REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Ecole Nationale Polytechnique

Département d'Electronique

MÉMOIRE
Présenté pour obtenir du diplôme de Magister

Option :
SIGNAL & COMMUNICATIONS

Par : Mr BOUCHAMEKH Mouslem
Ingénieur d'Etat en électronique

Ecole Nationale Polytechnique (E.N.P) – Alger.

Identification du locuteur indépendante du texte

Soutenue devant le jury :

Président :	**BERKANI Daoud**	**Professeur ENP**
Rapporteur :	**BOUSSEKSOU Boualem**	**CC ENP**
Examinateurs :	**BOUSBIA-Salah Hichem**	**CC ENP**
	GUERTI M'hania	**MC ENP**
	ZERGUI Rachid	**CC ENP**

E.N.P. 10, Avenue Hassen-Badi, El Harrach, ALGER

ملخص

يندرج هذا العمل في ميدان التشخيص الأوتوماتيكي للمتكلم, هذا الميدان الغني بالتطبيقات البالغة الأهمية بدءا من تأمين المعابر و تطبيقات التحريات الإجرامية إلى تقسيم الملفات الصوتية. نهتم في هذا العمل بالتشخيص الأوتوماتيكي المستقل عن النص المنطوق, وبالتحديد على إعطاء النماذج التشكيلية للمتكلمين, أي العمل على استخراج المعلومات الخاصة بكل شخص من خلال تحليل شاراته الصوتية ومحاولة إيجاد نماذج فعالة يمكن من خلالها تكوين نظام أوتوماتيكي فعال لتشخيص المتكلم.

كلمات مفتاحية : تشخيص المتكلم, وسائط مال سابستر (MFCC), الوسائط الطيفية الخطية LSP, التكميم الشعاعي, النموذج المتعدد الغوصيات (GMM), النموذج متعدد الغوصيات العمودية (OGMM),

Abstract

This work relates to the Automatic Speaker Recognition (ASR). The ASR is a field with many potential applications ranging from access security to audio document indexing. In this work, the text-independent speaker recognition is studied with a specific focus on speaker modeling and representation. We are especially interested to extract, from speech signals, the relative information to the identity and estimate with it the sufficiently robust speaker's model to permit speaker's recognition.

Keywords: Speaker identification, Mel Frequency Cepstral coefficients MFCC), Linear Spectral Pairs (LSP), Gaussian Mixture Model (GMM), Orthogonal Gaussian Mixture Gaussian (OGMM)

Résumé

Ce travail s'inscrit dans le domaine de la reconnaissance automatique du locuteur, domaine riche d'applications potentielles allant de la sécurisation d'accès et les applications d'ordre juridique à l'indexation de documents audio. Afin de laisser le champ à un large éventail d'applications, nous nous intéressons à la reconnaissance du locuteur en mode indépendant du texte. Nous nous intéressons plus particulièrement à la modélisation et à la représentation des locuteurs. Il s'agit d'extraire, à partir des signaux de parole, des informations relatives à l'identité, et d'estimer avec ces dernières un modèle du locuteur suffisamment robuste pour permettre son identification.

Mots clés : Identification du locuteur, paramètres Mel cepstraux (MFCC), paramètres LSP, Quantification Vectorielle (QV), GMM, OGMM.

Table des matières

Liste des figures

Acronymes

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

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

DTW: **D**ynamic **T**ime **W**arping.

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

FDR : **F**isher **D**iscriminant **R**atio.

FFT: **F**ast **F**ourier **T**ransform.

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

HMM: **H**idden **M**arkov **M**odel.

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.

MAP: **M**aximum **A** **P**osteriori.

LSP (LSF): **L**ine **S**pectral **P**airs (Line **S**pectral **F**requencies).

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

MV: **M**aximum de **V**raisemblance.

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

QV (VQ): **Q**uantification **V**ectorielle (**V**ector **Q**uantization).

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

SNR: **S**ignal to **N**oise **R**atio.

SV: **S**on **V**oisé.

SNV: **S**on **N**on **V**oisé.

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

UBM : **U**niversal **B**ackground **M**odel.

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

Introduction générale

Ce travail s'inscrit dans le domaine de la reconnaissance automatique du locuteur, domaine riche d'applications potentielles allant de la sécurisation d'accès à l'indexation de documents audio. Afin de laisser le champ à un large éventail d'applications, nous nous intéressons à l'identification du locuteur en mode indépendant du texte. Nous nous intéressons plus particulièrement à l'extraction des paramètres distinctifs et la modélisation des locuteurs.

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

Dans le deuxième chapitre nous avons présenté la modélisation et les différentes étapes nécessaires pour l'analyse du signal vocal, nous avons cités les différents paramètres acoustiques utilisés dans la majorité des systèmes de traitement de la parole. Ce chapitre donne une aidé générale sur le choix des paramètres acoustiques convenables.

Au troisième chapitre, nous nous sommes intéressés aux différentes modélisations des locuteurs, et plus particulièrement à la modélisation par mélange de gaussiennes où les locuteurs sont modélisés par une somme pondérée de gaussiennes. Dans le quatrième chapitre nous avons évalué ces différentes modélisations sur notre base de données de 60 locuteurs extraite de TIMIT.

Le dernier chapitre est consacré à l'utilisation du pitch, l'idée est de tenir en compte des informations propres à la source d'excitation, laquelle constitue un élément discriminatoire entre les locuteurs. Nous avons développé une méthodologie basé sur l'estimation de la probabilité a posteriori du pitch, l'ancien système est transformé en un nouveau, l'utilité en est dans les applications au le nombre de locuteurs est relativement important.

Chapitre 1

Introduction à la reconnaissance automatique du locuteur (RAL)

1.1 Introduction

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.

Dans ce chapitre nous introduisons au traitement de la parole, plus particulièrement a la reconnaissance du locuteur, nous donnons des informations générales sur les traits distinctifs entre locuteurs, et nous présentons les différentes étapes d'un système de reconnaissance du locuteur.

1.2 Niveaux d'information du message parlé

Du message parlé brut, tel qu'il frappe nos oreilles, nous extrayons tout d'abord des sons élémentaires; les phonèmes, définis par la possibilité qu'en ces sons, en s'opposant entre eux d'opposer des sens différents, et par la propriété d'être minimaux.

Ainsi dans l'opposition de sens entre les mots « pont » et « bon », les sons minimums opposés sont ceux correspondant aux phonèmes « p » et « b ».

Le premier niveau extrait du message brut est donc le niveau phonétique (ou phonologique). Dans un système de reconnaissance automatique, ce premier niveau est représenté par une phase de prétraitement et/ou d'analyse.

Le second niveau extrait est celui des mots, le niveau lexical. La difficulté de la reconnaissance automatique à ce niveau est liée à la taille du vocabulaire dans lequel le système est sensé retrouver les mots contenus dans la phrase prononcée. Ce second niveau est défini par le stade de la recherche lexicographique.

Le troisième niveau extrait est celui des phrases, qui est le niveau syntaxique. Ici le système se compose d'un analyseur syntaxique dont le rôle va être de constituer à partir des mots du vocabulaire proposés à la reconnaissance issu de l'étage précédent, des phrases dont la syntaxe soit cohérente avec celle du langage parlé utilisé. La complexité de cet étage est liée à la structure tolérée, qui peut aller depuis une syntaxe simple dans le cas d'un univers réduit, pour la commande par exemple d'un robot industriel, à la syntaxe extrêmement élaborée d'une langue naturelle.

Le quatrième niveau est le niveau sémantique proprement dit. L'information disponible à ce niveau est donc le sens de la phrase. L'analyseur sémantique dans un système automatique a pour tâche de sélectionner parmi toutes les phrases grammaticalement possibles celle (ou celles s'il subsiste des ambiguïtés) dont le sens est en accord avec le contexte.

Ces divers étages interagissent, la connaissance d'un des types d'information aidant à connaître les autres, en particulier en réduisant les ambiguïtés.

1.3 Reconnaissance automatique du locuteur (RAL)

La reconnaissance automatique du locuteur (RAL) est interprétée comme une tâche particulière de reconnaissance de formes. Ce domaine regroupe les problèmes relatifs à l'identification ou à la vérification du locuteur sur la base de l'information contenue dans le signal acoustique : il s'agit d'extraire du signal de parole la part relative à l'identité du locuteur. L'idée est de reconnaître une signature vocale, ou une empreinte vocale. L'utilité en est la possibilité de vérifier automatiquement l'identité d'une personne demandant d'accéder à des informations protégées. C'est le cas par exemple de toutes les transactions bancaires qui pourraient être réalisées par téléphone si les banques disposant d'un moyen de vérifier ainsi l'identité du correspondant d'après sa voix.

L'autre application est d'ordre judiciaire. Il s'agit d'identifier un criminel, d'après un échantillon de sa voix, parmi une population restreinte de suspects.
La tendance actuelle montre une évolution vers l'exécution de diverses transactions en utilisant les téléphones mobiles.

Un système de reconnaissance de locuteur procède en trois étapes : l'analyse acoustique du signal de parole, la modélisation du locuteur et une dernière étape de décision.

1.3.1 L'identification automatique du locuteur (IAL)

Il s'agit de reconnaître si, parmi une population donnée, se trouve la personne ayant prononcé un échantillon de voix inconnue, et si oui, reconnaître quelle est cette personne. Une variante de cette tâche consiste à déterminer parmi la population quelle est la personne la plus probable, en indiquant le taux de confiance associé.

1.3.2 La vérification automatique du locuteur (VAL)

Une personne fournit son code à la machine par quelque moyen que se soit (vocalement éventuellement) et la machine a pour tâche de vérifier l'identité de la personne, en fournissant une réponse soit par oui/non, soit en attribuant un niveau de confiance à son évaluation.

Dans la pratique ces deux tâches s'effectuent de la même manière, seul le test final diffère : comparaison du résultat du traitement de la donnée test avec la référence, en fonction d'un seuil dans le cas de la vérification, comparaison du test avec plusieurs références et sélection de la plus proche dans le cas de l'identification.

1.4 Mode dépendant et indépendant du texte

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

- Systèmes à texte libre (ou free-text) : le locuteur est libre de prononcer ce qu'il veut. Dans ce mode, les phrases d'apprentissage et de test sont différentes.

- Systèmes à texte suggéré (ou text-prompted) : un texte différent à chaque session et pour chaque personne, est imposé au locuteur et déterminé par la machine. Les phrases d'apprentissage et de test peuvent être différentes.

- Systèmes dépendants du vocabulaire (ou vocabulary- dependent) : le locuteur prononce une séquence de mots issus d'un vocabulaire limité. Dans ce mode, l'apprentissage et le test sont réalisés sur des textes constitués à partir du même vocabulaire.

- Systèmes personnalisés dépendants du texte (ou user-specific text dependent) : chaque locuteur a son propre mot de passe. Dans ce mode, l'apprentissage et le test sont réalisés sur le même texte.

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

1.5 Les sources d'erreurs

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

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

1.6 Traits distinctifs du locuteur

Le premier problème qui se pose lors de la réalisation d'un système de reconnaissance du locuteur est évidemment le choix des paramètres. Il s'agit de pouvoir extraire du signal vocal les caractéristiques de la voix de chaque locuteur, d'une part en sélectionnant les traits acoustiques signifiants (chargés d'une quantité suffisante d'informations sur l'identité du locuteur) et d'autre part de les coder.

La mise en oeuvre d'une tâche de reconnaissance de locuteur (ou de parole) est loin d'être facile, et ce pour deux raisons majeures. La première tient au fait que l'on ne maîtrise pas l'espace acoustique et en particulier la fonction de production d'un signal de parole. Aucune méthode analytique ne permet de prédire quelle va être la forme du signal de parole correspondant à l'émission d'un symbole donné par un locuteur particulier. La seconde, qui n'est qu'un corollaire de la première, est que la concrétisation acoustique d'un symbole donné n'est pas unique.
Intéressons-nous pour l'instant à la recherche des traits distinctifs. D'où provient donc la possibilité d'identifier un locuteur par sa voix ?

1.6.1 Production de la parole

La réponse à la question posée surgit de l'examen du mode de production du signal vocal. L'appareil phonatoire humain est constitué d'un organe respiratoire (source d'énergie), des cordes vocales (qui jouent le rôle d'oscillateurs) et des cavités buccales et nasales qui tiennent lieu de résonateur et antirésonateur. La respiration n'apparaissant pas dans le tracé du signal vocal, elle sera désormais négligée.

Dans la voix parlée, le signal est produit par une excitation acoustique du canal vocal. Ce canal peut être considéré comme un tube acoustique partant des cordes vocales (région du larynx) et allant jusqu'aux lèvres. Sur ce tube, peut se brancher en dérivation le tube constitué par les cavités nasales qui est normalement occulté et n'agit que lors de la production des sons nasalisés, par l'ouverture du velum (palais mou).

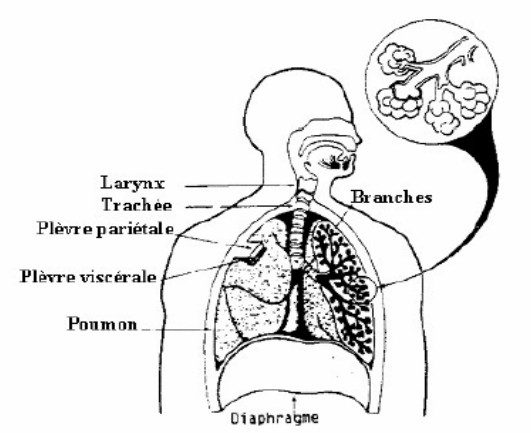

Fig. 1.1 L'appareil phonatoire humain.

Deux modes d'excitation du canal vocal existent :

1. par les vibrations des cordes vocales.
2. Par des bruits.

1.6.1.1 Les sons voisés

La vibration des cordes vocales produit les sons voisés (voyelles, semi voyelles, consonnes, nasales…etc). Les cordes vocales sont des replis musculaires recouverts d'une muqueuse, attachés aux trois cartilages (thyroïde, aryténoïdes) situés à l'extrémité de la trachée artère et constituant le larynx. Leur vibration est en fait leur accolement, puis leur séparation sous l'effet de la pression de l'air provenant des poumons, et de nouveau leur accolement sous l'effet des forces de Bernouilli produites par le passage de l'air. Les cartilages sur lesquels s'accrochent les cordes vocales régularisent la tension des cordes. Donc la fréquence des vibrations, au moyen des muscles du larynx s'appelle la fréquence fondamentale ou F_0 .

1.6.1.2 Les sons non voisés

Le second mode d'excitation est obtenu par divers bruits produits par le passage de l'air en un point de resserrement du canal vocal ou par des bruits d'occlusion ou de plosion provoqués par la fermeture ou l'ouverture des lèvres, ou des chocs de la langue contre le palais. Dans cette catégorie de sons les cordes vocales ne vibrent pas.

1.6.2 Le timbre

La caractéristique première de la voix d'un locuteur est son timbre, qui n'est perceptible que sur les sons voisés, et surtout les voyelles et semi-voyelles. Le timbre est totalement exprimé par le spectre du signal. Or, le signal vocal est constitué de la convolution du signal glottique (excitation) et de la réponse impulsionnelle du canal vocal, son spectre est le produit du spectre de la source et de celui du canal.

1.6.2.1 Spectre de la source

Les cordes vocales vibrent périodiquement en laissant échapper l'air durant un temps qui est court devant la période du phénomène. Le spectre de la source est donc composé par une fréquence fondamentale, et un grand nombre d'harmoniques, ayant une enveloppe dont la forme est proche d'une exponentielle. Une part de l'identité du locuteur est sans doute associée à cette forme [1].

1.6.2.2 Spectre du canal ou conduit vocal

Il s'agit en fait de la fonction de transfert du canal vocal en tant que tube acoustique. L'examen des spectres de signal vocal (éventuellement lissés pour éliminer l'influence de la source) montre la présence d'un certain nombre de pics dans le spectre des voyelles. Ce sont les zones correspondant aux fréquences renforcées par les différents résonateurs couplés. Ces zones sont appelées zones formantiques, ou « formants ». On numérote les formants dans l'ordre croissant des fréquences. On observe au plus 5 formants. Les deux premiers $F1$ et $F2$ caractérisent les voyelles, les formants $F3, F4, F5$ par contre, portent surtout l'information relative au locuteur [1]. Encore ici, ces deux informations ne se séparent pas totalement, $F1$ et $F2$ pour une même voyelle dépendent du locuteur, et les formants $F3$ à $F5$ varient en fonction de la consonne.

Les paramètres fréquentiels des voyelles (spectre ou formants …) sont liés à l'anatomie du sujet (donc à son identité), mais aussi à son état (fatigue musculaire par exemple, qualité de l'articulation suivant l'état émotif) ce qui peut gêner leur utilisation en tant que traits caractéristiques du locuteur.

D'autres sons peuvent se prêter à une caractéristique par le spectre, ce sont en particulier les consonnes nasales, qui font appel dans leur production à la cavité nasale, dont les caractéristiques anatomiques sont plus stables dans le temps pour chaque locuteur.

1.6.3 La mélodie

La source (impulsion de glotte) est caractérisée non seulement par son spectre, mais aussi par la période de vibration. La fréquence de la source est la fréquence fondamentale, $F0$, qui est le pitch. Cette fréquence n'est pas stable. Elle varie très rapidement en fonction du temps (mélodie), et porte une information sémantique au moyen des patrons intonatifs ou de la micromélodie (évolution du fondamental d'un phonème à un autre, ou même au sein d'un même phonème).

La mélodie porte également une information sur l'identité du locuteur qui apparaît dans la distribution statistique de la fréquence (pitch moyen, variance de pitch, …) et dans l'évolution temporelle de l'élocution, chaque locuteur ayant des patrons intonatifs favoris, mais cette dernière caractéristique est très sensible à l'imitation. Le pitch moyen permet de discriminer aisément la

voix des hommes de celles des femmes et des enfants, dont la tessiture est en moyenne plus élevée d'une octave.

1.6.4 L'articulation

Une troisième classe de traits distinctifs du locuteur est liée aux phénomènes d'articulation, et concerne non plus l'activité de la source ou des cavités, mais l'activité musculaire du locuteur.

1.6.4.1 La coarticulation

La coarticulation est l'influence d'un son sur un autre son contiguë ou voisin. Le locuteur prononçant une phrase produit une suite de phonèmes qui sont enchaînés les uns des autres de façon continue, en reliant les parties stables du signal (canal vocal en équilibre, donc signal quasi périodique) par des zones de transition. Suivant la qualité de l'articulation du locuteur, les transitions sont plus ou moins longues, et les zones stables peuvent ne pas être atteintes. La dynamique du canal vocal représenté par les variations de la fonction de transfert, est donc un ensemble de traits distinctifs du locuteur. Elle lié à la musculature. Néanmoins l'articulation est très variable suivant l'état physique ou émotif.

Ceux ci se manifestent lorsqu'un phonème est influencé par celui qui le précède et celui qui le suit. C'est le phénomène de coarticulation, lié à l'inertie du canal vocal. Le déplacement que font subir les voyelles antérieures et postérieures à une consonne voisée est caractéristique de l'élocution d'un individu, elle fournit un bon support d'identification.

1.6.4.2 Occlusives

Un second aspect de l'articulation peut servir à identifier le locuteur, c'est la durée du silence précédant l'explosion dans les plosives [p], [t], [k]. Il s'agit donc d'un paramètre temporel qui est très difficile à imiter, d'ailleurs comme la coarticulation, car régulé par des mécanismes plutôt réflexes.

1.6.4.3 Enveloppe énergétique

Fréquemment utilisée par les expérimentateurs, l'énergie du signal, en tant qu'énergie de tout ou partie du spectre à court terme du signal est également, liée dans son évolution le long d'une phrase, à l'identité du locuteur. On conçoit qu'il s'agisse aussi d'une donnée assez facile à imiter,

ce qui explique qu'elle ne soit utilisée que conjointement à d'autres paramètres, moins sensible à l'imitation.

1.6.5 Qualité des traits distinctifs

Wolf définit les conditions que doivent remplir des paramètres pour l'identification du locuteur [2]:

1 Etre aptes à représenter l'information utile sur l'identité du locuteur.

2 Etre faciles à mesurer.

3 Etre stables dans le temps.

4 Apparaître naturellement et fréquemment dans la parole.

5 Etre peu modifiables par un changement de l'environnement.

6 Ne pas être imitables.

Ce sont ces considérations qui vont guider le choix des paramètres lors de l'élaboration du système.

1.7 Description d'un système type

Les systèmes de reconnaissance du locuteur se présentent sensiblement de la même manière, les variantes étant dans le choix des solutions apportées.

1.7.1 Extraction des paramètres

Le signal analogique obtenu à partir d'un microphone ou d'un capteur téléphonique d'abord doit être numérisé, éventuellement après préaccentuation. C'est sur ce signal numérique que s'effectue l'extraction des paramètres échantillonnés à cadence fixe, ou sélectionnés par des procédures de segmentation. La numérisation peut aussi se faire après extraction des paramètres (Ex : banc de filtres analogiques).

1.7.2 Réduction des paramètres

En général les paramètres obtenus sont en nombre trop élevé pour pouvoir servir directement à la reconnaissance. Aussi faut il effectuer sur eux soit une sélection, soit une réduction par des méthodes statistiques. Les méthodes les plus utilisées sont l'expansion de Karhunen Loeve (Analyse en composantes principales ACP, ou factorielles), orthogonalisation de Gram – Schmidt,

réduction à des histogrammes, ou modélisation d'une distribution statistique (histogrammes, moments, covariances), sélection des paramètres les plus performants (F- ratios, divergence).

1.7.3 Décision

Elle se fait par le calcul d'une distance entre le lot de paramètres réduits, et la (ou les) référence(s) modélisée(s) de la même manière. C'est ici que s'introduit la distinction entre la vérification pour laquelle on calcule uniquement la distance entre les paramètres mesurés, et ceux qui ont été calculés et mémorisés pour le locuteur dont l'identité est vérifiée. Cette distance est ensuite comparée à un seuil, l'identité de locuteur est acceptée si la distance est inférieure au seuil, rejeté sinon [1].

1.8 Conclusion

L'expression vocale est une caractéristique propre au locuteur, ainsi est il possible, dans des conditions normales, de reconnaître son correspondant au cours d'une conversation téléphonique. Les variations individuelles entre locuteurs ont deux origines essentielles ?

En premier lieu, les caractéristiques de l'appareil de phonation influencent les formants, la valeur du pitch, et cela indépendamment de la phrase prononcée. D'autre part, une même phrase n'est pas prononcée de la même façon, par deux locuteurs, on observe des différences dans les débits d'élocution, dans l'étendue des variations du pitch, etc.

Chapitre 2

Modélisation et analyse du signal de parole

2.1 Introduction

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 essentiel de la modélisation d'un signal est de permettre la description de son spectre par un ensemble très limité de paramètres.

L'analyse de signal de parole consiste à chercher de mettre en évidence les caractéristiques du signal vocal tel qu'il est produit, ou parfois tel qu'il est perçu (*Analyseur Perceptuel*). L'analyseur est utilisé comme composant de base de système de reconnaissance [3].

Dans ce chapitre nous détaillons la modélisation autoregressive, et nous présentons les différents paramètres acoustiques permettants la discrémination entre locuteurs.

2.2 Modélisation autoregressive du signal vocal

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'impulsions (sons voisés).

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

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 un bruit blanc gaussien centré, soit un train périodique d'impulsions ayant pour fréquence le pitch.

En terme de transmittance, on obtient :

$$X(z) = U(z)\frac{\sigma}{A(z)} \tag{2.1}$$

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.

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

Avec $a_p(0) = 1$

$a_p(i)$: 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 :

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

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

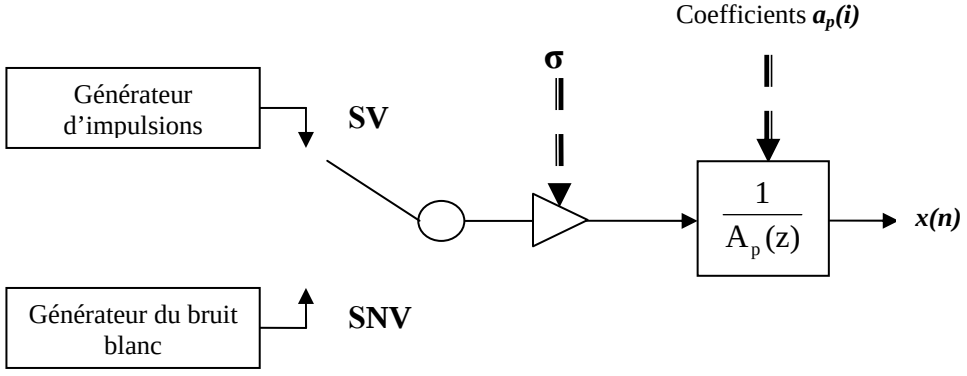

Fig. 2.1 Modèle autorégressif de production de la parole

Comme l'illustre la figure 2.1, la définition du modèle AR revient à chercher les paramètres suivants : le pitch, la décision V/NV, le gain et les coefficients de prédiction.

La modélisation autorégressive du signal vocale 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 (chevauchement de 10 *ms*).

2.3 Analyse 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.

2.3.1 Pré-traitement acoustique

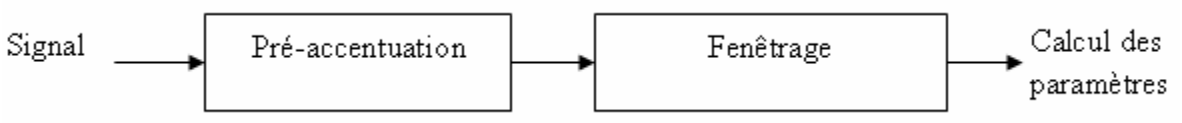

Fig. 2.2 Pré-traitement et extraction des paramètres

2.3.1.1 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 [4]. Pour pouvoir compenser cette distorsion, et accentuer les hautes fréquences, on applique un filtre de pré-accentuation passe haut de transmittance :

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

Avec $0.9 \le \alpha \le 1$.

2.3.1.2 Le fenêtrage

L'étape de fenêtrage consiste à appliquer au signal vocal 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.

Cette fenêtre est donnée par l'expression :

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

Avec $0 \le n \le N-1$

N : Le nombre d'échantillons dans une fenêtre.

2.3.2 Les paramètres acoustiques

2.3.2.1 L'énergie du signal

L'énergie du signal est un indice qui peut contribuer à augmenter les performances d'un système de reconnaissance, cette énergie correspond à la puissance du signal. Elle est souvent évaluée sur plusieurs trames de signal successives pour pouvoir mettre en évidence ses variations. La formule de calcul de ce paramètre est :

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

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

$$E = \ln\left(\sum_{n=0}^{N-1} s^2(n)\right) \tag{2.7.b}$$

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 ; généralement elle est normalisée, exprimée en décibels (par rapport à un niveau de référence).

2.3.2.2 Les coefficients de prédiction linéaire LPC

Le principe fondamental de la prédiction linéaire est qu'un échantillon donné peut être prédit à partir d'une combinaison linéaire des échantillons finis qui le précèdent [5]. Un seul jeu de coefficients du prédicteur est déterminé en minimisant les différences entre les échantillons actuels et ceux prédits. La technique de prédiction linéaire est basée sur le modèle de la production de la parole.

La fonction de transfert du modèle de la production de la parole est décrite par :

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \tag{2.8}$$

Ainsi, chaque échantillon de la parole $s(n)$ est constitué par une combinaison liénaire de p échantillons passés de la parole. Le prédicteur est défini comme un système dont la sortie est:

$$\tilde{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{2.9}$$

L'erreur de la prédiction est donnée par :

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k). \tag{2.10}$$

On cherche à trouver un ensemble de coefficients a_k de façon à minimiser l'erreur de prédiction $e(n)$ dans un certain intervalle.

La moyenne de l'erreur est donnée :

$$E = \sum_n e^2(n) = \sum_n \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right]^2 \tag{2.11}$$

$$\partial E / \partial a_i = 0 \qquad \text{pour } i = 1,..., p .$$

Alors :

$$\frac{\partial E}{\partial a_i} = -2 \sum_n \left\{ \left[s(n) - \sum_{k=1}^{p} a_k s(n-k) \right] s(n-i) \right\} = 0 \tag{2.12}$$

Cette dernière équation nous conduit à écrire :

$$\sum_n s(n)s(n-i) = \sum_n \sum_{k=1}^{p} a_k s(n-k)s(n-i). \tag{2.13}$$

On définit : $\phi(i,k) = \sum_n s(n-i)s(n-k)$.

Alors :

$$\sum_{k=1}^{p} a_k \phi(i,k) = \phi(i,0), \qquad i = 1,..., p. \tag{2.14}$$

Cet ensemble de p équations à p inconnus peut être résolu d'une manière efficace pour les coefficients de prédiction inconnus $\{a_k\}$.

On suppose que le segment de la parole est nul en dehors de l'intervalle 0 < n < L_a-1, ou L_a est la longueur de la fenêtre de l'analyse LPC. Ceci est équivalent à multiplier le signal parole d'entrée par une fenêtre de longueur finie.

$e(n)$ est non nulle uniquement sur l'intervalle 0 < n < L_a +p-1.

Ainsi

$$\phi(i,k) = \sum_{n=0}^{L_a+p-1} s(n-i)s(n-k) \qquad i = 1,..., p \tag{2.15}$$
$$k = 0,..., p.$$

On pose $m = (n-i)$, $\phi(i,k) = \sum_{m=0}^{L_a-1-(i-k)} s(m)s(m+i-k)$. (2.16)

Donc, $\phi(i,k)$ est l'autocorrélation de $s(m)$ évaluée sur $(i-k)$. D'où

$\phi(i,k) = R(i-k)$.

Donc $\sum_{k=1}^{p} a_k R(|i-k|) = R(i)$,

On obtient :

$$\begin{pmatrix} R(0) & R(1) & R(2) & \cdots & R(p-1) \\ R(1) & R(0) & R(1) & \cdots & R(p-2) \\ R(2) & R(1) & R(0) & \cdots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \cdots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{pmatrix} \tag{2.17}$$

La matrice, de dimension $p \times p$, des valeurs d'autocorrelations est une matrice de Toeplitz symétrique, tous les éléments d'une diagonale donnée sont égaux. Cette propriété peut être exploitée pour obtenir un algorithme efficace de résolution du système d'équations.

La solution la plus efficace est une méthode itérative connue sous le nom de l'algorithme de Wiener Livinson Durbin [6].

$$\left.\begin{aligned} &E_0 = R_0 \\ &k_i = -\left[R_i + \sum_{\substack{j=1 \\ \forall 1\le i\le p}}^{i-1} a_j^{(i-1)} R_{i-j}\right] / E_{i-1} \\ &a^{(i)}_i = k_i \\ &a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \quad \forall 1 \le j \le i-1 \\ &E_i = (1-k_i^{\,2}) E_{i-1} \end{aligned}\right\} \quad \forall\ i = 1,2,..., p \qquad (2.18)$$

$$a_j = a_j^{(p)} \quad \forall 1 \le j \le p.$$

$H(z)$ peut se mettre sous la forme :

$$H(z) = \frac{\sigma}{A(z)} \qquad (2.19)$$

avec

$$A(z) = 1 + \sum_{i=1}^{P} a_i \ z^{-i} \qquad (2.20)$$

2.3.2.3 Les paramètres LSP (*Line Spectral Pair* ou *Line Spectral Frequencies LSF)*

Les paramètres LSP (*Line Spectral Pair)* ont été présentés la première fois par itakura comme représentation alternative d'information spectrale du LPC. Ils contiennent exactement la même information que Les coefficients LPC [18].

En analyse par prédiction linéaire, un segment de parole est supposé être généré comme sortie d'un filtre tous pôles $H(z) = 1/A(z)$. Où $A(z)$ est un polynôme en z appelé le *filtre inverse* dont l'expression est donnée par:

$$A(z) = 1 + a_1 \ z^{-1} + \ldots + a_p \ z^{-p}. \qquad (2.21)$$

Par définition, un filtre stable, tous ses pôles sont à l'intérieur du cercle unité sur le plan complexe des z. Par conséquence, son filtre inverse est à minimum de phase, parce qu'il ne

possède aucun zéro ou pôle à l'extérieur du cercle unité. Le polynôme $A_p(z)$ associé à l'ordre p d'analyse LPC, vérifie la relation suivante:

$$A(z) = \frac{1}{2}\left[P(z) + Q(z)\right]$$
$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}) \qquad (2.22)$$
$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1})$$

Les LSP sont les fréquences des racines des polynômes $P(z)$ (symétrique), et $Q(z)$ (antisymétrique).

Parce que les coefficients LPC sont réels, le théorème fondamental de l'algèbre garantit que les racines de $A(z)$, $P(z)$ et $Q(z)$ sont des pairs conjugués, et à cause de cette propriété le demi plan complexe supérieur est redondant.

Les polynômes *P (z)* et *Q(z)* possèdent des racines sous la forme e^{jw_i} pour i = 0, 2, ..., p+1. Les paramètres $\{w_i\}$ i=0,2,...,p+1, définissent alors les " *Line Spectral Frequencies*" (LSF). Il est important de noter que $w_0 = 0$ et $w_{p+1} = \pi$, sont des racines fixées, des polynômes Q *(z)* et *P(z)* respectivement et seront exclus de l'ensemble des paramètres LSF.

Les polynômes *P(z)* et *Q(z)* possèdent des propriétés très intéressantes et importantes :

1- les racines des polynômes *P (z)* et Q *(z)* sont sur le cercle unité.

2- Les racines des polynômes *P (z)* et Q *(z)* sont entrelacées, c'est à dire dans un ordre ascendant et se trouvent dans le premier et le second quadrants du plan complexe Z ce qui se traduit par la relation suivante:

$$0 = w_0^{(Q)} < w_1^{(P)} < w_2^{(Q)} < \ldots < w_p^{(Q)} < w_{p+1}^{(P)} = \pi. \qquad (2.23)$$

Cette dernière relation exprime la propriété *d'ordonnancement* des LSF [6].

2.3.2.4 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 :

$$c_k = - a_k - \sum_{i=1}^{k-1} (1 - \frac{i}{k}) a_i \; c_{k-i} \, , \; k > 0 \tag{2.24}$$

2.3.2.5 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 [6].

$$s(t) = e(t) * h(t) \tag{2.25}$$

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 du logarithme sur le module de la Transformée de Fourier de $s(t)$ dans l'équation (2.25) donne :

$$log\,|S(f)| = log\,|E(f)| + log\,|H(f)| \tag{2.26}$$

Par une transformée de Fourier inverse, on obtient :

$$s'(cef) = e'(cef) + h'(cef) \tag{2.27}$$

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* [10].

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 l'application d'un simple fenêtrage dans le domaine quéfrentiel (liftrage passe-bas) pour le conduit vocal.

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

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'informations utiles dans le signal de parole. Les échelles perceptives les plus utilisées sont l'échelle Mel et l'échelle Bark[6] [13].

- Echelle Mel

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{2.28}$$

- Echelle Bark

$$Bark(f) = 6\, Arc\sinh\left(\frac{f}{1000}\right) \tag{2.29}$$

f représente la fréquence [Hz].

La procédure de calcul des coefficients MFCC est illustrée dans la figure 2.3

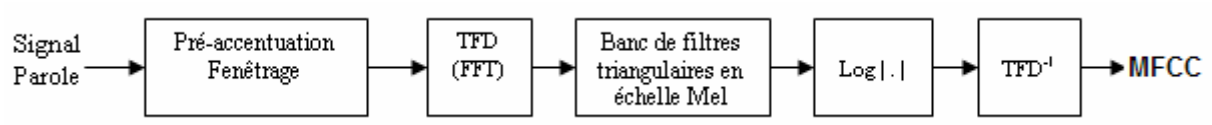

Fig. 2.3 Calcul des coefficients MFCC

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_e est la fréquence d'échantillonnage, la Transformée de Fourier Discrète à court terme $S(k)$ est obtenue avec la formule :

$$S(k) = \sum_{n=0}^{N-1} s(n)\, exp\left(\frac{-j\,2\pi\, n\, k}{N}\right),\ 0 \le k \le N-1 \tag{2.30}$$

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 :

$$B_m = B_b + m \frac{B_h - B_b}{M + 1} \quad , \quad 0 \le m \le M + 1 \tag{2.31}$$

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 (2.31), les points f_m discrets correspondants sont calculés d'après :

$$f_m = B^{-1}\left(B_b + m \frac{B_h - B_b}{M + 1} \right) \tag{2.32}$$

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

$$B^{-1}(x) = 700 \left(10^{\frac{x}{2595}} - 1 \right) \tag{2.33}$$

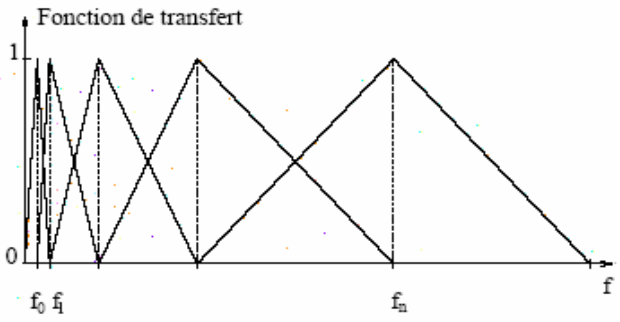

Fig. 2.4 Banc de filtres sur l'échelle linéaire

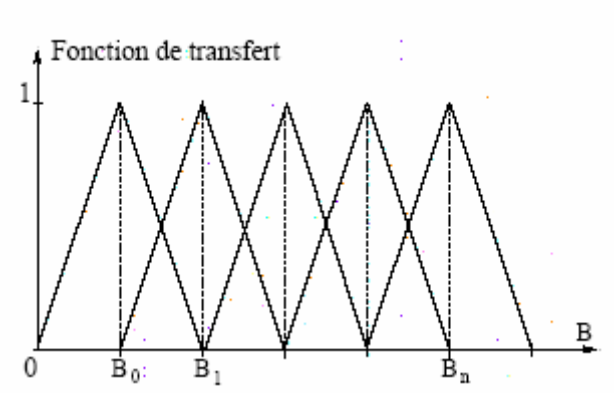

Fig. 2.5 Le banc de filtres sur l'échelle Mel

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 par :

$$c_k = \sum_{i=1}^{M} log\ E_i\ cos\left[\frac{\pi k}{M}\left(i-\frac{1}{2}\right)\right] \quad , 1 \le k \le d \tag{2.34}$$

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.

2.3.2.6 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 [8].

2.3.2.7 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 :

$$\Delta c_k(t) = \frac{\sum_{i=-n_\Delta}^{n_\Delta} i\, c_k(t+i)}{\sum_{i=-n_\Delta}^{n_\Delta} i^2} \tag{2.35}$$

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

$$\Delta E(t) = \frac{\sum_{i=-n_\Delta}^{n_\Delta} i\, E(t+i)}{\sum_{i=-n_\Delta}^{n_\Delta} i^2} \tag{2.36}$$

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 $n_\Delta = n_{\Delta\Delta} = 2$).

2.3.3 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*).

2.3.3.1 Le rapport discriminant de Fisher (FDR)

Ce rapport estime le pouvoir discriminant 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 :

$$FDR = \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} \left(\bar{c}[i] - \bar{c}[j] \right)^2}{\sum_{i=1}^{k} Var(\bar{c})[i]} \tag{2.37}$$

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. L'inconvinient de ce critère est qu'il n'intègre pas les relations de corrélation entre les paramètres.

2.3.3.2 Analyse Discriminante Linéaire (ADL)

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

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

$$P_{tr} = A\,P \tag{2.38}$$

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

La matrice A est appelée « la matrice de covariance de Fisher », elle prend la forme suivante :

$$A = \left[u_1, u_2, ..., u_D \right]^t \tag{2.39}$$

Où u_i sont les valeurs propres du rapport entre la matrice de dispersion inter-locuteurs S_b et la matrice de dispersion intra-locuteur S_w [20].

$$A = S_w^{-1}\ S_b \tag{2.40}$$

2.3.3.3 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 :

$$P_{tr} = \mathrm{A}\ \mathrm{P} \tag{2.41}$$

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

La matrice A prend la forme suivante :

$$A = \left[\mathrm{u}_1, \mathrm{u}_2, ..., \mathrm{u}_\mathrm{D} \right]^t \tag{2.42}$$

Où les u_i désignent les vecteurs propres de la matrice de covariance ordonnés de manière croissante [1].

2.4 Conclusion

Les méthodes d'analyse de la parole dépendent de plusieurs paramètres, selon l'application envisagée, dans notre cas l'identification. On a choisi pour notre étude l'analyse cepstrale issue du modèle de perception de l'oreille, étant donné qu'elle fournit des paramètres discriminants étendus sur tout le spectre de la bande de perception de l'oreille humaine, et caractérise efficacement les coordonnées du locuteur dans l'espace spectral. Aussi la prédiction linéaire nous fournit des coefficients qui peuvent caractériser le conduit vocal.

Dans notre étude on va évaluer l'utilisation des coefficients MFCC et les LSP pour un système d'identification de locuteur.

Chapitre 3
Identification du locuteur par mélange de gaussiennes (GMM).

3.1 Introduction

Les mélanges de gaussiennes sont utilisés pour modéliser un locuteur donné par une somme pondérée de gaussiennes. On peut assimiler un modèle GMM (*Gaussien Mixture Models*) à un HMM (*Hidden Markov Model*) à un seul état. On ne modélise donc pas les aspects temporels du signal. Cette méthode est la plus utilisée en ce qui concerne la reconnaissance du locuteur en mode indépendant du texte.

Dans ce chapitre, on donne des idées générales sur les différentes modélisations déjà existantes, puis la modélisation GMM est exposée en détail, et on termine par une conclusion.

3.2 Modélisation des locuteurs

Comme dans le cas de la reconnaissance de la parole, le problème de la reconnaissance du locuteur peut se formuler selon un problème de classification. Différentes approches ont été développées, néanmoins on peut les classer en quatre grandes familles [18].

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

3.2.1 L'approche vectorielle

3.2.1.1 Reconnaissance du locuteur à base de DTW

La reconnaissance par DTW (*Dynamique Time Warping*) repose sur le principe que chaque mot est représenté par une prononciation de référence (*template*). Compte tenue des

décalages temporels entre les différentes prononciations d'un même mot, l'algorithme met en correspondance des séquences de paramètres par distorsion temporelle (*Time Warping*). La programmation dynamique permet d'aligner temporellement une phrase de test avec une phrase d'apprentissage ce qui signifie que c'est une technique exclusivement utilisée en mode dépendant du texte [21][22].

3.2.1.2 Quantification vectorielle (QV)

Il s'agit de représenter l'espace acoustique par un nombre fini de vecteurs acoustiques. Cela consiste à faire un partitionnement de cet espace en régions, qui seront représentées par leur vecteur centroide. Pour déterminer la distance d'un vecteur acoustique à cet espace, on effectue une mesure de distance avec chacun des centroides des régions et on retient la distance minimale. Si le vecteur acoustique provient du même locuteur pour lequel on a établi le dictionnaire de quantification, la distorsion sera en général moins grande que si ce vecteur provient d'un autre locuteur. Ainsi, on va représenter un locuteur par son dictionnaire de quantification [23], [24][26].

3.2.2 L'approche statistique

3.2.2.1 Modèles de Markov cachés

Les modèles de Markov cachés (ou HMM pour *Hidden Markov Models*) ont été initialement introduits en reconnaissance de la parole. Puis leur utilisation s'est étendue peu à peu au domaine de la reconnaissance du locuteur. Dans cette dernière approche, il ne s'agit plus d'une mesure de distance d'une forme acoustique à une référence, mais de la probabilité que la forme acoustique ait été engendré par le modèle de référence du locuteur. Le modèle d'un locuteur est constitué de l'association d'une chaîne de Markov, une succession d'états avec des probabilités (probabilité d'observation d'un vecteur acoustique dans un état) [25], [27][28].

3.2.2.2 Les mélanges de gaussiennes

La reconnaissance du locuteur par mélange de gaussiennes (ou GMM pou *Gaussien Mixture Models)* consiste à modéliser le signal d'un locuteur par une somme pondérée de composantes gaussiennes [13]. Ainsi une large gamme de distributions peut être parfaitement représentée. Chaque composante des gaussiennes est supposée modéliser un ensemble de classes acoustiques. L'utilisation de ce type de modèles semble être prometteuse. Il semble bien modéliser les caractéristiques spectrales des voix des locuteurs, et il est relativement

simple à mettre en oeuvre. Les mélanges de gaussiennes sont considérés comme un cas particulier des HMM et une extension de la quantification vectorielle.

3.2.2.3 Mesures statistiques du second ordre.

Cette partie présente une famille de mesures de similarité entre locuteurs. Ces mesures reposent sur les caractéristiques du second ordre d'une séquence de vecteurs, c'est à dire sur le vecteur moyen et la matrice de covariance de cette séquence. Plusieurs mesures de distance peuvent être utilisées : Le rapport de vraisemblance, la distance de *Kullbak-Leibler*, maximum de vraisemblance, test de sphéricité, déviation absolue des valeurs propres. Ces mesures donnent des résultats encourageants sur la parole propre, et, naturellement, voient leurs performances se dégrader sur la parole téléphonique. De part leur relative simplicité, ces mesures peuvent également servir de référence pour évaluer la qualité d'une base de donnée [29][31].

3.2.3 L'approche connexionniste

Les réseaux de neurones ont été assez largement utilisés en reconnaissance du locuteur. Ils offrent en effet une bonne alternative au problème de la discrimination entre les locuteurs. Ces outils de classification permettent de séparer des classes, dans un espace de représentation donné, de façon non linéaire. L'inconvénient important de l'application de cette technique en identification du locuteur est le coût important lié à l'ajout d'un nouveau locuteur dans la base de référence (ce n'est pas le cas en vérification du locuteur). On peut aussi utiliser les réseaux de neurones en les couplant à d'autres techniques, comme par exemple les modèles de Markov cachés. On parle alors de méthodes hybrides [30], [32].

3.2.4 L'approche relative

Cette nouvelle technique consiste à modéliser un locuteur non plus de façon absolue mais relativement à un ensemble de locuteurs bien appris.

3.3 Modèle du mélange de gaussiennes

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 définit 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 $\sum_m^s$:

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

et les π_m^s représentent les poids du mélange, avec $\sum_{m=1}^{M} \pi_m^s = 1$.

Un locuteur est donc modélisé par un ensemble de paramètres noté λ_s :
$\lambda_s = \left\{\pi_m^s, \mu_m^s, \sum_m^s\right\} m = 1,...,M$

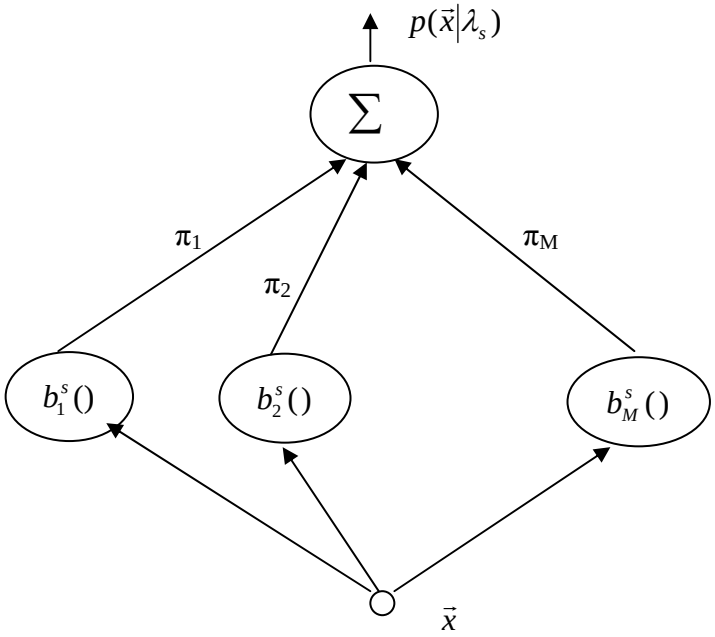

Fig. 3.1 Modèle de Mélange de Gaussiennes

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

3.4 Apprentissage du modèle

Il s'agit, lors de la phase d'apprentissage, d'estimer l'ensemble λ des paramètres d'un modèle GMM de locuteur. La méthode conventionnelle est celle du Maximum de vraisemblance (MV) dont le but est de déterminer les paramètres du modèle qui maximisent

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

$$p(X|\lambda) = \prod_{n=1}^{N} p(x_n|\lambda) = \prod_{n=1}^{N} \sum_{m=1}^{M} p(x_n|\pi_m, \mu_m, \Sigma_m) \quad (3.3)$$

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

3.4.1 Apprentissage par Maximum de Vraisemblance (MV)

L'algorithme expectation maximisation (EM)

L'algorithme EM (*Expectation Maximisation*) fait intervenir à la fois des observations *X* et des variables manquantes (l'indice de la gaussienne *m = 1,...* , M). Cet algorithme maximise, de façon itérative, la fonction de la vraisemblance. Cette maximisation n'est pas directe, elle fait intervenir la fonction auxiliaire $Q(B, B^{(t)})$ qui est définie comme étant l'espérance mathématique du logarithme de la vraisemblance jointe (incluant les variables observées et les variables cachées) sur l'ensemble complet des variables d'entraînement, calculée à la base des paramètres courants [3], à savoir :

$$Q(\theta, \theta^{(t)}) = \sum\sum p(m|x_n, \theta^{(t)}).\log p(x_n, m|\theta) \quad (3.4)$$

où θ désigne l'ensemble des paramètres à estimer $(\pi_m, \mu_m \; et \; \Sigma_m)$ et $\theta^{(t)}$ l'ensemble des paramètres estimés à l'itération *t*. Ce qui donne, après calcul :

$$Q(\theta, \theta^{(t)}) = \sum\sum \gamma_{n,m}^{(t)} \left[\log \pi_m - \frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_m| \right] - \sum_{m=1}^{M}\sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\frac{1}{2}(x_n - \mu_m) \right] \quad (3.5)$$

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

$$\gamma_{n,m}^{(t)} = \frac{\pi_m^{(t)} p(x_n \left| \mu_m^{(t)}, \Sigma_m^{(t)} \right.)}{\sum_{k=1}^{M} \pi_k^{(t)} p(x_n \left| \mu_k^{(t)}, \Sigma_k^{(t)} \right.)} \tag{3.6}$$

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

$$Q(\theta, \theta^{(t)}) = \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \log \pi_m - \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[Cste + \log \sigma_m^2 + \frac{(x_n - \mu_m)^2}{\sigma_m^2} \right] \tag{3.7}$$

où σ_m^2 est un élément diagonal de la matrice de covariance.

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

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

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

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

et pour les variances :

$$\sigma_m^{2(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} (x_n - \mu_m^{(t)})^2}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{3.10}$$

3.4.2 Apprentissage par Maximum A Posteriori (MAP)

L'algorithme EM est un des algorithmes les plus importants et les plus puissants en estimation statistique. De plus, il bénéficie d'une preuve de convergence garantissant que l'itération de l'étape d'estimation et de maximisation converge vers un maximum de la

fonction de vraisemblance [3]. Cependant, ses limites apparaissent lorsqu'on dispose de peu de données. Donc, il est important d'introduire de l'information a priori. Par conséquent, on ne cherche plus à maximiser la vraisemblance des données mais plutôt la probabilité a posteriori.

Les formules de ré-estimation, pour une gaussienne *m*, sont les suivantes [33] :

- Les poids des gaussiennes :

$$\pi_m = \frac{n_m^0 + n_m}{\sum_{k=1}^{M}(n_k^0 + n_k)} \tag{3.11}$$

- Les vecteurs des moyennes :

$$\mu_m = \frac{n_m^0 \overline{X_m^0} + n_m \overline{X_m}}{n_m^0 + n_m} \tag{3.12}$$

- Les variances :

$$\sigma_m^2 = \frac{n_m^0 \overline{X_m^0 X_m^{0\prime}} + n_m \overline{X_m X_m'}}{n_m^0 + n_m} - \mu_m \mu_m' \tag{3.13}$$

où n (respectivement n^0) représente le poids, $\overline{X}$ (respectivement $\overline{X^0}$) le moment d'ordre 1 et $\overline{XX'}$(respectivement $\overline{X^0 X^{0\prime}}$) le moment d'ordre 2 des données à adapter X (respectivement des données initiales X^0).

L'apprentissage incrémental consiste à effectuer quelques itérations d'apprentissage sur les données d'adaptation en conservant l'information apportée par les données initiales X^0. Dans le cas où de nombreuses données sont disponibles, l'apprentissage incrémental (ou plus généralement l'estimateur MAP) converge vers les estimateurs du maximum de vraisemblance. Il permet d'obtenir de nouveaux modèles avec peu de données. Ces estimées seront plus fiables que celle obtenue par MV étant donné qu'elles intègrent des connaissances p priori.

Cette approche est la plus utilisée en reconnaissance du locuteur en mode indépendant du texte [15].

3.4.3 Initialisation

Les valeurs initiales d'une densité multi-gaussienne peuvent être obtenues par différentes méthodes comme par exemple, la QV (Quantification Vectorielle) ou par éclatement de

gaussiennes. Cette initialisation est suivie par apprentissage EM ou par une adaptation incrémentale. En GMM, le modèles initial correspond au modèle du monde UBM (*Universal Background Model*) [15].

3.5 Décision

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

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

$$\hat{s} = \arg\max_{1 \le s \le S} p(\lambda_s | X) \tag{3.14}$$

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

$$\hat{s} = \arg\max_{1 \le s \le S} \frac{p(X|\lambda_s)}{p(X)} p(\lambda_s) \tag{3.15}$$

en supposant l'équiprobabilité d'apparition des locuteurs $p(\lambda_s) = \frac{1}{S}$, la loi de classification devient :

$$\hat{s} = \arg\max_{1 \le s \le S} p(X|\lambda_s) \tag{3.16}$$

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

$$\hat{s} = \arg\max_{1 \le s \le S} \sum_{n=1}^{N} \log p(x_n|\lambda_s) \tag{3.17}$$

3.6 Conclusion

Les mélanges de gaussiennes constituent l'état de l'art en reconnaissance automatique du locuteur, en mode indépendant du texte. Il existe plusieurs techniques pour faire apprendre un modèle GMM. En premier lieu, l'algorithme EM permet d'estimer les paramètres du modèle tout en offrant un formalisme théorique et une preuve de convergence. Malheureusement, ses limites apparaissent lorsqu'on dispose de peu de données. Dans ce cas, il est plus judicieux

d’utiliser un apprentissage par adaptation MAP. Cette estimation sera plus fiable que celle obtenue par MV sachant qu’elle intègre des connaissances a priori.

Chapitre 4

Évaluation expérimentale de reconnaissance par GMM

4.1 Introduction

Ce chapitre présente l'évaluation expérimentale de deux approches de l'identification de locuteur : la quantification vectorielle (QV) et la modélisation mixture gaussienne (GMM).
Avant de commencer la simulation il faut toujours préparer une base de données bien organisée, et sur cette base que le système d'identification va être testé. Dans la première partie on donne la description de la base de données utilisée dans cette simulation, dans les autres parties on donne les différentes étapes pratiques pour entraîner et tester un système d'identification de locuteur, ainsi que les résultats et les interprétations de notre simulation.

4.2 Description de la base de données utilisée

Dans le cadre de ce travail, on a utilisé une base de données composée de 60 locuteurs (25 hommes et 35 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 les 3 autres phrases sont utilisées pour le test.

4.3 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, les systèmes de reconnaissance utilisent sauvent 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.

Les paramètres LSF (*Line Spectral Pairs*) contiennent des informations sur le conduit vocal, et peuvent être utilisés pour caractériser un locuteur. Dans ce travail on va évaluer leur utilisation dans un système d'identification de locuteur.

Dans nos expériences, une analyse est appliquée toutes les 10 ms sur des fenêtres d'analyse de 20 ms (par glissement et recouvrement des fenêtres d'analyse). A chaque trame, on associe un vecteur de représentation acoustique.

4.3.1 Extraction des paramètres

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

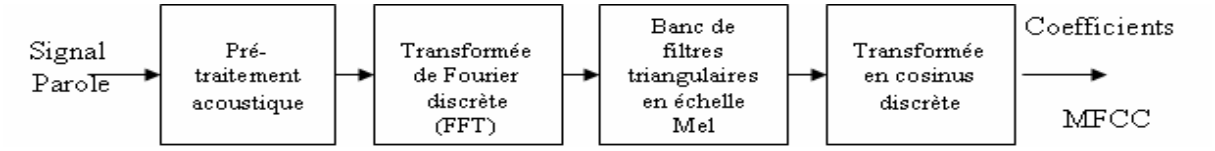

Fig. 4.1 Extraction des coefficients MFCC

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 transmittance $H(z) = 1 - 0.95\, z^{-1}$.

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 glissante de durée de 20 ms avec déplacement de 10 ms.

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

$$w(n) = 0.54 + 0.46 \cos\left[\frac{2\pi n}{N-1}\right] \quad \text{et} \quad 0 \le n \le N-1 \tag{4.1}$$

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

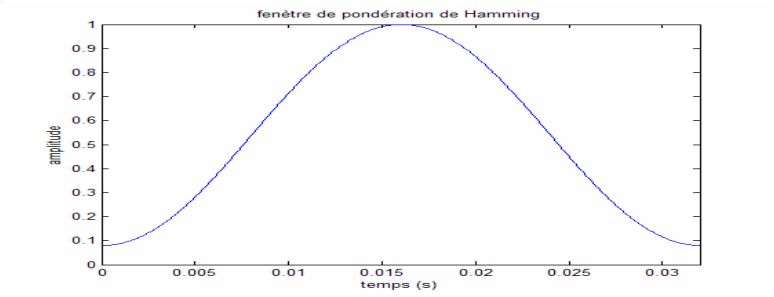

Fig. 4.2 Fenêtre de pondération de Hamming.

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

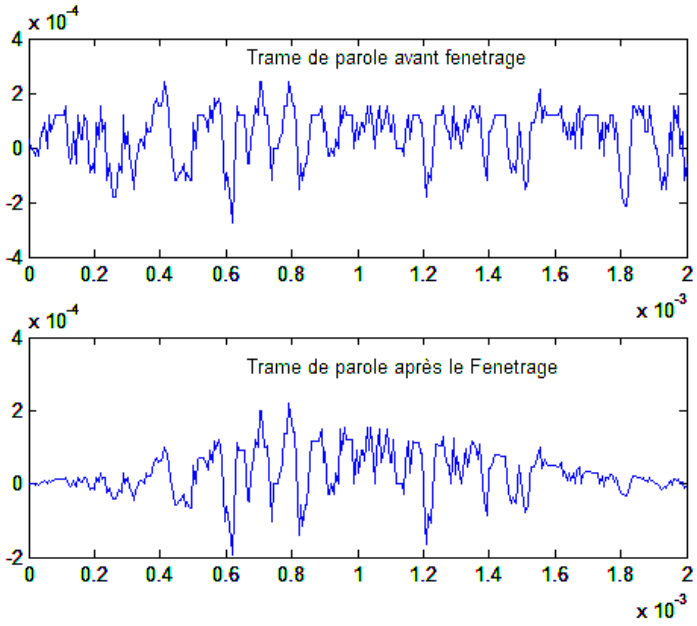

Fig. 4.3 Fenêtrage d'une trame de parole

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 :

$$c_k = \sum_{i=1}^{M} log\, E_i \cos\left[\frac{\pi k}{M}\left(i - \frac{1}{2}\right)\right] \quad (4.2)$$

4.3.2 Détection et élimination de silence

Les périodes de silences ne portent aucune information et peuvent diminuer les performances d'un système de reconnaissance, 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 portion du signal de niveau énergétique inférieur au seuil prédéterminé sera éliminée. La figure 4.4 illustre une trame de parole avant et après élimination de silence.

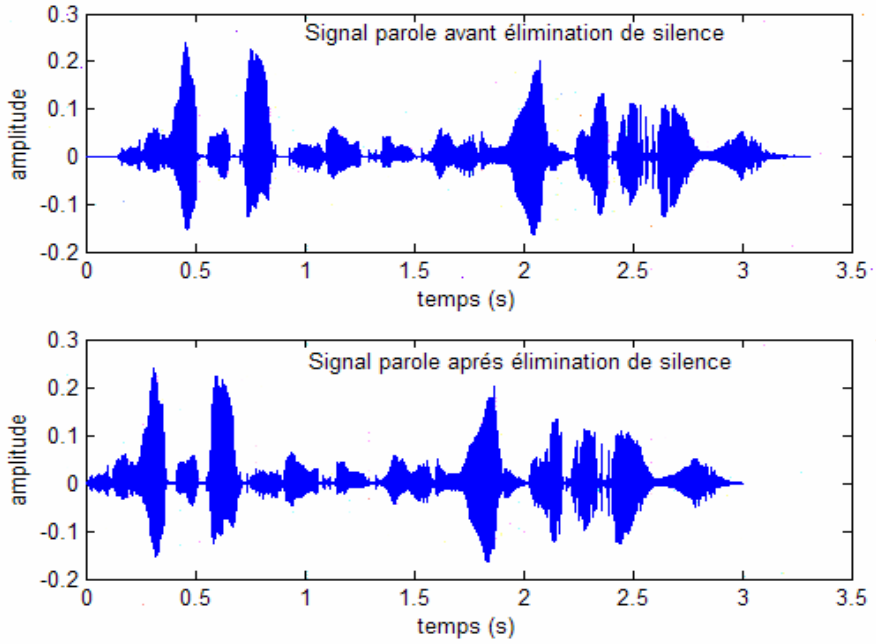

Fig. 4.4 Elimination de silence

4.4 Protocole d'évaluation

Nous allons évaluer les performances des deux approches GMM et OGMM sur un ensemble de 60 locuteurs (ensemble fermé). Il s'agit d'identifier un locuteur parmi les 60 locuteurs et de calculer le taux d'identification correcte défini par :

$$I_c = \frac{\text{Nombre de segments de test correctement identifiés}}{\text{Nombre total de segments de test}} \times 100 \quad (4.3)$$

4.5 Evaluations expérimentales

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

Pour ces trois techniques, nous étudions l'influence des paramètres suivants sur le taux d'identification :

4.6.2 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 (la bande téléphonique).

4.5.2 L'ordre du modèle

On varie l'ordre du modèle (nombre de centroïdes pour la quantification vectorielle) ou le nombre de composantes gaussiennes de 2 jusqu'à 60 gaussiennes.

4.5.3 La qualité des paramètres d'identification

Le vecteur des paramètres joue un rôle principal pour l'identification de locuteur. Dans ce travail on va évaluer l'utilisation des vecteurs MFCC, des vecteurs LSP et des vecteurs hybrides (MFCC & LSP).

4.6 Les résultats obtenus

4.6.1 La Quantification vectorielle

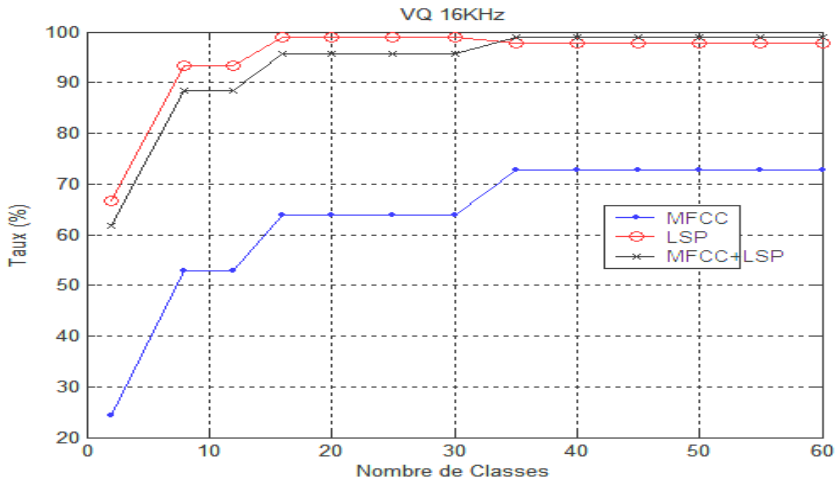

Fig. 4.5 La Quantification Vectorielle à Fe= 16 kHz

La figure 4.5 trace les variations des taux d'identification en fonction du nombre de classes (nombre de centroïdes extraits) pour la base de données échantillonnées à 16 kHz, le taux d'identification est croissant avec le nombre de classes, il devient presque stable au delà de 16 classes, un léger basculement entre les taux des vecteur LSP et MFCC plus LSP obtenu pour 35 centroïdes, le taux maximal est de 98% obtenu entre 16 et 35 classes avec les vecteurs LSP, et de 35 à 60 classes avec les vecteurs MFCC plus LSP, avec les MFCC on obtient presque la même allure que les vecteurs LSP, sauf que les valeurs de taux d'identification correcte sont inférieures (Valeur maximale 74%).

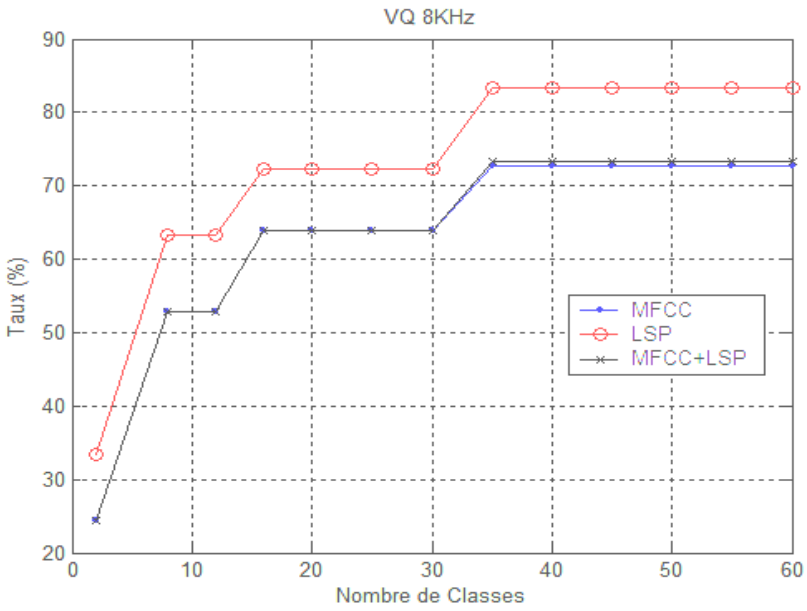

Fig. 4.6 La Quantification Vectorielle à Fe= 8 kHz

Sur la figure 4.6 on a tracé les résultats obtenus sur la base de données de *8* kHz, l'impact de la réduction de la fréquence d'échantillonnage est remarquable en comparant les deux figures (4.5 et 4.6), les performances obtenue avec les vecteurs LSP se dégradent quand on diminue la fréquence, la valeur maximale est de 84%. Mais avec les vecteurs MFCC on garde les mêmes valeurs de taux d'identification correcte trouvées sur 16 kHz. Les valeurs de taux d'identification correctes obtenue en utilisant les vecteurs MFCC et MFCC plus LSP sont presque les mêmes, la valeur maximale est de 74%.

Il est clair que les LSPs sont efficaces pour différencier entre les locuteurs. Dans ce cas on peut bien remarquer que les résultats obtenus sont meilleurs avec les LSPs que ceux obtenus avec les MFCCs.

4.6.2 Les mélanges de gaussiennes GMM

Avec les mélanges de gaussiennes, chaque classe est présentée par une gaussienne, donc le nombre de classes est le nombre de gaussiennes utilisées pour modéliser les références de chaque locuteur.

On varie le nombre de gaussiennes de 2 jusqu'à 60, la figure suivante montre les valeurs des taux d'identification correcte en fonction du nombre de gaussiennes utilisées. Le système d'identification est évalué sur la base de données de 16 kHz.

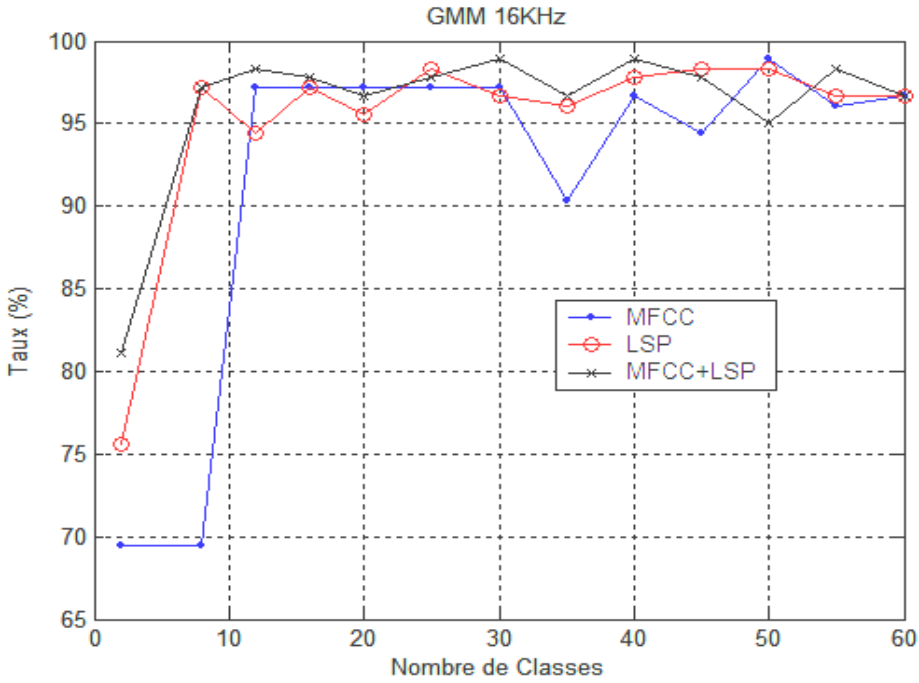

Fig. 4.7 GMM à Fe= 16 kHz

Pour 2 gaussiennes, le meilleur taux d'identification obtenu est de 82% avec les vecteur de MFCC plus LSP, un régime stable s'établit avec les vecteurs MFCC au-delà de 12 gaussiennes (taux de 97%), puis une dégradation jusqu'à 90% pour 35 gaussiennes, avec les vecteurs LSP et MFCC plus LSP on voit l'oscillation de la valeur de taux d'identifications correctes entre 95% et 98%.

On dit qu'un maximum local de la valeur de taux d'identification correcte est obtenu avec les vecteurs des MFCC (97% pour 12 à 30 gaussiennes).

Pour trouver l'impact de la fréquence d'échantillonnage sur le taux d'identification correcte, on a évalué le mélange de gaussiennes sur la base de données de 8 kHz. Les résultats obtenus sont tracés sur la figure 4.8.

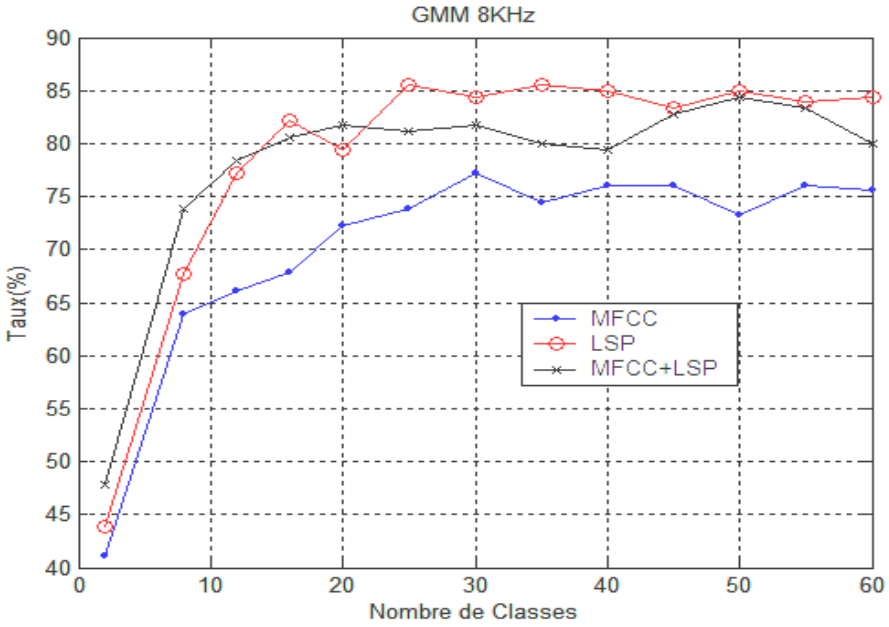

Fig. 4.8 GMM à Fe= 8 kHz

Le système d'identification à base de GMM est sensible à la réduction de la fréquence d'échantillonnage (8 kHz), le taux d'identification correcte se dégrade jusqu'à 75% en moyenne avec les vecteurs MFCC, une légère amélioration de l'identification est obtenue en utilisant les vecteurs LSP, avec lesquels le taux d'identification correcte atteint 85%, le nombre de gaussiennes nécessaires est de 25 gaussiennes.

L'augmentation de l'ordre du modèle permet d'affiner la séparation des classes acoustiques, ce qui se traduit par un accroissement de taux d'identification correcte.

4.6.3 Les mélanges de gaussiennes orthogonales (OGMM)

Les vecteurs d'entraînement (d'apprentissage) subissent une orthogonalisation avant qu'ils ne soient utilisés pour générer les gaussiennes.

Sur la figure 4.9, on voit les graphes tracés pour l'orthogonal GMM, et pour la base de données échantillonnée à 16 kHz.

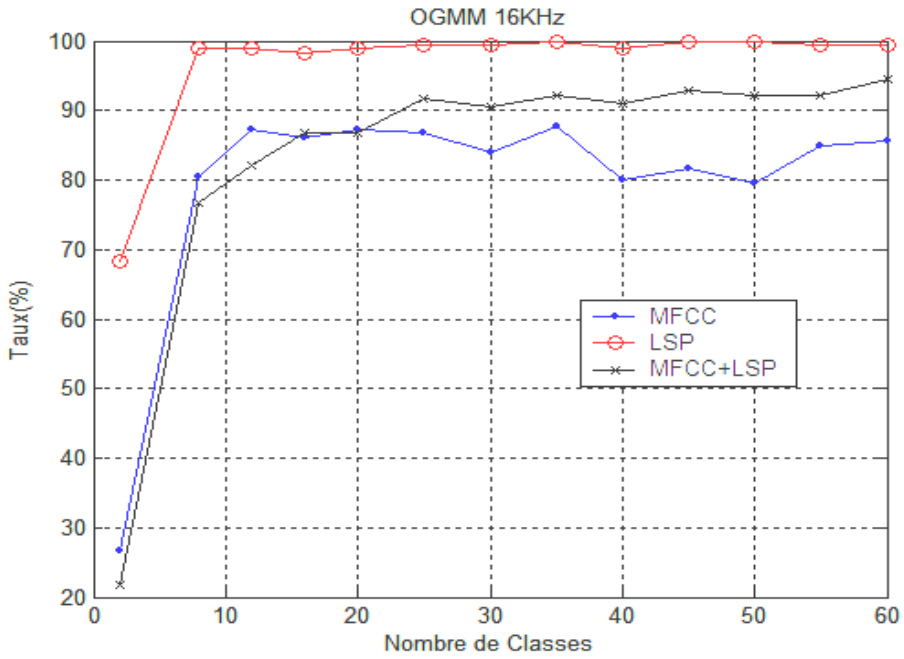

Fig. 4.9 OGMM à Fe= 16 kHz

L'allure obtenue en utilisant les vecteurs MFCC plus LSP, est croissante avec le nombre de gaussiennes, la valeur maximale est de 94% pour 60 gaussiennes.

En utilisant les vecteurs MFCC, le taux d'identification est toujours inférieur à 90%, il est de 88% pour un nombre de gaussienne varie entre 12 et 35, et se dégrade légèrement (80%) au-delà de 35 gaussiennes.

Les meilleures valeurs de taux d'identification correcte sont obtenues en utilisant les vecteurs des LSP, un nombre de 8 gaussiennes est suffisant pour trouver un maximum local d'une valeur de 99%.

On constate que l'orthogonal GMM est une approche intéressante, le temps de calcul est moins important par rapport au GMM, le nombre de gaussiennes nécessaires est de 8 en utilisant les vecteurs LSP, le taux d'identification correcte est de 99%.

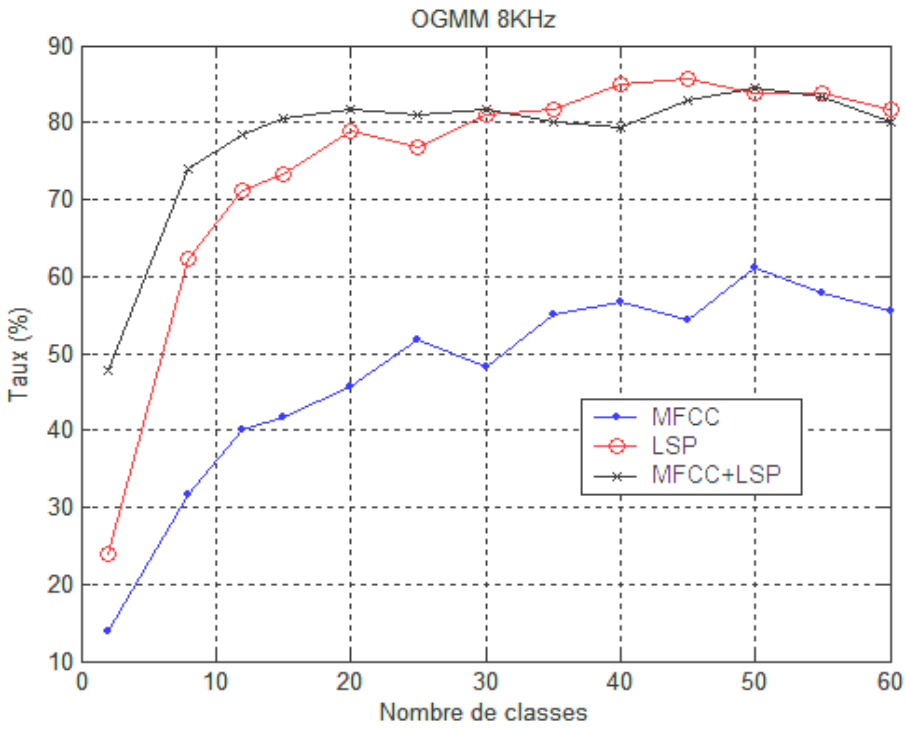

Fig. 4.10 OGMM à Fe= 16 kHz

Avec Fs=8 kHz (Fig. 4.10) on obtient un taux de 86% en utilisant 40 gaussiennes, c'est un nombre relativement grand, qui nécessite un espace mémoire important pour le stockage des références pour chaque locuteur. L'OGMM est plus sensible à la réduction de la fréquence d'échantillonnage que le GMM.

4.7 Conclusion

- La quantification Vectorielle donne de bonnes performances en terme de taux d'identification correcte, mais elle est coûteuse en temps de calcul et espace mémoire nécessaire pour le stockage des références des locuteurs.

- Pour une fréquence d'échantillonnage de 16 kHz, 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, en particulier en utilisant les vecteurs LSP. On remarque que pour les mêmes performances, l'OGMM utilise un nombre de gaussiennes réduit (8 Gaussiennes), ce qui permet une réduction considérable de temps de calcul.

- Pour une fréquence de 8 kHz, les taux de l'OGMM se dégradent, mais en utilisant les LSP les résultats sont toujours meilleurs, le nombre de gaussiennes nécessaires est relativement grand (environ 40 gaussiennes).

- D'après les expériences effectuées, les vecteurs LSP permettent une bonne discrimination entre les locuteurs, les performances obtenues en utilisant les LSP sont toujours les meilleures.

Chapitre 5
Utilisation du Pitch.

5.1 Introduction

La fréquence de vibration des cordes vocales est appelée fréquence fondamentale (ou pitch) est un paramètre très important pour caractériser le locuteur, l'oreille est en effet très sensible à ses variations, lesquelles constituent un élément essentiel de la prosodie.
Une caractéristique très importante du pitch est sa robustesse au bruit. Différentes grandeurs liées au pitch, telles que sa valeur, sa moyenne, son contour, jitter et l'histogramme sont proposés par les chercheurs pour la reconnaissance de locuteurs [34][35][36].

5.2 Motivations

La plupart des systèmes utilisent des traits distinctifs caractérisants le conduit vocal, mais la contribution de la glotte à ces traits est en grande partie ignorée. Même si les paramètres cepstraux (MFCC) possèdent la propriété de déconvoluer entre la glotte et le conduit vocal; mais dans la pratique, ces coefficients cepstraux sont affectés par les voix aigues (femmes et enfants).
On peut illustrer le rôle du pitch quand la dépendance de la source et le conduit vocal est maintenue. Sur figures 5.1 et 5.2 il est montré quatre spectrogrammes et histogrammes du pitch; chaque colonne correspond à une locutrice différente. Toutes les locutrices ont prononcé le même mot, « zéro ». Les spectrogrammes montrent une ressemblance considérable de distributions des formants entre les locutrices. La distribution spatiale des formants dépend de la variabilité inters locuteurs. Cependant, les histogrammes du pitch sont différents et varient d'un locuteur à un autre pour le même contexte. Si on compare les histogrammes en prenant en considération leur amplitude de la fréquence fondamentale et sa largeur, il est observé que la locutrice 2 et 3 ont une distribution du pitch semblable.

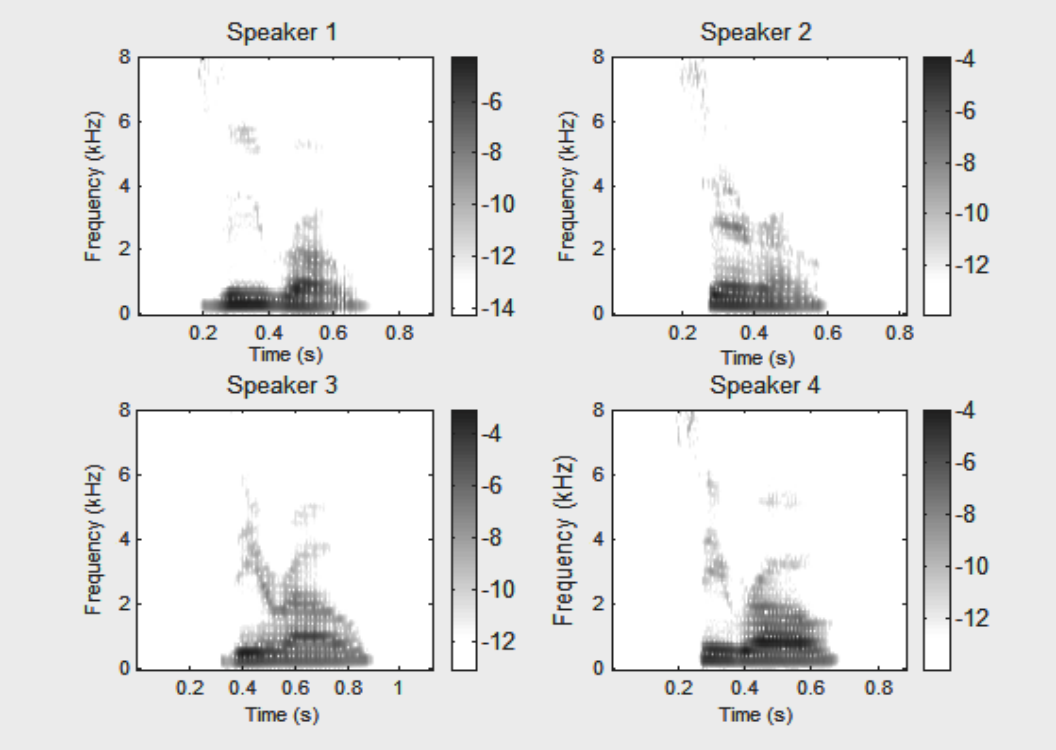

Fig 5.1 Spectrogrammes de 4 locuteurs

D'autre côté, les locutrices 1 et 3 est caractérisé par les histogrammes du pitch dissemblables. Par conséquent, si on prend en considération les informations du pitch, la variabilité inter-locuteurs peut être restreinte aux locuteurs possédants une distributions du pitch semblables, et les autres locuteurs seront considérés comme appartiennent aux autres groupes. Les locuteurs avec pitch semblable seront reconnus en se basant sur leurs caractéristiques spectrales.

En résumé, la fréquence fondamentale de courte durée et les traits vocaux peuvent être exploités conjointement pour établir un modèle de probabilité de vecteurs des traits qui suppose la connaissance a priori sur la distribution du pitch.

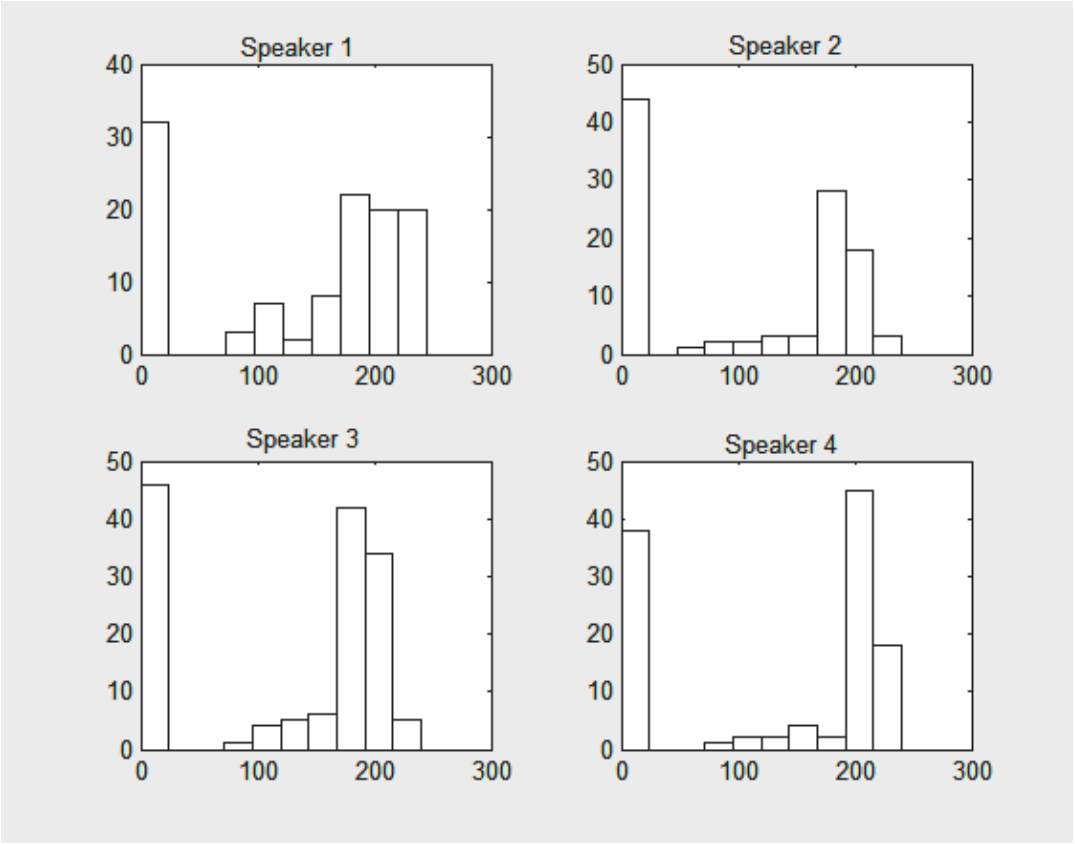

Fig 5.2 Histogrammes des 4 locuteurs

5.3 Le modèle proposé

5.3.1 La théorie

Nous supposons que le pitch et les traits vocaux sont deux processus aléatoires notés respectivement $X(t)$ et $Y(t)$. $\hat{X}(n)$ (Resp. $\hat{Y}(n)$) est le pitch discret estimé à l'instant $n\Delta t$ (resp. le vecteur des paramètres vocaux discret estimé à l'instant $n\Delta t$). $Y(n)$ est un vecteur de dimension D. En pratique, $Y(n)$ est un vecteur des LPC, LSP ou MFCC extraits a partir d'une fenêtre centrée sur $n\Delta t$. Nous supposons que l'ensemble des réalisations de $X(n)$ et $Y(n)$ sont indépendants de temps. En conséquence, $\hat{X}(n+1)$ (resp. $\hat{Y}(n+1)$) est supposée indépendante de la réalisation de $\hat{X}(n)$ (resp. $\hat{Y}(n)$). Nous conservons la crosscorrélation entre $\hat{X}(n)$ et $\hat{Y}(n)$.

Soit $\{x_1, x_2, ..., x_n\}$, l'ensemble des réalisations croissantes de $\hat{X}$, avec $x_i \in [40Hz, 700Hz]$.

Pour simplifier, on va supposer que l'ensemble des réalisation de $\hat{Y}$ est fini (en utilisant le codebook de la Quantification vectorielle par exemple), soit $\{y_1, y_2, ..., y_m\}$, avec $y_i \in \Re^D$.

Soit f est la probabilité conjointe de $\hat{X}$ et $\hat{Y}$.

$$f(x_i, y_j) = P(\hat{X} = x_i, \hat{Y} = y_j) \tag{5.1}$$

Avec $0 \le f(x_i, y_j) \le 1$ et $\sum_{i=1}^{n}\sum_{j=1}^{m} f(x_i, y_j) = 1$

Les fonctions de probabilité marginales sont respectivement :

$$f(x_i) = \sum_{j=1}^{m} f(x_i, y_j) \text{ et } f(y_j) = \sum_{i=1}^{n} f(x_i, y_j) \tag{5.2}$$

Chaque locuteur s est supposé défini par sa fonction de probabilité f_s, qui tient en compte l'accouplement entre X et Y :

$$f_s(x_i, y_j) = P_s(\hat{X} = x_i, \hat{Y} = y_j). \tag{5.3}$$

On observe que :

$$f_s(x_i, y_j) = f_s(y_j / x_i) f_s(x_i). \tag{5.4}$$

$f_s(x_i)$ est la probabilité a priori que la fréquence du pitch soit égal à x_i et $f_s(y_j / x_i)$ la probabilité a posteriori que l'observation du vecteur de paramètre soit égale à y_j sachant que la valeur du pitch est égale à x_i. L'estimation de la probabilité a priori du pitch relativement simple, l'estimation de $f_s(y_j / x_i)$ peut être longue.

5.3.2 La distribution des vecteurs de paramètres basée sur la connaissance du pitch

Dans ce travail on vise l'estimation et l'intégration de la proabilité à posteriori, $f_s(y_j / x_i)$ dans le système de l'identification de locuteur. La considération de $f_s(x_i)$ est laissé en perspective.

Nous proposons de subdiviser l'espace (x, y) en petites sous espaces H_k où $f_s(y_j / x_i)$ supposée localement indépendante du pitch. On défini $I_k, k = 1,2,..., N$, comme les sous intervalles des réalisations $\{x_1, x_2, ..., x_n\}$. On rappelle que $x_1 = 40\,Hz$ et $x_n = 700\,Hz$, N est le nombre des intervalles I_k avec $I_1 \cup \cdots \cup I_N = \{x_1, x_2, ..., x_n\}$. Chaque sous-espace H_k est associé avec un intervalle de pitch I_k. Pour chaque H_k, on suppose que la fonction $f_s(y_j / x_i)$ est stationnaire et indépendante de pitch à l'intérieur de I_k (c.a.d le pitch est supposé le même sur cet intervalle).

Donc $f_s(y_j / x_i) = P(\hat{Y} = y_j / I_k, Locuteur = s \; avec \; x_i \in I_k)$.

Théoriquement, le nombre des modèles $f_s(y_j / x_i) = \lambda_{s,k}$ vaut n. Par la subdivision de l'espace en N sous espaces, le nombre de modèles est réduit à N. La figure suivante illustre la notion de sous espaces et le modèle de la fonction de probabilité $f_s(y_j / x_i)$. L'intervalle de I_k est basé sur les histogrammes du pitch.

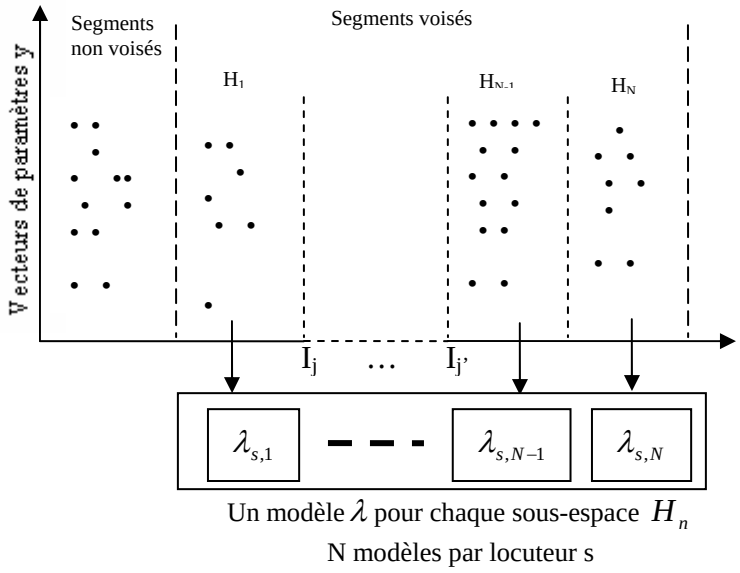

Fig 5.3 L'approche proposée pour générer des sous-modèles

5.4 Le modèle d'identification

On utilise Le modèle GMM avec M gaussiennes. Chaque GMM est défini pour un locuteur spécifique s et pour un intervalle du pitch I_k. Soit $p(y/\lambda_{s,k})$, la densité de m :élanges gaussiennes associée avec la fonction de probabilité $f_s(y_j / I_k)$ pour le locuteur s :

$$p(y/\lambda_{s,k}) = \sum_{m=1}^{M} \pi_{m,k}^{s} . b_{m,k}^{s}(y) \tag{5.5}$$

$$\text{Avec } b_{m,k}^{s}(y) = \frac{1}{(2\pi)^{D/2} \left| \sum_{m,k}^{s} \right|^{1/2}} e^{-\frac{1}{2}(y-\mu_{m,k}^{s})'(\Sigma_{i,k}^{s})^{-1}(y-\mu_{m,k}^{s})} \tag{5.6}$$

M est l'ordre du modèle GMM, y est le vecteur de paramètre vocaux de dimension D (MFCC), $b_{m,k}^{s}$ est la $i^{éme}$ densité gaussienne avec la moyenne $\mu_{m,k}^{s}$ et la matrice de covariance $\sum_{,k}$, les $\pi_{m,k}^{s}$ sont les poids des gaussiennes. $b_{m,k}^{s}, \mu_{m,k}^{s}, \sum_{m,k}^{s}$ et $\omega_{m,k}^{s}$ sont définis pour l'intervalle du pitch I_k et pour le locuteur s. Chaque locuteur est caractérisé par N modèles $\lambda_{s,k}$ correspondant aux intervalles du pitch I_k. $\lambda_{s,k} = \left\{ \pi_{m,k}^{s}, \mu_{m,k}^{s}, \sum_{m,k}^{s} \right\}$ $m = 1...M, k = 1...K$.

5.5 La reconnaissance

Pendant l'apprentissage on estime un modèle par locuteur, le signal parole est découpé en trames de $20\,ms$ chacune, avec décalage de fenêtre de $10\,ms$.

Pour la reconnaissance, deux stratégies sont évaluées, la première basée sur les segments voisés, et la deuxième basée sur l'estimation de la probabilité à posteriori.

5.5.1 Reconnaissance basée sur les segments voisés

On a inclus, dans la modélisation GMM classique, un nouveau module qui estime le pitch et donne si la trame en cours présente un voisement ou non, avec ce module les zones de silences et les segments non voisés seront exclus. Pendant l'apprentissage (l'entraînement du système), une fenêtre de Hamming est appliquée sur des trames de *32 ms*, puis un vecteur des paramètres MFCC est extrait.

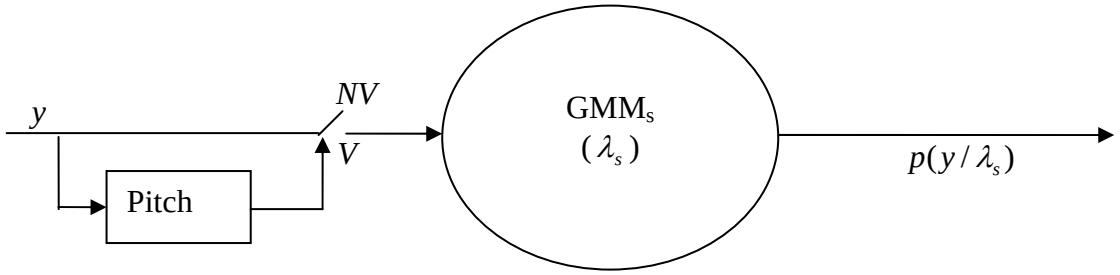

Fig 5.4 Modèle de reconnaissance basée sur les segments voisés

5.5.2 Reconnaissance basée sur l'estimation de la probabilité a posteriori

D'après les histogrammes du pitch, nous avons remarqué que plus de 90% des valeurs du pitch intervienent à l'intervalle $[100Hz, 250Hz]$. Nous avons divisé l'espace $I = (x, y)$ en quatre sous espaces :

$I_1 = [100,150]Hz$, $I_2 = [150,200]Hz$, $I_3 = [200,250]Hz$ et $I_4 = [40,100] \cup [250,700]Hz$

Ce choix de quatre intervalles est conduit par une étude statistique sur la répartition du pitch des locuteurs de notre base de données, on dit que l'identification de locuteur se fait sur bande étroite, ou ce qu'on appelle l'effet de loupe. En conséquence chaque locuteur est représenté par quatre mixtures gaussiennes (GMM).

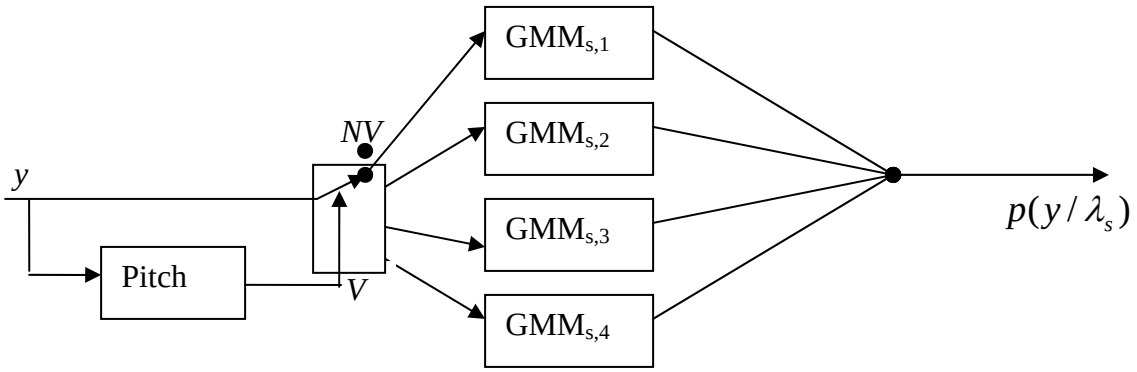

Fig. 5.5 Reconnaissance basée sur l'estimation de la probabilité a posteriori

5.6 Evaluation expérimentale

La base de données de 60 locuteurs est échantillonnée à 16 KHz. on va étudier l'effet de nombre de gaussiennes et la période d'apprentissage sur le taux d'identification pour les deux stratégies appliquées sur l'OGMM.

5.6.1 La première stratégie (Reconnaissance basée sur les segments voisés)

La figure 5.6 trace le taux d'identification en fonction de nombre de gaussiennes utilisées, le taux est toujours croissant avec le nombre de gaussiennes, on voit que la reconnaissance avec un vecteur de paramètres LSP est meilleure que celle obtenue avec les vecteur des MFCC, il y'a une différence de 30% avec 2 gaussiennes. En moyenne les performances du système d'identification sont augmentées en 2% par rapport au GMM classique.

On dit donc que la majorité des traits distinctifs entre les locuteurs se trouve sur les tranches voisées de la parole, ce qui explique les résultats encourageants obtenus en considérant des informations sur le pitch (le voisement).

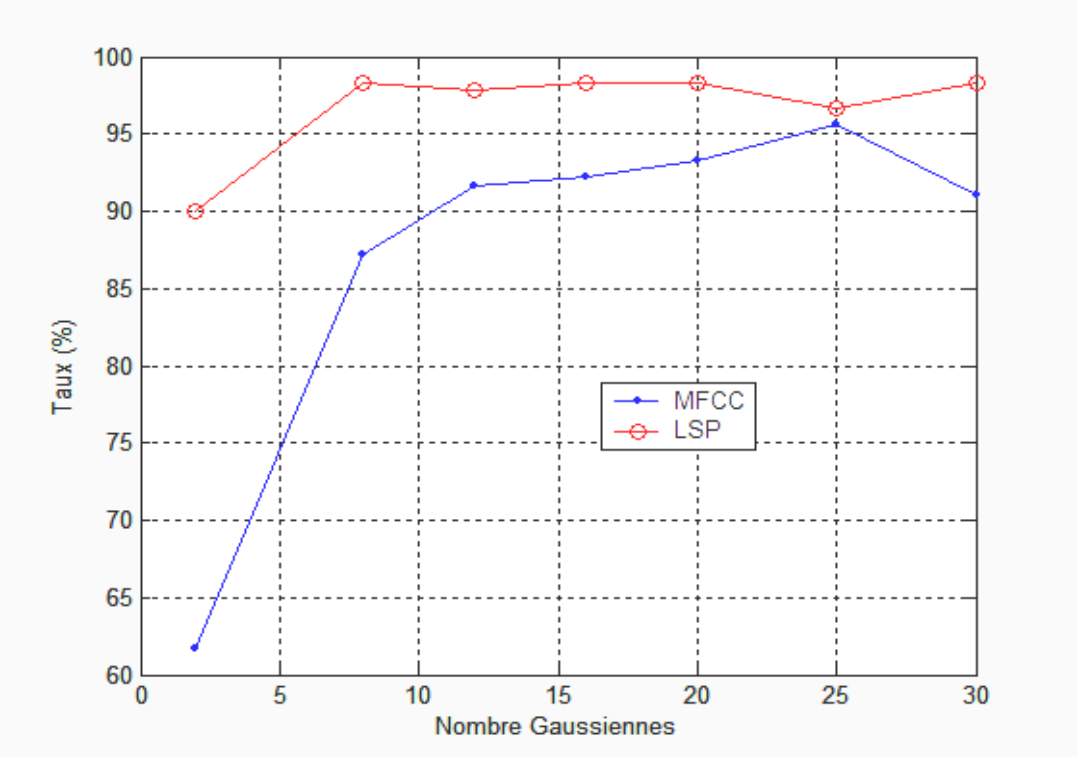

Fig. 5.6 Taux d'identification basée sur les segments voisés

5.6.2 La deuxième stratégie (Estimation de la probabilité a posteriori)

Le signal parole est découpé en trames de 20 ms chacune avec un chevauchement de 10ms entre elles. La moyenne du pitch est calculée pour chaque trame, les trames possédants des valeurs de moyennes du pitch appartiennent au même intervalle $I_i\, i = 1...4$ sont considérées de même groupe et vont être utilisées pour entraîner le même modèle GMM. On aura quatre GMMs par locuteur, et pour une base de donnée de N locuteurs, elle est modélisée par 4N modèle GMMs.

Sur la figure 5.7, on voit que le taux d'identification se dégrade au delà de 16 gaussiennes, un nombre de 8 à 16 gaussiennes est pratiquement suffisant pour représenter les locuteurs, cette foi l'espace des paramètres est partitionné en sous espaces dans lesquels la valeur du pitch est considéré constante, donc le nombre de gaussiennes qui modélisent ces petits espace est forcément plut petit que le nombre de gaussiennes nécessaires pour caractériser l'espace global.

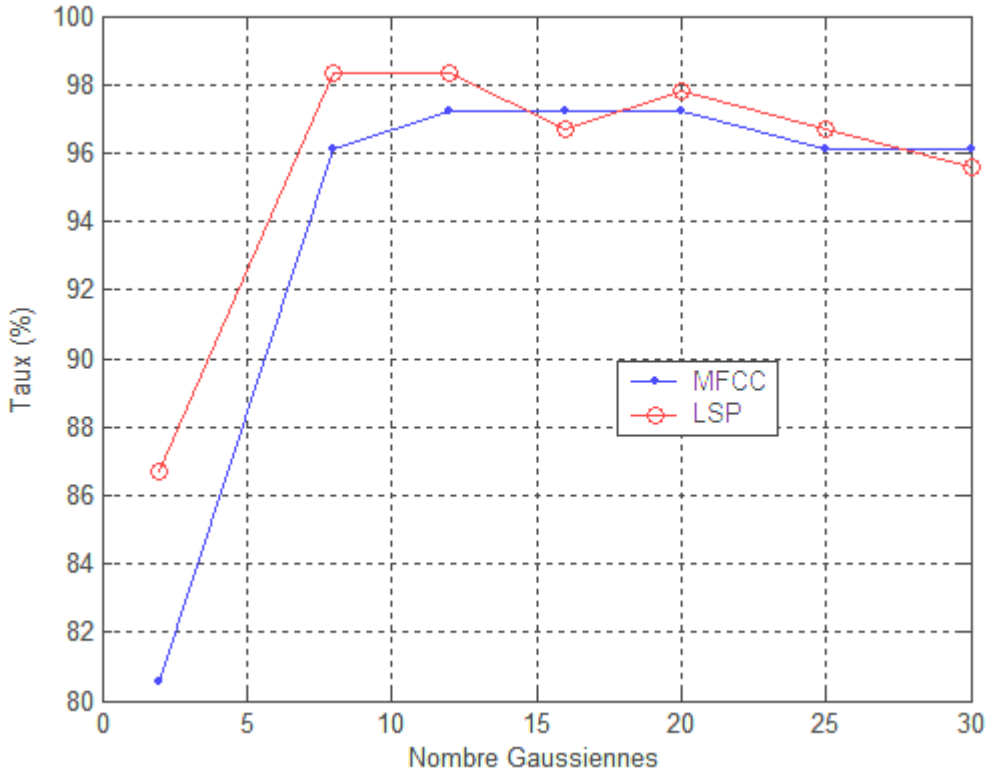

Fig. 5.7 Taux d'identification basée sur la probabilité a posteriori.

5.7 Comparaison entre le système classique et le nouveau système

En pratique il est souhaitable que le système d'identification de locuteurs garde ses performance en augmentant le nombre de locuteurs dans la base de données, ce point constitue un inconvénient majeur pour les systèmes qui utilisent le GMM avec les vecteurs MFCCs sans tenir compte des informations sur la source, en intégrant le pitch (basant sur l'estimation de la probabilité a posteriori) on peut remédier partiellement à ce problème. Dans ce paragraphe on va comparer entre le système classique et le nouveau système basé sur l'estimation de la probabilité a posteriori du pitch développé dans les paragraphes précédents.

On utilise une autre base de données extraite de TIMIT, contient 180 locuteurs, échantillonnée à 16 kHz, les mêmes techniques d'analyse expliquées précédemment sont conservées. Comme abréviation, le système classique qui utilise la modélisation GMM avec les vecteurs de paramètres MFCCs est noté GMM (MFCC), le nouveau système qui se base sur l'estimation de la probabilité a posteriori du pitch est noté GMM-Pitch (MFCC) (figure 5.8).

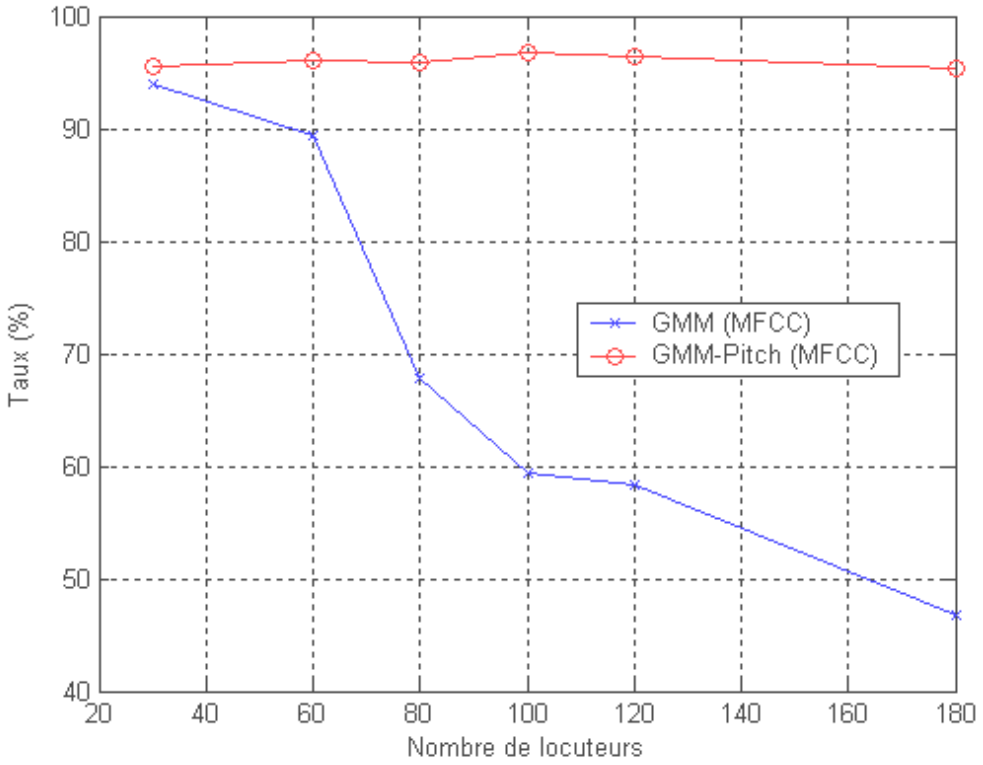

Fig. 5.8 Comparaison entre GMM(MFCC) et GMM-Pitch(MFCC).

La figure 5.8 trace le taux d'identification en fonction de nombre de locuteurs (de 30 à 180), il est clair sur la figure que le taux de GMM(MFCC) décroît rapidement en augmentant le nombre de locuteurs dans la base de données, par contre le nouveau système garde presque le même taux (≈ 95%) même avec 180 locuteurs, ce qui justifie la grande utilité de ne pas ignorer les informations sur la source d'excitation dans des applications où les locuteurs utilisent le système d'identification sont nombreux. Une redondance acceptable d'identification implique l'utilisation de plus d'informations caractéristiques de l'appareil phonatoire.

5.8 Conclusion

Le système classique utilise la modélisation GMM et les vecteurs de paramètre MFCC, est limité par le volume réduit de la base de données (nombre de locuteurs), avec une bonne estimation du pitch on peut réduire l'effet de cette limitation, dans ce travail on a développé et mis en œuvre une méthodologie pour tenir en compte la valeur du pitch, le système proposé est simulé sur MATLAB.

Mathématiquement chaque locuteur est supposé défini par sa fonction de probabilité conjointe f_s qui tien en compte le couplage entre le pitch et les caractéristiques du conduit vocal, cette fonction est le produit de la fonction de probabilité a priori et la fonction de la probabilité a posteriori, au cours de ce travail le nouveau système proposé est basé principalement sur l'estimation de la probabilité a posteriori de pitch, en perspective on estime que l'utilisation de la probabilité a priori du pitch donne encore des meilleures performances.

Pratiquement, des autres contraintes et difficultés apparaissent au cours de l'implémentation de n'importe quelle technique, donc il est de grande importance de viser une cible et implémenter toutes techniques.

Conclusion générale

Au cours de ce travail, nous avons traité le problème de l'identification de locuteur indépendante du texte. Elle consiste à extraire des vecteurs de paramètres a partir des signaux de paroles prononcés par les locuteurs concernés, qui servent à l'entraînement (l'apprentissage) des modèles mathématiques caractérisants la voix de chaque locuteur. Nous avons évalué l'utilisation des paramètres MFCC et LSP, et la modélisation par GMM et OGMM pour l'identification du locuteur, aussi nous avons contribué à l'intégration du pitch dans le système classique à base de GMM.

Ces travaux ont permis de se rendre compte que les vecteurs LSP sont utilisables pour l'identification de locuteur, leur capacité de discrimination est importante en modélisant le locuteur par un mélange de gaussiennes. Aussi l'utilisation du pitch permet d'augmenter le nombre de locuteur de la base de donnée. En intégrant le pitch on peut toujours caractériser l'espace propre à chaque locuteur. Nous avons montré que chaque locuteur est supposé défini par sa fonction de probabilité conjointe f_s qui tien en compte l'accouplement entre le pitch et les caractéristiques du conduit vocal, cette fonction est le produit de la fonction de probabilité a priori et la fonction de la probabilité a posteriori, au cours de ce travail le nouveau système proposé est basé principalement sur l'estimation de la probabilité a posteriori de pitch, en perspective on estime que l'utilisation de la probabilité a priori du pitch donne encore des meilleures performances.

Plusieurs points peuvent faire l'objet d'améliorations notables, notamment pour l'approche de séparer entre la source d'excitation et le conduit vocal et trouver des fonctions caractéristiques, et considérer plus d'informations sur la source d'excitation.

Pratiquement, des autres contraintes et difficultés apparaissent au cours de l'implémentation de n'importe quelle technique, donc il est de grande importance de concrétiser toutes ces modélisations sur une maquette électronique.

BIBLIOGRAPHIE

[1] Yves GRENIER, Thèse de docteur-ingénieur « Identification du Locuteur et Adaptation au Locuteur d'un système de reconnaissance phonémique » Octobre 1977 ENST -E- 77005.

[2] J.J. WOLF, « Effecient acoustic parameters for speaker recognition ». JASA – Vol 51 – part 2 – Juin 72.

[3] René Boite, Hervé Bourlard, Thierry Dutoit, Joël Hancq et Henri Leich : « Traitement de la parole » – Presses Polytechniques et universitaires Romandes 2000.

[4] Bouchefra Khelifa, Thèse de magister à l'ENP : « Contribution à la reconnaissance automatique de la parole continue : Etude et réalisation d'un système de reconnaissance acoustico-phonétique », 1995.

[5] J.Makhoul, « Linear prediction: A tutorial review » *Proc, IEEE,* vol. 63, pp. 561-580. April 1975.

[6] JOSEPH P. CAMPBELL, « Speaker Recognition : A Tutorial» *Proc. IEEE*, Vol. 85, NO. 9, September 1997.

[7] H.HADJ-ALI & M.BOUCHAMEKH, Projet de fin d'étude à l'ENP « Identification du locuteur indépendante du texte» Département d'Electronique, Juin 2004.

[8] H.TAKHEDMIT & N.AIT SAADI, Projet de fin d'étude à l'ENP « Identification du locuteur en mode indépendant du texte », Dép. d'Electronique. Juin 2005.

[9] R.Boite, M.Kunt, « Traitement de la parole ». Presses polytechniques romandes, Lausanne, 1987.

[10] M.Kunt, « Traitement numérique des signaux ». Presses polytechniques romandes, Lausanne, 1980.

[11] Calliope, « La parole et son traitement automatique ». Edition Masson, Paris, 1989.

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

[13] D.A. REYNOLDS, C.ROSE, « Robust Text – Independent Speaker Identification Using Gaussian Mixture Speaker Models » IEEE transaction on Speech and Audio Processing, Vol. 3, No. 1. January 1995.

[14] D.A REYNOLDS, « An Overview of Automatic Speaker Recognition Technology », IEEE 2002.

[15] D.A REYNOLDS, T.F QUAIERI, and R.B DUNN « Speaker Verification Using Adapted Gaussian Mixture Models » M.I.T. Lincolon Laboratory 2000.

[16] G. SINGH, A. PANDA, S. BHATTACHARYYA, and T. SRIKANTHAN « Vector Quantization Techniques for GMM Based Speaker Verification » ICASSP 2003.

[17] S. FURUI « Recent advances in speaker recognition » Tokyo Institute of Technology – Elseiver Science B.V. 1997.

[18] F. Itakura, « Line Spectrum Representation of Linear Predictive Coefficients of Speech Signals » J. Acoust. Soc. Am, 57, 535(a), s35(A), 1975.

[19] Y. Mami, Thèse de doctorat, « Reconnaissance de locuteurs par localisation dans un espace de locuteurs de référence », Ecole Nationale Supérieure des Télécommunication de Paris, Octobre 2003.

[20] R. D. Zilca & Y. Bistritz « Feature Concatenation for Speaker identification ».

[21] Furui, S. « Cepstral analysis technique for automatic speaker verification », *IEEE Tans Acoustics, Speech and signal Processing.* Vol 29, p 254-272 (1981).

[22] Booth, I, Barlow M, and Watson, B. « Enhancement to DTW and VQ decision algorithms for speaker recognition », Speech Communication 13(3-4), 427-433 (1993).

[23] Matsui, T et Furui S. « Comparison of text-independent speaker recognition methods using VQ Distorsion and Discrete /Continuous HMMs » *International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Volume 2, p157-160, (1992).

[24] Matsui, T and Furui S. « Comparison of text-independent speaker recognition methods using VQ Distorsion and Discrete /Continuous HMMs » *IEEE Transactions on speech and Audio Processing,*, 2(3), p 456-459, (1994).

[25] Savic, M and Gupta, S. K. « Variable parameter Speaker verification System based on Hidden Markov Modeling » *ICASSP,* Volume 1, p281-284 (1990).

[26] Yu, K, Mason, J and Oglesby, J. C « Speaker Recognition using Hidden Markov Models, Dynamic time Warping and Vector Quantization ». *Vision, Image and Signal Processing,* 142(5) p 313-316 (1995).

[27] Rosenberg, A. E. Lee, C.-H and Gokcen, S. « Connected word talker verification using whole word hidden Markov Models » *ICASSP*, volume 1, p 381-384 (1991).

[28] Rissanen, E. L. and Webb J. J. « Speaker identification experiments using HMMs » *ICASSP*, volume 2, p 387-390 (1993).

[29] Bimbot F. Magrin-Chagnalleau I. and Mathan L. « Second-order statistical measures for text-independent speaker identification » *Speech Communication*, 17: 177-192 (1995).

[30] Oglesby J. and Mason J. S. « Speaker recognition with neural classifier » *First IEE International Conference*, p306-309 (1989).

[31] Magrin-Chagnolleau I. Bonastre J.F and Bimbot F. « effect of uterence duration and phonetic content on speaker identification using second-order statistical methods » *Conference on speech Communication and Technology (EUROSPEECH),* vol 1, p337-340 (1995).

[32] Homayounpour M. and Chollet G. « Neural network approaches to speaker verification: Comparison with second order statistic measures » *ICASSP,* vol 1, p353-356 (1995).

[33] Hassan Ezzaidi, Jean Rouat and Douglas O`Shaughnesy « Combining pitch and MFCC for speaker recognition systems » ERMETIS, Université du Québec à Chicoutimi, Québec, Canada, G7H 2B1. INRS-Télécommunications, Université du Québec.

[34] B.S. Atal, « Automatic recognition of speakers from their voices », in *Proc. IEEE*, 1976, vol. 64, pp. 460-475.

[35] Douglas O'Shaughnessy and Hesham Tolba, "Towards a robust/fast continuous speech recognition system using a voiced-unvoiced decision", pp.413-416, ICASSP, 1999.

[36] Kemel Sönmez, Elisabeth Shriberg, Larry Heck, and Mitchel Weintraub, "Modeling dynamic prosodic variation for speaker verification", in *Proc. Of international Conference on Spoken Language Processing*, 1998, pp. 3189-3192.

[37] Mokbel C. and Collin O. « Incremental enrollment of speech recognizers » *ICASSP,* p 453-456 (1999).