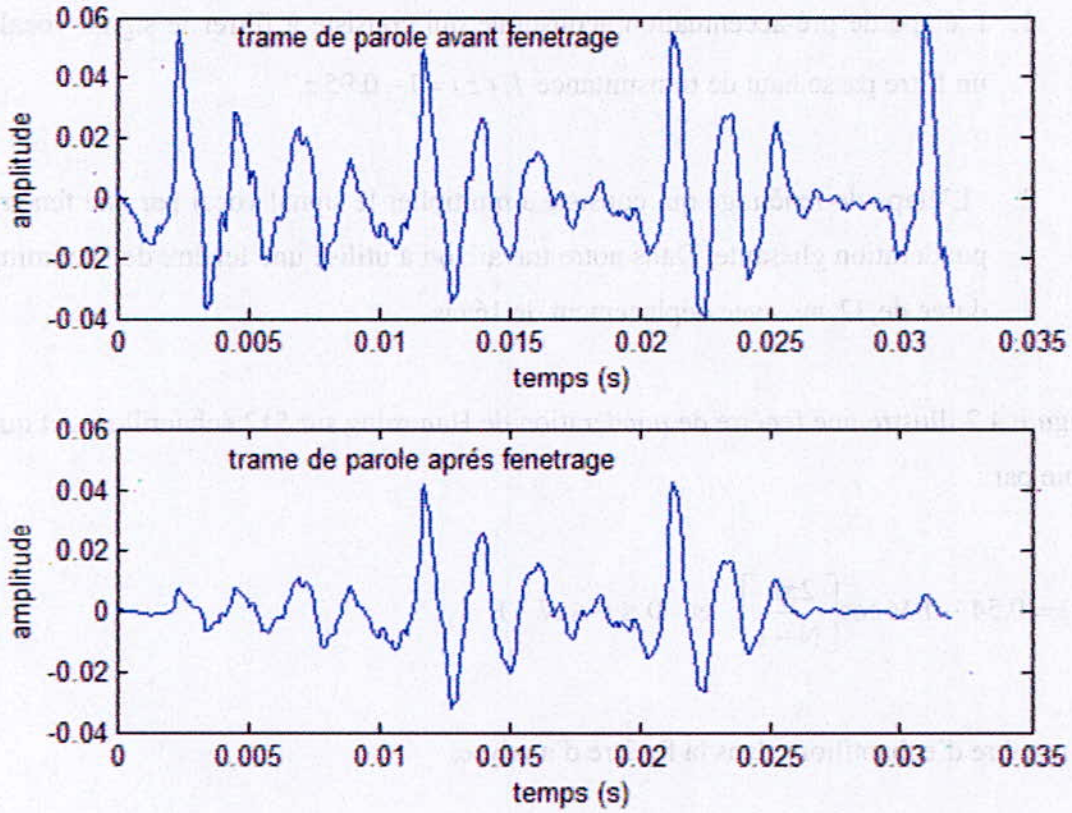


FIG. 4.3 Fenêtrage d'une trame de parole

Une fois la phase de pré-traitement terminée, on applique aux trames de parole résultantes les traitements suivants :

a. Transformée de Fourier discrète

Elle permet le passage du domaine temporel au domaine fréquentiel. Pour un traitement rapide, on utilise la transformée de Fourier rapide (FFT).

b. Banc de filtres triangulaire en échelle Mel

Le spectre du signal est filtré par un banc de filtres triangulaires, dont les bandes passantes sont de même largeur sur une échelle perceptive de type Mel. Chaque filtre opère sur une bande de fréquence bien déterminée.

c. Transformée en cosinus discrète

Les premiers coefficients cepstraux c_k sont calculés directement à partir du logarithme des énergies E_i à la sortie d'un banc de M filtres par la transformée en cosinus discrète qui permet l'obtention de coefficients fortement décorrélés et qui est définie par :

$$c_k = \sum_{i=1}^{M} log\, E_i \, cos\left[\frac{\pi\, k}{M}\left(i - \frac{1}{2}\right)\right]$$

4.1.4 Détection et élimination de silence

Avant d'aborder l'étape d'analyse acoustique, on a tout d'abord éliminé les périodes de silence. Pour cela, on a effectué une étude statistique sur la base de données utilisée, et à partir de laquelle on a déterminé un seuil d'énergie. Toute trame de niveau énergétique inférieur au seuil prédéterminé sera éliminée. La figure 4.4 illustre une trame de parole avant et après élimination de silence.

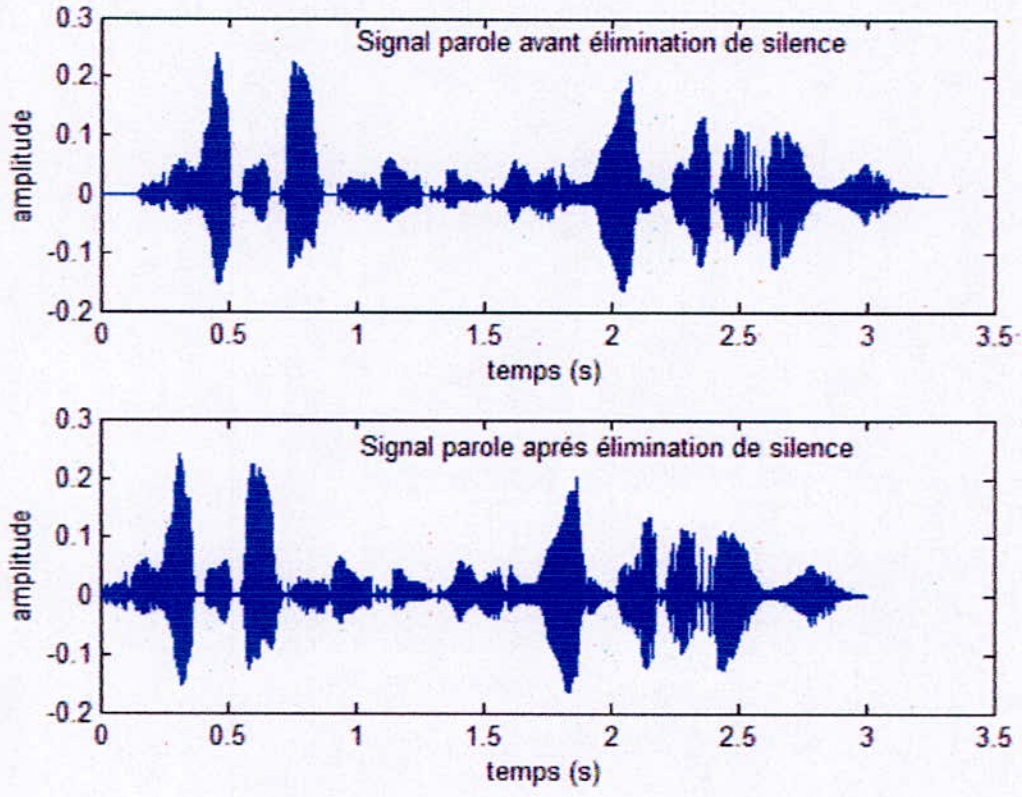


FIG. 4.4 Elimination de silence

4.1.5 Filtrage dans la bande téléphonique et ré-échantillonnage

La base de données TIMIT est échantillonnée à 16 KHz et les données sont sur 16 bits, elle est dédiée pour le développement et l'évaluation des systèmes de reconnaissance automatique de la parole.

L'évaluation d'un système ou d'une technique de reconnaissance du locuteur sur une telle base est peu objective. Pour cette raison, on a introduit une dégradation sur les données à utiliser afin d'obtenir une qualité réseau téléphonique commuté (RTC).

Tout d'abord, on a filtré par un filtre passe bande dans la bande 300-3400 Hz. Ensuite, on ré-échantillonne à 8 KHz. Enfin, on ajoute un bruit blanc gaussien avec un rapport signal sur bruit de 50 dB.

4.1.6 Apprentissage des modèles

Dans ce contexte expérimental, on va utiliser trois algorithmes d'apprentissage. En premier lieu, on fait l'apprentissage par maximum de vraisemblance en utilisant l'algorithme EM. Dans un second lieu, on va adapter les deux algorithmes de quantification vectorielle : LBG et K-moyennes pour l'apprentissage des modèles GMM et on compare ensuite leurs performances avec celles obtenues par l'algorithme EM.

4.1.7 Protocole d'évaluation

Nous allons évaluer les performances des deux approches GMM et OGMM sur un ensemble de 48 locuteurs (ensemble fermé). Il s'agit d'identifier un locuteur parmi les 48 locuteurs et de calculer le taux d'identification correcte défini par :

$$I_c = \frac{\text{Nombre de segments de test correctement identifiés}}{\text{Nombre total de segments de test}} \times 100$$

Pour chaque locuteur de l'ensemble, nous disposant de 50 segments de test (les segments de test sont sur 3 secondes avec décalage de 32 ms). Pour l'ensemble des locuteurs, nous effectuons un test par segment soit 2400 tests.

4.1.8 Langage utilisé

On a utilisé **MATLAB** version **6.5** qui possède des boites à outils spécialisées. L'ensemble des fonctions de ces boites à outils facilitent beaucoup la simulation.

Dans ce travail, on a utilisé principalement deux boites à outils, la première « Signal Processing Toolbox » orientée traitement du signal et la seconde « Voicebox Toolbox » orientée traitement de la parole.

4.1.9 Description de l'interface graphique

Pour mieux présenter notre travail et rendre facile l'utilisation du système d'identification, nous avons réalisé une interface graphique sous MATLAB 6.5. Comme le montre la figure 4.5, cette interface comprend principalement deux volets : un volet pour l'introduction des différents paramètres et un autre volet pour l'affichage des résultats.
Dans le premier volet, on choisit les paramètres suivants : le nombre de locuteurs, l'ordre du modèle, la fréquence d'échantillonnage, le rapport signal sur bruit, le locuteur de test, l'approche de modélisation ainsi que l'algorithme d'apprentissage désirés. Dans le second volet, le système affiche les durées d'apprentissage et de test ainsi que le locuteur le plus vraisemblable, c'est-à-dire le locuteur identifié.

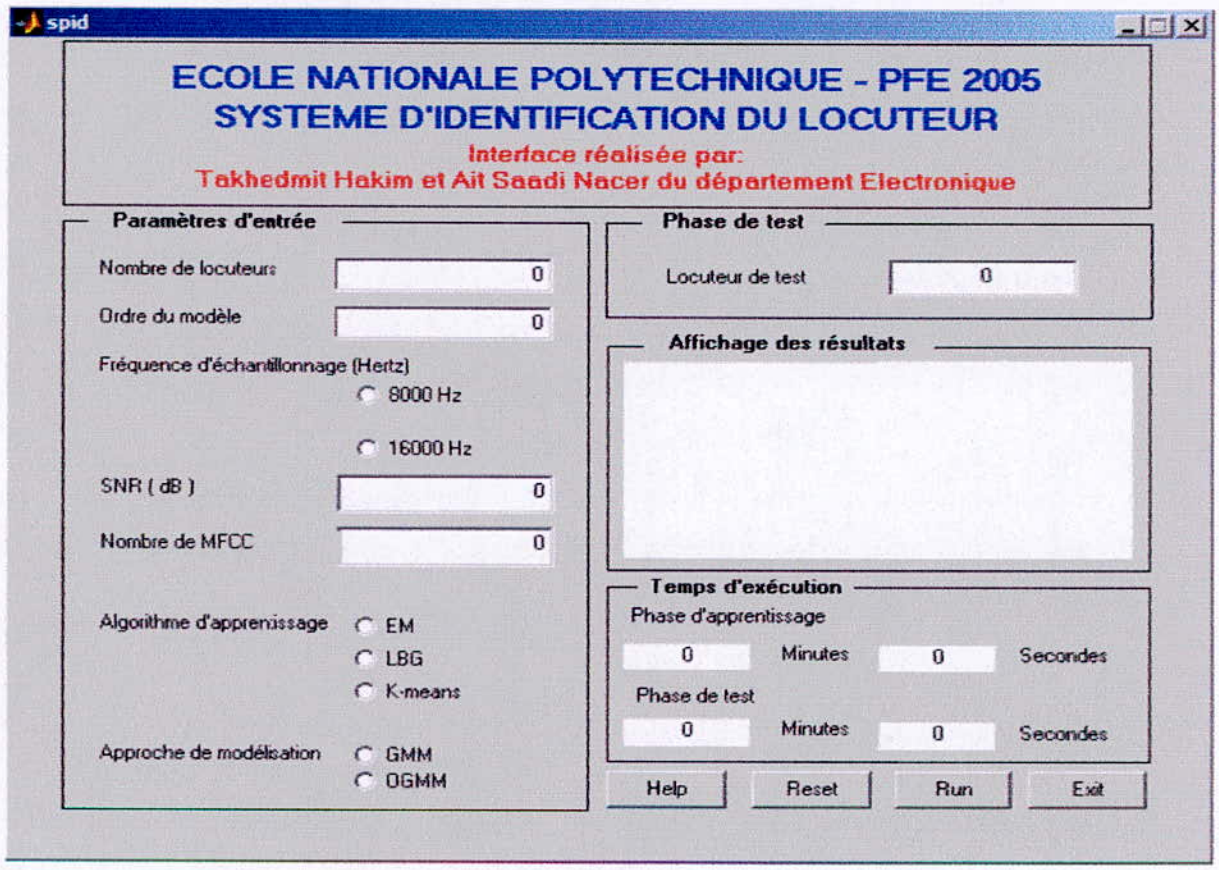


FIG. 4.5 Interface graphique du système d'identification

4.2 Evaluations expérimentales

En premier lieu, nous présentons et commentons les résultats expérimentaux obtenus par les deux techniques de modélisation GMM et OGMM. Ensuite, nous comparons et expliquons les résultas obtenus. Enfin, nous donnons quelques conclusions.

Pour la technique GMM, nous étudions l'influence des paramètres suivants sur le taux d'identification :

a. Qualité des données d'apprentissage et de test

On commence avec une fréquence d'échantillonnage de 16 KHz, et ensuite on essaie de travailler dans la bande téléphonique avec une fréquence d'échantillonnage de 8 KHz et un rapport signal sur bruit variant de 5 dB à 100 dB.

b. La dimension du vecteur de paramètres MFCC

Pour voir l'apport de la dimension du vecteur acoustique sur le taux d'identification, on va varier le nombre de coefficients MFCC de 5 à 40.

c. L'ordre du modèle

On varie l'ordre du modèle ou le nombre de composantes gaussiennes de 1, qui correspond au cas mono-gaussienne jusqu'à 64 gaussiennes.

d. Le nombre de locuteurs

Pour voir l'influence de nombre de locuteurs sur le taux d'identification correcte, des expériences ont été menées sur 16, 32 puis 48 locuteurs de la base de données.

e. L'algorithme d'apprentissage

On va évaluer et comparer les résultats obtenus avec les trois algorithmes d'apprentissage : EM, LBG et K-moyennes.

f. La quantité des données de test

Dans la pratique, on a généralement peu de données de test. Pour évaluer les performances de la GMM dans de telles conditions et voir l'influence de la quantité de données de test, on a testé le système sur 1 seconde puis sur 3 secondes.

Pour la technique OGMM, nous étudions l'influence du rapport signal sur bruit ainsi que l'ordre du modèle sur le taux d'identification correcte.

L'étude de l'influence des différents paramètres cités sur les performances des deux techniques GMM et OGMM terminée, une étude comparative entre la GMM et l'OGMM s'impose.

4.2.1 Les mélanges de gaussiennes standards (GMM)

4.2.1.1 La fréquence d'échantillonnage : 16 KHz

4.2.1.1.1 Etude de l'influence de l'ordre du modèle

Nombre de classes		1	2	4	8	16	32	64
Taux d'identification correcte (%)	16 locuteurs	93.5	96.88	99.13	99.50	100	100	100
	32 locuteurs	94.13	96	97.5	97	100	100	100
	48 locuteurs	92.04	95.67	99.21	99.42	100	100	98.58

TAB. 4.2 GMM - 16 KHz : Influence de l'ordre du modèle

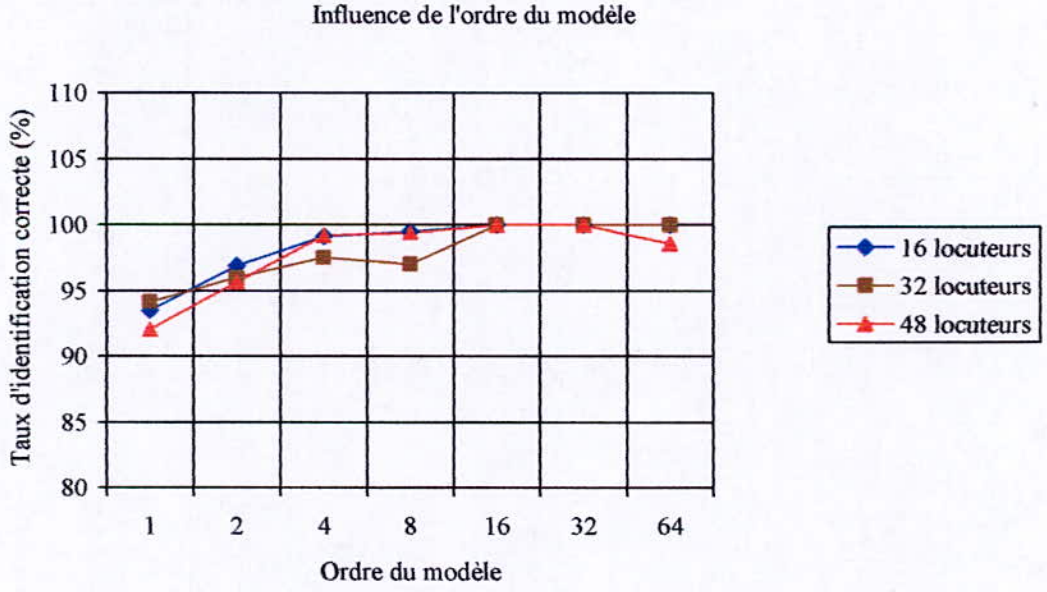


FIG. 4.6 GMM - 16 KHz : Influence de l'ordre du modèle

Commentaires et conclusions

- Comme l'illustre la figure 4.6, l'ordre du modèle apporte une amélioration significative au taux d'identification correcte. Néanmoins, au delà de 16 gaussiennes où le taux d'identification atteint les 100 %, on remarque qu'un régime permanent s'établit.
- Pour ce qui concerne le nombre de locuteurs, on remarque que le nombre de segments correctement identifiés diminue avec l'augmentation du nombre de locuteurs que le système doit identifier.

- L'augmentation de l'ordre du modèle permet d'affiner la séparation des classes acoustiques, ce qui se traduit par un accroissement du taux d'identification.
- On constate qu'avec une fréquence d'échantillonnage de 16 KHz, 16 composantes gaussiennes sont amplement suffisantes pour modéliser un locuteur.
- Lorsque le nombre de locuteurs que le système doit identifier augmente, les similitudes entres locuteurs augmentent, ce qui accroît la probabilité d'une fausse identification. Cet accroissement se traduit par la dégradation des performances.

4.2.1.1.2 Etude de l'influence de la dimension du vecteur acoustique

Dans cette partie, on a choisi de travailler sur 32 locuteurs qui, d'après les résultats obtenus précédemment, constitue le cas le plus défavorable.

Nombre de coefficients MFCC	5	10	15	20	25	30	35	40
Taux d'identification correcte (%)	91.44	96.88	98.63	98.31	100	100	98.88	99.56

TAB. 4.3 GMM - 16 KHz : Influence du la dimension du vecteur acoustique

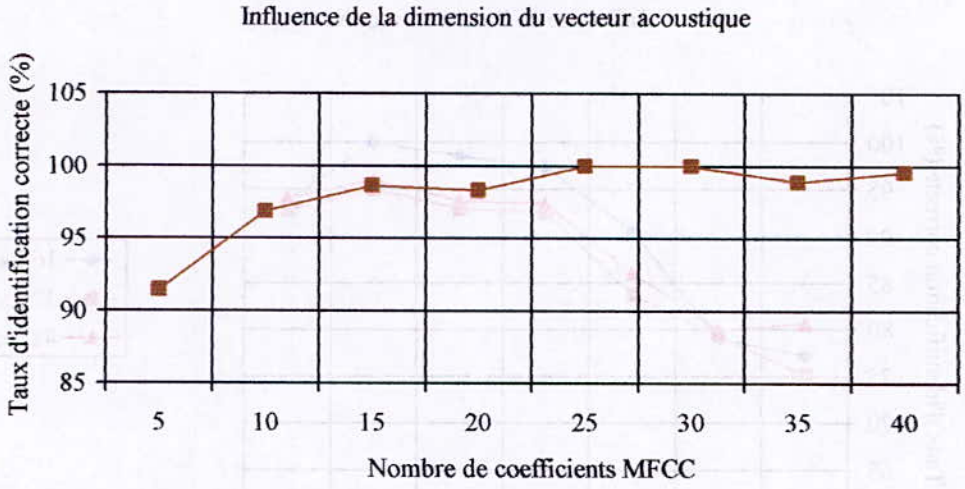


FIG. 4.7 GMM - 16 KHz : Influence de la dimension du vecteur acoustique

Commentaires et conclusions

- La figure 4.7 illustre que le taux d'identification correcte augmente avec le nombre de coefficients MFCC utilisé, avec un maximum entre 25 et 30 coefficients. La pente de la courbe est beaucoup plus importante entre 5 et 25 coefficients. Au delà de 25 coefficients MFCC, on atteint pratiquement le régime permanent.
- On constate que la quasi-totalité de l'énergie du signal parole utilisé est contenue dans les 25 premiers MFCC, et les coefficients MFCC d'ordre supérieurs n'apportent pratiquement pas un plus d'information sur l'identité du locuteur.
- Pour une fréquence d'échantillonnage de 16 *KHz* , on constate qu'il faut utiliser entre 25 et 30 coefficients MFCC pour avoir de bons taux d'identification.

4.2.1.2 La fréquence d'échantillonnage : 8 KHz

4.2.1.2.1 Etude de l'influence de l'ordre du modèle

a. Algorithme EM

Ordre du modèle		1	2	4	8	16	32	64
Taux d'identification correcte (%)	**16 locuteurs**	77.25	79.13	90.50	97.38	98.63	100	93.75
	32 locuteurs	75.13	79.38	83.75	92.56	92.94	95	92.75
	48 locuteurs	80.71	79.96	86.13	93.46	93.80	96.04	94.04

TAB. 4.4 GMM - 8 KHz - EM : Influence de l'ordre du modèle

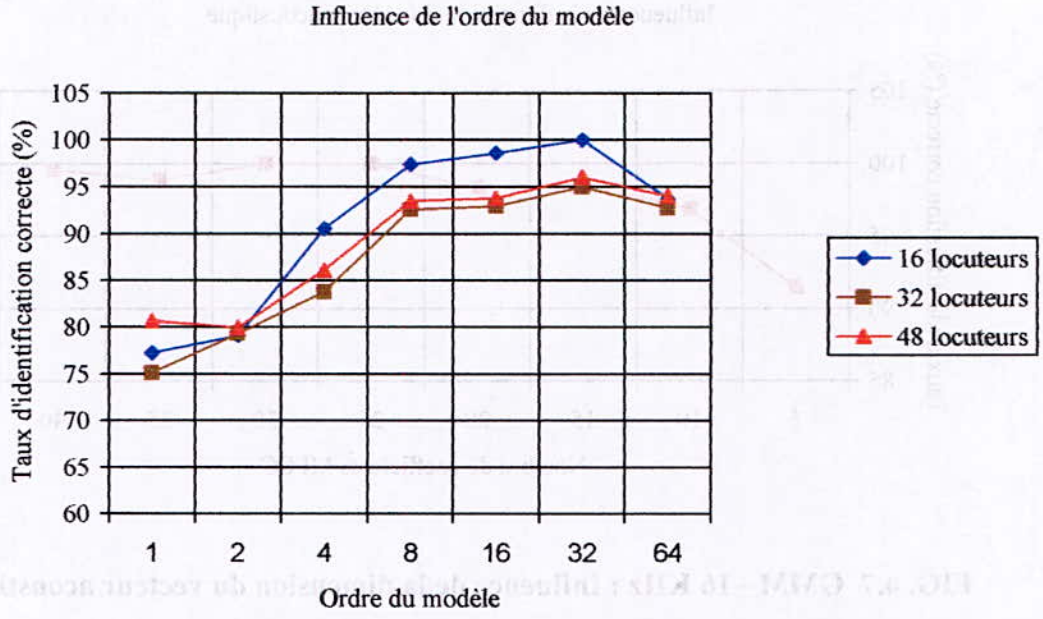


FIG. 4.8 GMM - 8 KHz - EM : Influence de l'ordre du modèle

Commentaires

Sur les courbes de la figure 4.8, on remarque que l'ordre du modèle améliore le taux d'identification correcte jusqu'à 32 gaussiennes, au delà de lesquelles le pourcentage d'identification diminue. On remarque aussi que, pour 32 et 48 locuteurs, le taux d'identification ne dépasse pas les 96 %.

b. L'algorithme LBG

Ordre du modèle		1	2	4	8	16	32	64
Taux d'identification correcte (%)	**16 locuteurs**	78.63	81.5	88.13	86.38	97.75	100	99
	32 locuteurs	65.94	74.88	86.25	89.81	97	95.88	98.13
	48 locuteurs	71.42	77.63	87.46	89.75	94.21	93.29	95.79

TAB. 4.5 GMM - 8 KHz - LBG : Influence de l'ordre du modèle

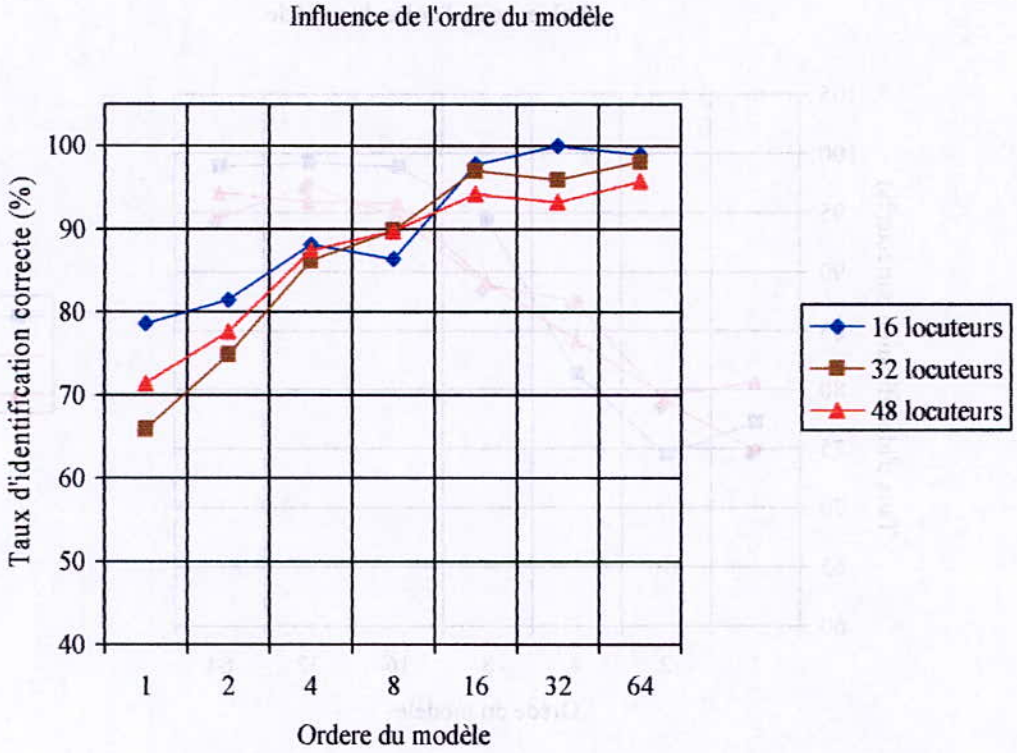


FIG. 4.9 GMM - 8 KHz - LBG : Influence de l'ordre du modèle

Commentaires

Sur les courbes de la figure 4.9, on remarque que l'ordre du modèle améliore le taux d'identification correcte. Pour 16 locuteurs, le maximum est atteint pour 32 gaussiennes. Cependant, pour 32 et 48 locuteurs, il faut aller jusqu'à 64 gaussiennes pour atteindre le maximum.

c. L'algorithme K-moyennes

Ordre du modèle		1	2	4	8	16	32	64
Taux d'identification correcte (%)	**16 locuteurs**	77.25	74.5	81.38	94.5	98.88	99.25	98.88
	32 locuteurs	74.94	78.94	87.50	88.88	94.63	97.38	94.56
	48 locuteurs	80.71	79.54	84.17	89.17	95.79	95.92	96.71

TAB. 4.6 GMM - 8 KHz - K-moyennes : Influence de l'ordre du modèle

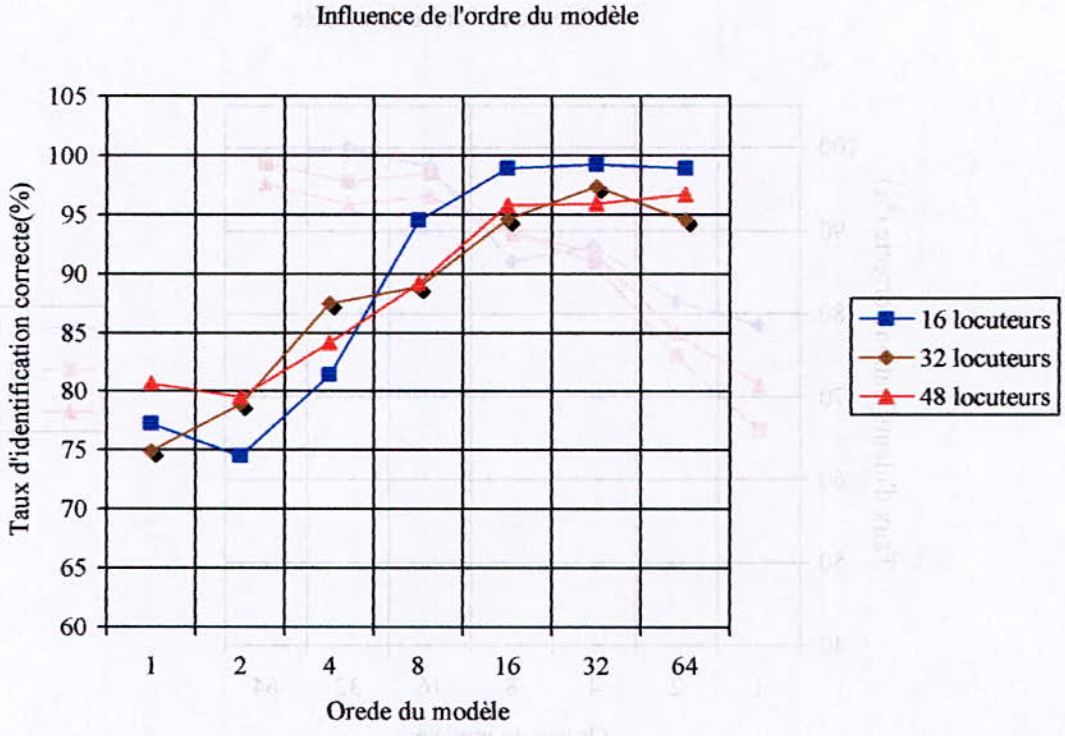


FIG. 4.10 GMM - 8 KHz – K-moyennes : Influence de l'ordre du modèle

Commentaires

Sur les courbes de la figure 4.10, on remarque que l'ordre du modèle améliore le taux d'identification correcte. Pour 16 et 32 locuteurs, le maximum est atteint pour 32 gaussiennes. Cependant, pour 48 locuteurs, il faut aller jusqu'à 64 gaussiennes pour atteindre le maximum.

Conclusions

- La diminution du taux d'identification correcte au delà d'un certain nombre de composantes gaussiennes s'explique par le fait qu'on dispose de peu de données d'apprentissage.
- On constate d'après les résultats trouvés, que lorsque on dégrade les données de test et d'apprentissage, il faut augmenter considérablement l'ordre du modèle pour avoir de bonnes performances.

Comparaisons

En comparant les taux d'identification obtenus par les trois algorithmes d'apprentissage utilisés, on a trouvé que les deux algorithmes de quantification vectorielle : LBG et K-moyennes donnent respectivement 98.74 % et 99.53 % de la précisions de l'algorithme EM. Cependant, ces deux algorithmes permettent de diminuer de façon considérable les temps de calculs (phase d'apprentissage) et la complexité relativement par rapport à l'algorithme EM, ce qui les rend le choix idéal pour les applications grand public qui ne nécessitent pas un niveau de sécurité élevé.

Dans tout ce qui suit, on va travailler sur un ensemble de 32 locuteurs et l'apprentissage des modèle va se faire avec l'algorithme EM qui, d'après les résultats obtenus précédemment, constitue le cas le plus défavorable.

4.2.1.2.2 Etude de l'influence de la dimension du vecteur acoustique

Nombre de coefficients MFCC	5	10	15	20	25	30	35	40
Taux d'identification correcte (%)	75.69	84.75	91.32	97.75	94.75	95.94	97.81	94.06

TAB. 4.7 GMM - 8 KHz : Influence du la dimension du vecteur acoustique

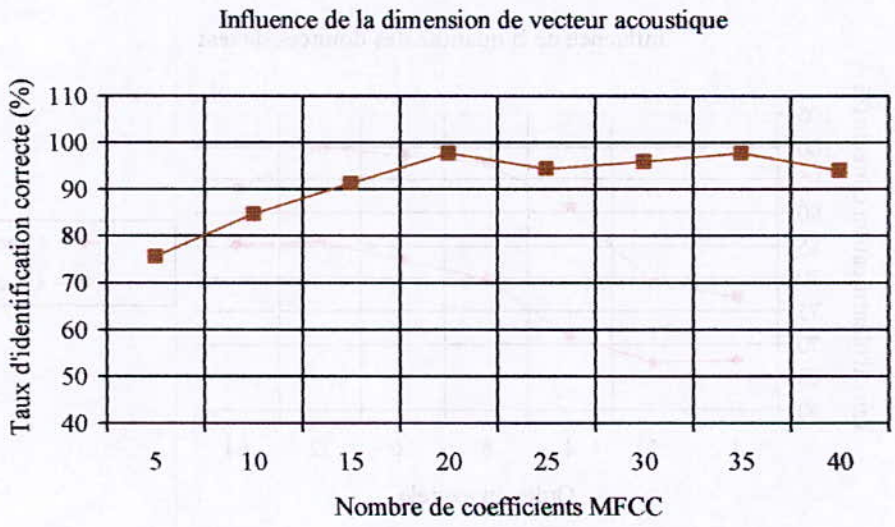


FIG. 4.11 GMM - 8 KHz : Influence de la dimension du vecteur acoustique

Commentaires et conclusions

- Sur la courbe de la figure 4.11, on peut remarquer que l'accroissement du nombre de coefficients apporte une amélioration sur le taux d'identification correcte. A partir du maximum atteint pour 35 coefficients MFCC, on remarque une certaine diminution du taux d'identification.
- On constate que la quasi-totalité de l'énergie du signal parole utilisé est contenue dans les 35 premiers MFCC, et les coefficients MFCC d'ordre supérieurs n'apportent pratiquement pas d'information sur l'identité du locuteur.
- Pour une fréquence d'échantillonnage de 8 KHz et un rapport signal sur bruit de 50 dB, on constate qu'il faut utiliser 35 coefficients MFCC pour avoir de bons taux d'identification.

4.2.1.2.3 Etude de l'influence de la quantité des données de test

Ordre du modèle		1	2	4	8	16	32	64
Taux d'identification correcte (%)	3 s de test	77.25	79.13	90.50	97.38	98.63	100	93.75
	1 s de test	67.5	67.25	71	79.75	82.75	85.38	85.25

TAB. 4.8 GMM - 8 KHz : Influence de la quantité des données de test

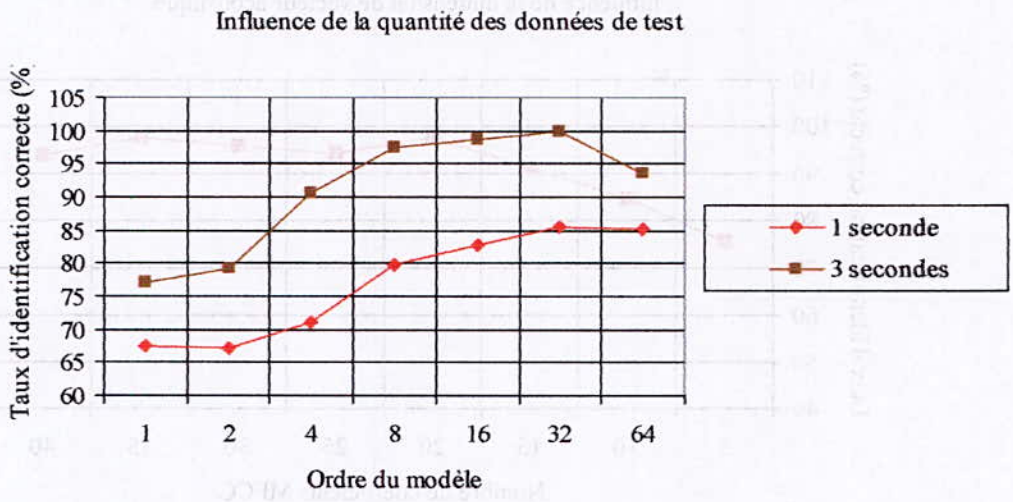


FIG. 4.12 GMM - 8 KHz : Influence de la quantité des données de test

Commentaire et conclusions

- La figure 4.12 montre que le taux d'identification augmente avec la quantité des données de test. On remarque aussi que les deux courbes ont pratiquement la même allure.
- On constate que la quantité des données de test est un facteur déterminant des performances d'un système d'identification.
- On constate aussi que le taux d'identification augmente de façon linéaire avec la quantité des données de test.

4.2.1.2.4 Etude de l'influence du rapport signal sur bruit

Rapport SNR (dB)	10	20	25	30	35	40	45	50	55
Taux d'identification correcte (%)	4.56	15.94	48.06	63.06	84.06	88.44	93	92.94	94.38

Rapport SNR (dB)	60	65	70	75	80	85	90	95	100
Taux d'identification correcte (%)	93.75	98.56	96.69	95.19	95.88	96.69	96.50	95.38	97.19

TAB. 4.9 GMM - 8 KHz : Influence du rapport signal sur bruit

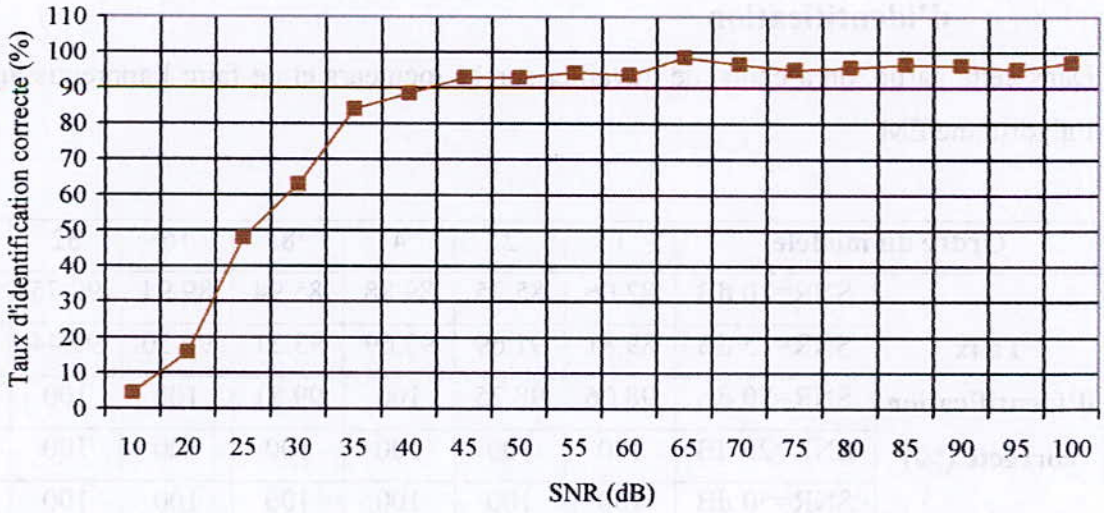


FIG. 4.13 GMM - 8 KHz : Influence du rapport signal sur bruit

Commentaires et conclusions

- La figure 4.13 montre que le taux d'identification augmente en améliorant la qualité des données, c'est-à-dire en augmentant le rapport signal sur bruit. La pente de la courbe est plus importante entre 10 et 45 dB, et au delà de 45 dB le régime permanent s'installe.

- D'après les résultats obtenus, on constate que la technique de modélisation par mélanges de gaussiennes standards (GMM) n'est pas adaptée pour les milieux fortement bruités.

Conclusions sur la GMM

D'après l'ensemble des expériences que nous avons effectué, on constate que :

- Lorsqu' on dégrade la qualité des données de test et d'apprentissage, il faut augmenter le nombre de coefficients MFCC ainsi que l'ordre du modèle.
- Pour augmenter le nombre de composantes gaussiennes, il faut disposer de beaucoup de données d'apprentissage.
- La technique GMM standard n'est pas robuste au bruit, elle est adaptée seulement pour les milieux faiblement bruités.

4.2.2 Les mélanges de gaussiennes orthogonale (OGMM)

4.2.2.1 Etude de l'influence de l'ordre du modèle et du SNR sur le taux d'identification

Dans cette partie, on a choisi de travailler sur 16 locuteurs et de faire l'apprentissage avec l'algorithme EM.

Ordre du modèle		1	2	4	8	16	32	64
Taux d'identification correcte (%)	SNR=10 dB	82.06	85.25	89.88	85.94	89.94	90.75	91.86
	SNR=15 dB	89.81	91.69	93.69	93.31	99.50	99.44	98.44
	SNR=20 dB	98.06	98.75	100	99.81	100	100	100
	SNR=25 dB	100	100	100	100	100	100	100
	SNR=50 dB	100	100	100	100	100	100	100

TAB. 4.10 OGMM : Influence de l'ordre de modèle et du rapport signal sur bruit

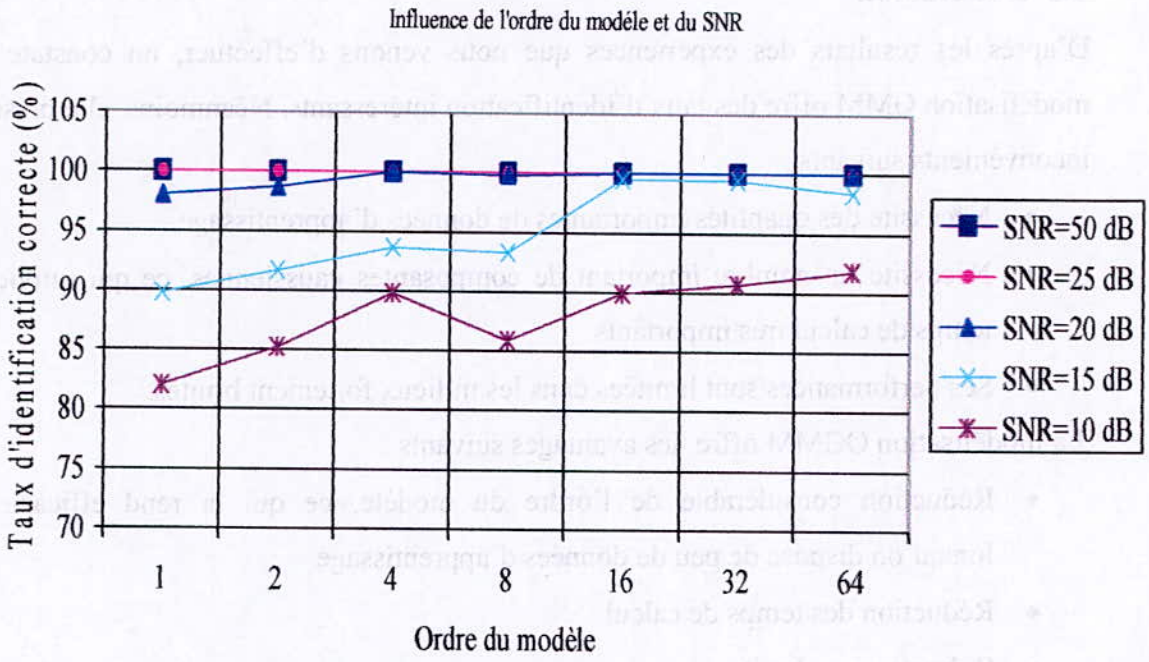


FIG. 4.14 OGMM : Influence de l'ordre du modèle et du rapport signal sur bruit

Commentaires et conclusions

- La figure 4.14 montre que l'augmentation de l'ordre du modèle ainsi que le rapport signal sur bruit améliore le taux d'identification du système. On remarque que pour un $SNR \geq 25\,dB$, on obtient des taux d'identification de 100 %.

- D'après ces expériences, on constate que la modélisation OGMM est très robuste au bruit.

- La modélisation OGMM permet réduit l'ordre du modèle et permet d'avoir de bons taux d'identification.

4.2.3 Etude comparative entre GMM et OGMM et conclusions

- En comparant les performances des deux approches GMM et OGMM, on constate que l'OGMM offre des résultats largement meilleurs que ceux obtenus par la GMM, surtout en milieu fortement bruité.
- On remarque que pour les mêmes performances, l'OGMM utilise un nombre de gaussiennes réduit, ce qui permet une réduction considérable des temps de calculs, d'autant plus que ce dernier augmente de façon exponentielle avec l'ordre du modèle.

4.3 Conclusion

D'après les résultats des expériences que nous venons d'effectuer, on constate que la modélisation GMM offre des taux d'identification intéressants. Néanmoins, elle présente les inconvénients suivants :

- Nécessite des quantités importantes de données d'apprentissage.
- Nécessite un nombre important de composantes gaussiennes, ce qui implique des temps de calcul très importants.
- Ses performances sont limitées dans les milieux fortement bruités.

La modélisation OGMM offre les avantages suivants :

- Réduction considérable de l'ordre du modèle, ce qui la rend efficace même lorsqu'on dispose de peu de données d'apprentissage.
- Réduction des temps de calcul.
- Robustesse au bruit.

Ce qui lui permet de résoudre certains problèmes de la modélisation GMM.

Conclusions et Perspectives

Dans le cadre de ce projet de fin d'études, nous avons étudié et évalué les deux techniques statistiques de modélisation du locuteur, GMM et OGMM, appliquées à l'identification du locuteur en mode indépendant du texte.

Pour ce qui concerne la modélisation par mélange de gaussiennes standards (GMM), nous avons effectué un certain nombre d'expériences où nous avons examiné l'influence d'un certain nombre de paramètres sur le taux d'identification correcte, et à partir desquelles nous avons aboutit aux conclusions suivantes :

- La qualité et la quantité des données, ainsi que la taille de la population constituent le problème principal des systèmes d'identification du locuteur.
- La modélisation GMM fournit de bonnes performances. Néanmoins, elle nécessite beaucoup de données d'apprentissage, ce qui engendre des temps de calculs assez importants.
- La modélisation GMM n'est pas très robuste au bruit.

L'introduction des deux algorithmes de quantification vectorielle : LBG et K-moyennes pour l'apprentissage des modèles GMM permet une réduction significative des temps de calculs avec une diminution minime des performances.

Pour tenter de remédier à certains problèmes posés par la GMM, nous avons introduit la notion d'orthogonalisation de l'espace acoustique sur la GMM standard, ce qui a donné naissance à la modélisation par mélange de gaussiennes orthogonales (OGMM).

Nous avons effectué un certain nombre d'expériences où nous avons examiné l'influence de l'ordre du modèle et du rapport signal sur bruit sur la robustesse de ce type de modélisation, et à partir desquelles nous avons aboutit aux conclusions suivantes :

- L'OGMM donne des performances nettement supérieures par rapport à la GMM.
- L'OGMM permet de réduire de façon remarquable le nombre de composantes gaussiennes nécessaires pour une bonne modélisation. Par conséquent, elle permet un gain considérable en temps de calculs.
- L'OGMM est beaucoup plus robuste au bruit que la GMM.

Comme on vient de le voir, l'idée de base de la modélisation OGMM est une orthogonalisation de l'espace acoustique propre à chaque locuteur. Cela permet d'approximer les matrices de covariances par des matrices diagonales.

Une perspectives très intéressante serait d'orthogonaliser non pas l'espace acoustique propre à chaque locuteur, mais les sous espaces engendrés par les classes acoustiques constituants l'espace acoustique de chaque locuteur. Cela permet d'avoir réellement des matrices de covariances diagonales, ce qui permet de donner une modélisation plus rigoureuse du nuage des vecteurs constituants l'espace acoustique.

Pour ce faire, il faut proposer des algorithmes qui intègrent un module d'orthogonalisation dans la phase d'apprentissage.

Annexe A

Résultats des Evaluations Expérimentales

Cette annexe expose les résultats intermédiaires de toutes les expériences effectuées le long de ce travail de thèse. 50 segments de test sont utilisés pour chaque locuteur, et les tableaux ci-dessous donnent le nombre de segments correctement identifiés sur l'ensemble des segments de test.

1. Modélisation par GMM

1.1 Fréquence d'échantillonnage : 16 KHz

1.1.1 Etude de l'influence de l'ordre du modèle

Algorithme d'apprentissage : EM

a. Nombre de locuteurs : 16

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	9	30	43	46	50	50	50
2	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	48	48	50	50	50	50	50
9	50	50	50	50	50	50	50
10	50	50	50	50	50	50	50
11	50	50	50	50	50	50	50
12	50	50	50	50	50	50	50
13	50	50	50	50	50	50	50
14	50	50	50	50	50	50	50
15	50	50	50	50	50	50	50
16	41	47	50	50	50	50	50
I_c **(%)**	**93.5**	**96.88**	**99.13**	**99.50**	**100**	**100**	**100**

b. Nombre de locuteurs : 32

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	45	50	50	50	50	50
2	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	49	39	50	50	50	50	50
9	41	50	50	50	50	50	50
10	50	50	50	50	50	50	50
11	50	50	50	50	50	50	50
12	50	50	50	50	50	50	50
13	50	50	50	50	50	50	50
14	41	50	50	50	50	50	50
15	50	50	50	50	50	50	50
16	41	50	50	50	50	50	50
17	50	50	50	50	50	50	50
18	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50
20	49	50	50	50	50	50	50
21	50	50	50	50	50	50	50
22	50	50	50	50	50	50	50
23	35	9	10	2	50	50	50
24	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50
26	50	43	50	50	50	50	50
27	50	50	50	50	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	50	50	50	50	50	50	50
31	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50
I_c **(%)**	**94.13**	**96**	**97.5**	**97**	**100**	**100**	**100**

c. Nombre de locuteurs : 48

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	50	50	50	50	50	50
2	50	50	43	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50
7	24	11	50	50	50	50	16
8	49	50	50	50	50	50	50
9	41	50	50	50	50	50	50
10	50	50	50	50	50	50	50
11	50	50	50	50	50	50	50

12	50	46	50	50	50	50	50
13	50	50	50	50	50	50	50
14	41	50	50	50	50	50	50
15	34	45	50	50	50	50	50
16	41	50	50	50	50	50	50
17	50	50	50	50	50	50	50
18	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50
20	49	50	50	50	50	50	50
21	50	50	50	50	50	50	50
22	50	50	50	50	50	50	50
23	35	7	38	36	50	50	50
24	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50
27	50	50	50	50	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	48	50	50	50	50	50	50
31	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50
33	25	50	50	50	50	50	50
34	50	50	50	50	50	50	50
35	50	50	50	50	50	50	50
36	22	50	50	50	50	50	50
37	50	50	50	50	50	50	50
38	50	50	50	50	50	50	50
39	50	50	50	50	50	50	50
40	50	50	50	50	50	50	50
41	50	43	50	50	50	50	50
42	50	50	50	50	50	50	50
43	50	44	50	50	50	50	50
44	50	50	50	50	50	50	50
45	50	50	50	50	50	50	50
46	50	50	50	50	50	50	50
47	50	50	50	50	50	50	50
48	50	50	50	50	50	50	50
I_C **(%)**	**92.04**	**95.67**	**99.21**	**99.42**	**100**	**100**	**98.58**

1.1.2 Etude de l'influence du nombre de coefficients MFCC

Ordre du modèle : 16

Nombre de locuteurs : 32

Algorithme d'apprentissage : EM

	Nombre de coefficients MFCC							
locuteur	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**
1	49	50	50	50	50	50	50	50
2	11	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50	50

6	50	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50	50
8	50	50	50	50	50	50	50	50
9	50	50	50	50	50	50	50	50
10	50	50	50	50	50	50	50	50
11	50	50	50	50	50	50	50	50
12	50	50	50	50	50	50	50	50
13	5	50	50	50	50	50	50	50
14	50	50	50	50	50	50	50	50
15	50	50	50	50	50	50	50	50
16	26	50	50	50	50	50	50	50
17	50	50	50	50	50	50	50	50
18	50	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50	50
20	50	50	50	50	50	50	50	50
21	50	50	50	50	50	50	50	50
22	50	50	50	50	50	50	50	50
23	29	0	28	50	50	50	50	50
24	50	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50	50
27	43	50	50	23	50	50	32	43
28	50	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50	50
30	50	50	50	50	50	50	50	50
31	50	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50	50
I_c **(%)**	**91.44**	**96.88**	**98.63**	**98.31**	**100**	**100**	**98.88**	**99.56**

1.2 Fréquence d'échantillonnage : 8 KHz

1.2.1 Etude de l'influence de l'ordre du modèle

1.2.1.1 Apprentissage par l'algorithme EM

a. Nombre de locuteurs : 16

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	23	34	40	50	50	50
2	50	50	28	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	2	7	39	50	50	50	50
6	50	50	50	50	50	50	50
7	45	50	50	50	50	50	50
8	50	50	50	50	50	50	50
9	3	0	36	50	48	50	50
10	31	42	50	50	50	50	0
11	50	50	50	50	43	50	50
12	50	50	50	50	50	50	50
13	40	13	37	50	50	50	50
14	50	48	50	50	50	50	50
15	50	50	50	50	50	50	50
16	47	50	50	39	48	50	50
I_c **(%)**	**77.25**	**79.13**	**90.50**	**97.38**	**98.63**	**100**	**93.75**

b. Nombre de locuteurs : 32

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	23	37	50	50	50	49
2	50	50	32	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	2	6	42	50	50	50	50
6	50	50	50	50	50	50	50
7	45	50	50	50	50	50	50
8	12	45	50	50	50	36	34
9	2	0	50	33	49	50	50
10	33	42	19	50	50	50	50
11	50	42	50	0	0	0	0
12	50	40	0	50	50	50	50
13	20	25	50	19	3	50	11
14	32	12	0	39	50	50	50
15	50	50	35	50	50	50	50
16	49	50	50	50	50	50	50
17	49	50	50	40	50	50	50
18	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50
20	9	50	50	50	50	50	50
21	9	27	17	50	50	50	50
22	50	46	50	50	47	50	50
23	29	18	20	50	50	34	40
24	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50
26	50	50	50	50	38	50	50
27	31	47	45	50	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	48	40	50	50	50	50	50
31	50	50	50	50	50	50	50
32	32	7	43	50	50	50	50
I_C **(%)**	**75.13**	**79.38**	**83.75**	**92.56**	**92.94**	**95**	**92.75**

c. Nombre de locuteurs : 48

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	16	43	50	50	50	50
2	50	50	22	30	50	50	50
3	50	43	50	50	50	50	50
4	50	50	50	50	50	50	50
5	0	7	32	50	50	50	50
6	50	50	50	50	50	50	50
7	41	36	50	50	50	50	50
8	12	46	50	39	50	50	50
9	0	0	50	23	49	50	50
10	30	42	5	50	50	50	50
11	50	43	50	0	0	0	0

12	50	42	4	50	50	50	50
13	21	0	50	18	7	45	49
14	32	13	0	50	50	50	50
15	50	50	50	50	50	50	50
16	41	50	50	50	40	50	49
17	32	18	49	14	47	49	41
18	50	50	0	50	50	50	50
19	50	50	50	50	50	50	50
20	5	50	50	50	50	50	50
21	10	27	49	32	50	50	50
22	50	45	49	50	50	50	41
23	30	17	50	42	43	37	41
24	50	50	22	50	50	50	50
25	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50
27	30	42	50	28	47	49	50
28	50	50	40	50	50	50	50
29	50	50	50	50	50	50	50
30	48	40	50	50	40	48	50
31	50	50	39	50	50	50	50
32	31	50	50	50	50	50	50
33	50	50	50	50	50	50	50
34	50	44	50	50	50	50	50
35	50	50	50	50	50	50	50
36	50	50	50	50	50	50	50
37	50	50	50	50	50	50	50
38	50	50	50	50	50	50	50
39	50	50	50	50	50	50	50
40	50	50	50	50	50	50	50
41	48	49	50	50	50	50	50
42	50	50	50	50	50	50	50
43	50	0	50	50	50	50	50
44	50	50	50	50	50	50	50
45	50	50	49	50	50	50	50
46	28	0	38	48	30	31	0
47	48	49	26	47	48	46	36
48	50	50	50	50	50	50	50
I_C **(%)**	**80.71**	**79.96**	**86.13**	**93.46**	**93.79**	**96.04**	**94.04**

1.2.1.2 Apprentissage par l'algorithme LBG

a. Nombre de locuteurs : 16

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	21	50	9	50	50	50
2	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	50	50	50	50	50	50	50
9	9	0	0	0	38	50	50
10	0	38	50	50	50	50	50

11	50	50	50	49	50	50	50
12	50	35	50	50	50	50	50
13	50	49	50	50	50	50	50
14	50	36	49	50	50	50	50
15	50	50	50	50	50	50	50
16	20	23	6	33	44	50	42
I_C **(%)**	**78.63**	**81.50**	**88.13**	**86.38**	**97.75**	**100**	**99**

b. Nombre de locuteurs : 32

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	0	38	48	50	50	50
2	48	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	37	44	50	50	50	50	50
6	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	0	50	50	50	50	50	50
9	3	0	0	4	50	50	50
10	0	36	50	50	50	50	48
11	50	50	50	50	39	3	50
12	50	35	50	50	50	50	50
13	50	39	50	50	50	50	50
14	0	0	0	0	15	37	36
15	50	50	50	50	50	50	50
16	20	23	7	28	50	50	42
17	9	25	31	50	50	50	50
18	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50
20	0	4	46	47	50	50	50
21	1	9	11	20	50	50	50
22	9	26	50	50	50	50	50
23	0	44	48	50	48	44	44
24	50	50	50	50	50	50	50
25	40	36	50	50	50	50	50
26	50	50	50	50	50	50	50
27	47	38	49	50	50	50	50
28	46	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	45	39	50	40	50	50	50
31	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50
I_C **(%)**	**65.94**	**74.88**	**86.25**	**89.81**	**97**	**95.88**	**98.13**

c. Nombre de locuteurs : 48

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	0	49	6	50	50	50
2	47	50	50	50	50	50	50

3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	34	45	46	49	50	50	50
6	44	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	0	50	50	50	50	50	50
9	3	0	0	18	40	50	50
10	0	36	40	50	50	50	43
11	50	50	49	46	44	32	46
12	50	35	50	50	50	50	50
13	50	38	50	50	50	50	14
14	0	0	0	0	13	33	48
15	50	50	50	50	50	50	50
16	20	23	7	35	43	31	50
17	0	0	8	9	26	43	50
18	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50
20	0	4	50	49	50	50	50
21	1	9	9	37	50	50	50
22	9	11	21	50	50	50	50
23	0	44	50	50	44	48	42
24	50	50	50	50	50	50	50
25	40	37	50	50	50	50	50
26	50	50	50	50	50	50	50
27	46	19	47	50	50	50	50
28	25	41	50	50	50	50	50
29	50	50	50	50	50	50	50
30	40	39	50	36	50	50	50
31	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50
33	50	50	50	50	50	50	50
34	7	19	33	50	50	50	50
35	50	50	50	50	50	50	50
36	50	50	50	50	50	50	50
37	50	50	50	50	50	50	50
38	50	49	50	50	50	50	50
39	46	50	50	50	50	50	50
40	50	50	50	50	50	50	50
41	38	50	50	50	50	50	50
42	37	50	50	50	50	50	50
43	50	50	50	50	50	50	50
44	50	50	50	50	50	50	50
45	30	23	49	19	31	23	40
46	50	50	41	50	50	29	50
47	47	41	50	50	20	0	16
48	50	50	50	50	50	50	50
I_c **(%)**	**71.42**	**77.63**	**87.46**	**89.75**	**94.21**	**93.29**	**95.79**

1.2.1.3 Apprentissage par l'algorithme K-moyennes

a. Nombre de locuteurs : 16

	Ordre du modèle						
locuteur	1	**2**	**4**	**8**	**16**	**32**	**64**
1	0	4	6	43	50	50	50

2	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	2	7	7	50	50	50	50
6	50	50	50	50	50	50	50
7	45	47	50	50	50	50	50
8	50	50	50	50	50	50	50
9	3	2	9	19	41	50	50
10	31	50	50	44	50	50	50
11	50	50	50	50	50	50	41
12	50	50	50	50	50	50	50
13	40	4	50	50	50	50	50
14	50	50	50	50	50	50	50
15	50	50	50	50	50	50	50
16	47	32	29	50	50	44	50
I_c **(%)**	**77.25**	**74.50**	**81.38**	**94.50**	**98.88**	**99.25**	**98.88**

b. Nombre de locuteurs : 32

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	2	49	34	50	50	50
2	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	1	8	18	50	50	50	50
6	50	50	50	50	50	50	50
7	45	47	50	50	50	50	50
8	12	48	50	44	50	50	6
9	1	0	26	4	21	50	50
10	34	50	43	50	44	50	50
11	50	34	46	17	0	35	11
12	50	47	50	50	50	50	50
13	20	0	9	44	50	50	50
14	32	15	28	28	50	50	50
15	50	50	50	50	50	50	50
16	46	33	32	47	50	30	50
17	49	50	50	36	50	50	50
18	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50
20	9	50	50	50	50	50	50
21	9	24	10	50	50	50	50
22	50	50	50	50	50	50	50
23	29	25	50	43	49	43	46
24	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50
27	31	43	39	29	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	49	38	50	46	50	50	50
31	50	50	50	50	50	50	50
32	42	49	50	50	50	50	50
I_c **(%)**	**74.94**	**78.94**	**87.50**	**88.88**	**94.63**	**97.38**	**94.56**

c. Nombre de locuteurs : 48

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	0	19	50	50	50	50
2	50	50	50	50	50	50	50
3	49	33	40	50	50	50	50
4	50	50	50	50	50	50	50
5	1	5	1	9	50	50	50
6	50	50	50	50	50	50	50
7	41	22	13	50	50	50	50
8	12	48	50	37	50	50	42
9	0	0	20	2	45	50	50
10	35	50	43	50	44	50	50
11	50	33	49	26	16	6	41
12	50	46	50	50	50	50	50
13	21	0	8	22	42	50	50
14	33	16	27	33	50	50	50
15	50	50	50	50	50	50	50
16	42	24	22	34	31	46	50
17	31	22	30	29	34	50	50
18	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50
20	5	45	50	50	50	50	50
21	10	25	2	41	50	50	50
22	50	50	50	50	50	50	50
23	28	25	43	42	43	40	28
24	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50
27	30	32	39	30	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	46	39	50	44	50	50	50
31	50	50	50	50	50	50	50
32	30	50	50	50	50	50	50
33	50	50	50	50	50	50	50
34	50	47	50	50	50	50	50
35	50	50	50	50	50	50	50
36	50	50	50	50	50	50	50
37	50	50	50	50	50	50	50
38	50	50	50	50	50	50	50
39	50	50	50	50	50	50	50
40	50	50	50	50	50	50	50
41	48	50	50	50	50	50	50
42	50	50	50	50	50	50	50
43	50	50	50	50	50	50	50
44	50	50	50	50	50	50	50
45	50	50	50	50	50	50	50
46	27	0	28	41	45	23	12
47	48	47	36	50	49	37	48
48	50	50	50	50	50	50	50
I_C **(%)**	**80.71**	**79.54**	**84.17**	**89.17**	**95.79**	**95.92**	**96.71**

1.2.2 Etude de l'influence du nombre de coefficients MFCC

Ordre du modèle : 16

Nombre de locuteurs : 32

Algorithme d'apprentissage : EM

	Nombre de coefficients MFCC							
locuteur	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**
1	50	44	50	50	50	50	50	50
2	24	50	50	50	50	50	50	50
3	47	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50	50
7	0	48	50	50	50	50	50	50
8	50	50	50	50	50	50	49	50
9	11	3	18	49	21	50	50	50
10	50	39	50	50	50	50	50	50
11	0	28	0	48	0	0	19	0
12	50	50	50	50	50	50	50	50
13	0	8	3	29	40	44	48	48
14	50	50	50	50	50	49	50	47
15	39	50	50	50	50	50	50	50
16	27	38	50	50	50	50	50	50
17	50	50	50	50	50	50	50	50
18	50	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50	50
20	43	50	50	50	50	50	50	50
21	50	50	50	50	50	50	50	50
22	45	16	50	50	50	50	50	50
23	0	37	50	42	49	50	50	50
24	50	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50	50
26	49	50	50	50	50	50	50	50
27	0	14	40	46	50	50	49	50
28	26	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50	50
30	50	31	50	50	50	42	50	50
31	50	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50	10
I_C **(%)**	**75.69**	**84.75**	**91.31**	**97.75**	**94.38**	**95.94**	**97.81**	**94.06**

1.2.3 Etude de l'influence du rapport signal sur bruit

Ordre du modèle : 16

Nombre de locuteurs : 32

Algorithme d'apprentissage : EM

	Rapport signal sur bruit : SNR (dB)								
locuteur	**10**	**20**	**25**	**30**	**35**	**40**	**45**	**50**	**55**
1	0	0	2	1	43	33	50	50	50

2	0	0	0	0	45	42	50	50	50
3	0	48	50	50	50	50	50	50	50
4	0	20	50	50	50	50	50	50	50
5	0	0	0	7	24	50	50	50	50
6	0	0	0	7	42	50	50	50	50
7	0	0	0	44	50	50	50	50	50
8	0	0	42	32	50	45	50	50	50
9	0	0	0	0	0	11	40	49	50
10	0	0	49	47	50	50	50	50	50
11	0	0	0	0	0	0	0	0	2
12	0	0	0	0	36	42	50	50	50
13	0	0	0	0	0	1	7	3	12
14	0	0	7	26	50	50	50	50	50
15	0	0	0	9	50	50	50	50	50
16	0	50	50	50	50	49	50	50	50
17	50	46	45	50	50	50	49	50	50
18	23	0	50	38	50	50	50	50	50
19	0	6	50	50	50	50	50	50	50
20	0	0	0	47	50	50	50	50	50
21	0	0	43	50	50	50	50	50	50
22	0	0	50	50	50	50	50	47	50
23	0	0	0	0	36	42	43	50	47
24	0	0	27	50	50	50	50	50	50
25	0	50	50	50	50	50	50	50	50
26	0	0	0	50	40	50	50	38	50
27	0	0	0	28	29	50	49	50	49
28	0	0	12	50	50	50	50	50	50
29	0	25	50	50	50	50	50	50	50
30	0	0	42	50	50	50	50	50	50
31	0	0	50	43	50	50	50	50	50
32	0	10	50	50	50	50	50	50	50
I_C **(%)**	**4.56**	**15.94**	**48.06**	**63.06**	**84.06**	**88.44**	**93**	**92.94**	**94.38**

	Rapport signal sur bruit : SNR (dB)								
locuteur	**60**	**65**	**70**	**75**	**80**	**85**	**90**	**95**	**100**
1	50	50	50	50	50	50	50	50	50
2	50	50	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50	50	50
8	50	50	50	50	50	50	50	50	50
9	50	50	50	50	50	50	50	50	50
10	50	50	50	50	50	50	50	50	50
11	0	30	0	0	0	20	0	0	50
12	50	50	50	50	50	50	50	50	14
13	19	50	50	50	50	49	50	42	50
14	50	50	50	50	50	50	50	50	49
15	50	50	50	50	50	50	50	50	50
16	42	50	50	50	50	28	50	50	50
17	50	50	50	50	50	50	50	50	50
18	50	50	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50	50	50
20	50	50	50	50	50	50	50	50	50
21	50	50	50	50	50	50	50	50	50
22	50	50	50	50	50	50	50	50	50

23	43	47	47	44	43	50	44	50	50
24	50	50	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50	50	47
26	50	50	50	50	50	50	50	50	50
27	46	50	50	29	46	50	50	46	50
28	50	50	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50	50	45
30	50	50	50	50	45	50	50	38	50
31	50	50	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50	50	50
I_C (%)	**93.75**	**98.56**	**96.69**	**95.19**	**95.88**	**96.69**	**96.5**	**95.38**	**97.19**

1.2.4 Etude de l'influence de la quantité des données de test

Nombre de locuteurs : 16

Algorithme d'apprentissage : EM

1.2.4.1 Durée des données de test : 3 secondes

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	0	23	34	40	50	50	50
2	50	50	28	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	2	7	39	50	50	50	50
6	50	50	50	50	50	50	50
7	45	50	50	50	50	50	50
8	50	50	50	50	50	50	50
9	3	0	36	50	48	50	50
10	31	42	50	50	50	50	0
11	50	50	50	50	43	50	50
12	50	50	50	50	50	50	50
13	40	13	37	50	50	50	50
14	50	48	50	50	50	50	50
15	50	50	50	50	50	50	50
16	47	50	50	39	48	50	50
I_C (%)	**77.25**	**79.13**	**90.50**	**97.38**	**98.63**	**100**	**93.75**

1.2.4.2 Durée des données de test : 1 secondes

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	13	18	27	32	38	30	35
2	32	29	17	22	30	38	42
3	47	49	39	50	50	50	50
4	38	46	41	42	43	40	50
5	19	18	36	40	39	45	35
6	36	37	50	50	50	50	48
7	39	41	41	45	45	46	38

8	47	47	50	49	47	48	46
9	13	13	6	43	38	35	42
10	28	35	33	33	37	42	36
11	36	33	29	28	33	28	21
12	40	41	43	50	50	50	50
13	37	26	33	41	45	50	49
14	49	39	45	41	50	50	50
15	36	36	40	40	46	49	50
16	30	30	38	25	21	32	40
I_C **(%)**	**67.5**	**67.25**	**71**	**79.75**	**82.75**	**85.38**	**85.25**

2. Modélisation par OGMM

2.1 Etude de l'influence de l'ordre du modèle et du rapport signal sur bruit

Nombre de locuteurs : 32

Fréquence d'échantillonnage : 8000 Hz

Algorithme d'apprentissage : EM

2.1.1 Rapport signal sur bruit : 50 dB

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	50	50	50	50	50	50	50
2	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	50	50	50	50	50	50	50
9	50	50	50	50	50	50	50
10	50	50	50	50	50	50	50
11	50	50	50	50	50	50	50
12	50	50	50	50	50	50	50
13	50	50	50	50	50	50	50
14	50	50	50	50	50	50	50
15	50	50	50	50	50	50	50
16	50	50	50	50	50	50	50
17	50	50	50	50	50	50	50
18	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50
20	50	50	50	50	50	50	50
21	50	50	50	50	50	50	50
22	50	50	50	50	50	50	50
23	50	50	50	50	50	50	50
24	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50
27	50	50	50	50	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50

30	50	50	50	50	50	50	50
31	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50
I_c **(%)**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

2.1.2 Rapport signal sur bruit : 25 dB

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	50	50	50	50	50	50	50
2	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	50	50	50	50	50	50	50
9	50	50	50	50	50	50	50
10	50	50	50	50	50	50	50
11	50	50	50	50	50	50	50
12	50	50	50	50	50	50	50
13	50	50	50	50	50	50	50
14	50	50	50	50	50	50	50
15	50	50	50	50	50	50	50
16	50	50	50	50	50	50	50
17	50	50	50	50	50	50	50
18	50	50	50	50	50	50	50
19	50	50	50	50	50	50	50
20	50	50	50	50	50	50	50
21	50	50	50	50	50	50	50
22	50	50	50	50	50	50	50
23	50	50	50	50	50	50	50
24	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50
27	50	50	50	50	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	50	50	50	50	50	50	50
31	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50
I_c **(%)**	**100**	**100**	**100**	**100**	**100**	**100**	**100**

2.1.3 Rapport signal sur bruit : 20 dB

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	50	50	50	50	50	50	50
2	50	50	50	50	50	50	50
3	50	50	50	50	50	50	50
4	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50

6	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	50	50	50	50	50	50	50
9	50	50	50	50	50	50	50
10	50	50	50	50	50	50	50
11	50	50	50	50	50	50	50
12	50	50	50	50	50	50	50
13	50	50	50	50	50	50	50
14	50	50	50	50	50	50	50
15	50	50	50	50	50	50	50
16	50	50	50	50	50	50	50
17	50	50	50	50	50	50	50
18	19	30	50	47	50	50	50
19	50	50	50	50	50	50	50
20	50	50	50	50	50	50	50
21	50	50	50	50	50	50	50
22	50	50	50	50	50	50	50
23	50	50	50	50	50	50	50
24	50	50	50	50	50	50	50
25	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50
27	50	50	50	50	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	50	50	50	50	50	50	50
31	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50
I_C **(%)**	**98.06**	**98.75**	**100**	**99.81**	**100**	**100**	**100**

2.1.4 Rapport signal sur bruit : 15 dB

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	50	50	50	50	50	50	50
2	50	50	50	50	50	50	50
3	50	0	0	0	50	49	50
4	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	50	50	50	50	50	50	50
9	50	50	50	50	50	50	50
10	50	50	50	50	50	50	50
11	50	50	50	50	50	50	50
12	50	50	50	50	50	50	50
13	50	50	50	50	50	50	50
14	50	50	50	50	50	50	50
15	50	50	50	50	50	50	50
16	50	50	50	50	50	50	50
17	50	50	50	50	50	50	50
18	0	0	1	1	42	48	25
19	50	50	50	50	50	50	50
20	15	17	50	50	50	50	50
21	50	50	50	42	50	50	50
22	50	50	50	50	50	50	50

23	47	50	50	50	50	50	50
24	25	50	48	50	50	50	50
25	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50
27	50	50	50	50	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	50	50	50	50	50	50	50
31	50	50	50	50	50	50	50
32	50	50	50	50	50	50	50
I_C **(%)**	**89.81**	**91.69**	**93.69**	**93.31**	**99.50**	**99.44**	**98.44**

2.1.5 Rapport signal sur bruit : 10 dB

	Ordre du modèle						
locuteur	**1**	**2**	**4**	**8**	**16**	**32**	**64**
1	50	50	50	50	50	50	50
2	50	50	50	50	50	50	50
3	0	0	0	0	0	0	0
4	50	50	50	50	50	50	50
5	50	50	50	50	50	50	50
6	50	50	50	50	50	50	50
7	50	50	50	50	50	50	50
8	50	50	50	50	50	50	50
9	50	50	50	50	50	50	50
10	50	50	50	50	50	50	50
11	50	50	50	50	50	50	50
12	50	48	50	34	50	50	50
13	50	50	50	50	50	50	50
14	50	50	50	50	50	50	50
15	50	50	50	50	50	50	50
16	1	40	50	13	50	50	50
17	50	50	50	50	50	50	50
18	0	0	0	0	10	0	0
19	50	50	50	50	50	50	50
20	0	0	41	0	50	50	49
21	50	50	50	50	50	50	50
22	50	50	50	49	10	45	50
23	13	50	50	50	50	50	50
24	0	0	0	29	37	50	21
25	50	50	50	50	50	50	50
26	50	50	50	50	50	50	50
27	50	50	50	50	50	50	50
28	50	50	50	50	50	50	50
29	50	50	50	50	50	50	50
30	50	50	50	50	50	50	50
31	49	26	47	50	50	7	50
32	50	50	50	50	50	50	50
I_C **(%)**	**82.06**	**85.25**	**89.88**	**85.94**	**89.94**	**90.75**	**91.86**

Annexe B

1. Estimation du modèle autorégressif

1.1 Estimation du modèle AR de production de parole

On a vu que le système de production de la parole peut être modélisé par un système AR de transmitance :

$$H(z) = \frac{\sigma}{A(z)} = \frac{X(z)}{U(z)} \quad \text{Avec} \quad A(z) = 1 + \sum_{i=1}^{P} a_p(i)\ z^{-i}$$

Cela se traduit dans le domaine temporel par la récurrence suivante :

$$x(n) + \sum_{i=1}^{P} a_p(i)\ x(n-i) = \sigma\, u(n)$$

Si on essaie d'estimer l'échantillon $x(n)$ à partir des P échantillons qui le précèdent,

$$\hat{x}(n) = -\sum_{i=1}^{P} \hat{a}_p(i)\ x(n-i)$$

alors on commis une erreur de prédiction :

$$e(n) = x(n) - \hat{x}(n) = x(n) + \sum_{i=1}^{P} \hat{a}_p(i)\ x(n-i)$$

Lorsque $\hat{a}_p(i) = a_p(i)$ pour $i = 1, 2, ..., P$, l'erreur de prédiction coïncide avec l'excitation à un facteur prés :

$$e(n) = \sigma\, u(n)$$

1.1.1 Estimation des coefficients de prédiction

Le critère usuel pour l'optimisation des coefficients de prédiction est la minimisation de la variance de l'erreur de prédiction. Cette variance vaut :

$$\sigma_e^2 = \Phi_{ee}(0) = \sum_{i,j=0}^{P} a_p(i)\, a_p(j)\, \overline{x(n-i)\, x(n-j)}$$

$$= \sum_{i,j=0}^{P} a_p(i)\, a_p(j)\, \Phi_{xx}(i-j)$$

Φ_{xx} représente la fonction d'autocorrélation, définie comme suit :

$$\Phi_{xx}(k) = E[\, x(n)\; x(n+k)\,] = \overline{x(n)\;\; x(n+k)}$$

La minimisation par rapport aux coefficients $a_p(i)$ conduit au système :

$$\frac{\delta\, \sigma_e^2}{\delta\, a_p(i)} = \sum_{j=0}^{P} \Phi_{xx}(i-j)\; a_p(j) = 0 \quad , \quad i = 1, 2, \ldots, p$$

qui peut être mis sous forme matricielle suivante :

$$A \begin{bmatrix} a(1) \\ . \\ . \\ . \\ a(P) \end{bmatrix} = - \begin{bmatrix} \Phi_{xx}(1) \\ . \\ . \\ . \\ \Phi_{xx}(P) \end{bmatrix} \quad \text{avec} \quad A = \begin{bmatrix} \Phi_{xx}(0) & \Phi_{xx}(1) & . & . & \Phi_{xx}(P-1) \\ \Phi_{xx}(1) & \Phi_{xx}(0) & . & & \Phi_{xx}(P-2) \\ . & . & . & & . \\ . & . & . & & . \\ \Phi_{xx}(P-1) & . & . & . & \Phi_{xx}(0) \end{bmatrix}$$

La matrice A est une matrice de Toeplitz symétrique. En exploitant cette structure particulière de la matrice A et en utilisant l'algorithme de *Levinson-Durbin,* on diminue considérablement la complexité des calculs.

1.1.2 Algorithme de Levinson-Durbin

$E_0 = \phi_{xx}(0)$

pour m = 1, 2, ..., p

pour i = 1, 2, ..., p

$$k_i = \frac{\left[\phi_{xx}(i) + \sum_{j=1}^{i-1} a_j^{(m-1)} \phi_{xx}(i-j)\right]}{E_{i-1}}$$

$a_i^{(m)} = k_i$

pour j = 1, 2, ..., m - 1

$a_j^{(m)} = a_j^{(m-1)} + k_i \; a_{i-j}^{(m-1)}$

$E_i = (1 - k_i^2) E_{i-1}$

pour j = 1, 2, ..., p

$a_j = a_j^{(p)}$

1.1.3 Estimation du gain du modèle

Les coefficients de polynôme $A(z)$ étant estimé, il reste à choisir une valeur adéquate du gain du modèle σ. Le gain σ peut être estimé par la variance minimale de l erreur de prédiction.

$$\sigma_{e,m}^2 = \sum_{i=0}^{p} a_p(i) \Phi_{xx}(i)$$

On a aussi :

$$\sum_{j=1}^{p} \Phi_{xx}(i-j) a_p(j) = -\Phi_{xx}(i) \quad , i = 1,2....,p$$

Ce qui implique :

$$\sigma_x^2 = \Phi_{xx}(0) = \sigma_{e,m}^2 - \sum_{i=1}^{p} a_p(i) \Phi_{xx}(i)$$

$$\Phi_{xx}(k) = -\sum_{i=1}^{p} a_p(i) \Phi_{xx}(k-i) \quad , \; k = 1,2,....,p$$

D'où $\sigma_{\hat{x}}^2 = \Phi_{\hat{x}\hat{x}}(0) = \sigma^2 - \sum_{i=1}^{p} a_p(i) \Phi_{\hat{x}\hat{x}}(k-i)$

Si on choisit $\sigma = \sigma_{e,m}$, on aura :

$\Phi_{\hat{x}\hat{x}}(k) = \Phi_{xx}(k)$, k = 0,1,2...,p

avec $\sigma^2_{e,m} = -\Phi_{xe}(0) = -\overline{x(n)\ e(n+k)}$

2. Relation entre coefficients LPCC et les coefficients LPC

Il est possible d'estimer les coefficients cepstraux $c(n)$ à partir des coefficients de prédiction $a_p(n)$.

On peut en effet écrire :

$$ln\left(\frac{1}{A_p(z)}\right) = \sum_{n=1}^{\infty} c(n)\ z^{-n}$$

et si l'on dérive chaque membre par rapport à z^{-1}, il vient :

$$\frac{A'_p(z)}{A_p(z)} = \sum_{n=1}^{\infty} n\,c(n)\ z^{-(n+1)}$$

où : $$-\sum_{i=1}^{P} i\,a_p(i)\ z^{-i+1} = \left[\sum_{j=0}^{P} a_p(j)\ z^{-j}\right]\left[\sum_{n=1}^{\infty} n\,c(n)\ z^{-n+1}\right]$$

soit : $$-i\,a_p(i) = \sum_{n=1}^{i-1} n\ c(n)\ a_p(i-n) + i\,c(i) \qquad , i>0$$

On obtient donc la récurrence :

$$c(i) = -a_p(i) - \sum_{n=1}^{i-1} (1-n/i)\,a_p(n)\ c(i-n) \qquad , i>0.$$

Bibliographie

[1] A.V.Oppenheim, R.W.Shaffer, Digital signal processing. Prentice Hall, New Jersey, 1975.

[2] Calliope, La parole et son traitement automatique. Edition Masson, Paris, 1989.

[3] C.Barras, Reconnaissance de la parole continue : Adaptation au locuteur et contrôle temporel dans les Models de Markov Cachés, thèse de doctorat de l'Université ParisVI, Mai 1996.

[4] D.A.Reynolds, An overview of automatic speaker recognition technology, MIT Lincoln Laboratory, IEEE, USA, 2002, pp. 4072-4075.

[5] D.A.Reynolds and R.C.Rose, Robust text-independent speaker identification using gaussian mixture speakers models, IEEE Trans. Speech and Audio Processing, vol. 3, no. 1, 1995, pp. 72-82.

[6] D.A.Reynolds, M.A.Zissman, T.F.Quatieri, G.C.O'Leary and B.A.Carlson, The effects of telephone transmission degradations on speaker recognition performance, Massachusetts Institute of Technology, IEEE, USA, 1995, pp. 329-332.

[7] D.M.Istrate, Détection et reconnaissance des sons pour la surveillance médicale, thèse de doctorat à l'Institut Nationale Polytechnique de Grenoble. 2003.

[8] F.Cottet, Traitement des signaux et acquisition de données. Edition Dunod, Paris, 2002.

[9] F.Coulon, Théorie et traitement des signaux, Presses polytechniques romandes, Lausanne, Edition Georgi, 1984.

[10] G.Singh, A.Panda, S.Bhattacharyya and T.Srikanthan, Vector quantization techniques for GMM based speaker verification, Indian Institute of Technology. Kampur, India. IEEE, ICASSP, 2003, pp. 65-68.

[11] H.Hadjali, M.Bouchamekh, Identification du locuteur indépendante du texte, thèse d'ingéniorat à l' Ecole Nationale Polytechnique d'Alger, Juin 2004.

[12] J.Kharoubi, Etude de techniques de classement «Machines A Vecteurs Supports» pour la vérification automatique du locuteur, thèse de doctorat de l'Ecole Nationale Supérieure de Télécommunications de Paris, Juin 2002.

[13] J.P.Campbell and D.A.Reynolds, Corpora for the evaluation of speaker recognition systems, IEEE, 1999, pp. 829-832.

[14] J.P.Campbell, Speaker recognition: A tutorial, Proceedings of the IEEE, vol. 85, pp. 1437-1462, September 1997.

[15] J.Ppelecanos, S.Myers, S.Sridharan and V.Chandran, Vector quantization based gaussian modeling for speaker verification, Queensland University of Technology, Australia, IEEE, 2000, pp. 294-297.

[16] L.Lebart, A.Morineau et M.Piron, Statistique exploratoire multidimensionnelle, Edition Dunod, 1995.

[17] L.Liu, J.He, On the use of orthogonal GMM in speaker recognition, Hearing Science, Arizona State University, USA, 1999, pp.845-848.

[18] L.X.Hung, Extraction des traits non linguistiques pour l'indexation des documents audio-visuels, thèse de DEA d'informatique, Ecole Doctorale de Mathématiques et Informatique, Université Joseph Fourier, Grenoble, Juin 2003.

[19] M.Bellanger, Traitement numérique du signal, «théorie et pratique», Edition Masson, 1987.

[20] M.Kunt, Traitement numérique des signaux. Presses polytechniques romandes, Lausanne, 1980.

[21] R.Boite et M.Kunt, Traitement de la parole. Presses polytechniques romandes, Lausanne, 1987.

[22] S.Furui, Recent advances in speaker recognition, Tokyo Institute of Technology, Japan, 1997, pp. 859-872.

[23] S.Hayakawa and F.Itakura, School of Engineering, Nagoya University, Japan, IEEE, 1995, pp. 321-324.

[24] www.mathworks.com.

[25] Y.Mami, Reconnaissance du locuteurs par localisation dans un espace de locuteurs de référence, thèse de doctorat d'état à l'Ecole Nationale Supérieure des Télécommunications de Paris, Ooctobre 2003.