RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Ecole **N**ationale **P**olytechnique

Département d'Electronique

PROJET DE FIN D'ETUDES

Thème

IDENTIFICATION DU LOCUTEUR EN MODE INDEPENDANT DU TEXTE

Proposé et dirigé par :

Mr.B.BOUSSEKSOU

Etudié par :

HAMOUDA Cherif

KACI Rabah

Promotion Juin 2007

E.N.P 10 Avenues 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 فهذا 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 مفتاحيه : تشخيص المتكلم , الوسائط الطيفية الخطية LSP,التكميم ألشعاعي, النموذج المتعدد الغوصيات (GMM), (OGMM),وسائط مال سابسالنموذج متعدد الغوصيات العمودية

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.

Remerciements

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

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

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

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

Nous remercions tous ceux, qui de prés ou de loin, nous ont apporté leur contribution pour la réalisation de ce travail.

Dédicaces

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

A nos frères et sœurs

A nos proches

A tous nos amis

Table des matières

Chapitre 1:Généralités sur l'identification du locuteur et la parole

Chapitre 2: Paramétrisation du signal

Chapitre 3: Classification des vecteurs acoustiques

Chapitre: Évaluations expérimentales

Liste des figures

Liste des tableaux

Acronymes

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

AR: **A**uto-**R**égressif.

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

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

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 (**L**ine **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.

PLP : **P**erceptual **L**inear **P**redictive.

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

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

SV: **S**on **V**oisé.

SNV: **S**on **N**on **V**oisé.

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

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

Introduction générale

La reconnaissance automatique du locuteur s'inscrit dans le domaine plus général du traitement de la parole. C'est une tâche particulière de la reconnaissance de forme .Ce domaine regroupe les problèmes relatifs à l'identification et la vérification du locuteur sur la base de l'information contenue dans le signal vocal. Afin de laisser le champ à un large éventail d'applications, nous nous intéressons à l'identification du locuteur en mode indépendant du texte (reconnaître la voix d'un locuteur parmi une population composée de N locuteurs connus indépendamment du texte). Nous nous intéressons plus particulièrement à l'extraction des paramètres distinctifs et à 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 permet de présenter le contexte général de la reconnaissance du locuteur et de comprendre la terminologie de l'identification et de la vérification du locuteur ainsi que les notions générales de la parole.

Dans le deuxième chapitre nous avons présenté les 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.

Dans la troisième chapitre nous avons présenté les techniques de modélisation, où plusieurs approches existent : approche vectorielle, connexioniste, statistique et relative. De cette large gamme d'approches, l'approche statistique demeure la plus utilisée.

Le dernier chapitre est consacré à l'évaluation des différentes modélisations sur notre base de données avec les modélisations VQ, GMM, OGMM, et GMMpitch et cela avec les coefficients MFCC et LPS. Nous comparerons notre base de données à celle de TIMIT. Nous examinerons l'influence d'un certain nombre de paramètres (la qualité de données d'apprentissage et de test, le nombre de cœfficients acoustiques, le nombre de locuteurs et

la quantités de données de test) sur le taux d'identification correcte et sélectionner par la suite l'ensemble des paramètres qui donne les meilleurs performances pour une éventuelle conception d'un système d'identification du locuteur.

Chapitre 1

Généralités sur l’identification du locuteur et la parole

1.1 Introduction

Le but de ce chapitre est de définir les principales techniques employées en biométrie, puis de passer en revue les difficultés rencontrées dans l’identification automatique du locuteur, pour bien appréhender le problème et de comprendre les différents niveaux de complexité et les différents facteurs qui rendent le problème difficile. Enfin, en va présenter les notions générales de la parole utilisée dans le domaine d’identification.

1.2. La biométrie

La Biométrie est définit comme une science qui étudie à l'aide de mathématiques (statistiques, probabilités) les variations biologiques à l'intérieur d'un groupe déterminé. C’est donc une discipline qui s'intéresse à la mesure de caractéristiques physiques d'êtres vivants et à leur traitement statistique.

Les termes "biométrie" et "biométrique" se rapportent donc à des dispositifs destinés à reconnaître des êtres humains à partir de mesures effectuées automatiquement. L’authentification peut concerner le visage, la forme de la main, les empreintes digitales, l’iris, la rétine, la voix, …etc.

Principe de fonctionnement

Le principe des techniques biométriques consiste à

- recueillir l'information à analyser,
- traiter cette information et créer un fichier de référence, puis le mettre en mémoire,
- créer un fichier test à l’image du fichier de référence,
- comparer les deux fichiers et déterminer leur taux de similitude,

- prendre la décision qui s'impose.

1.3. Caractéristiques du signal acoustique de la parole

Le signal acoustique de la parole est un peu particulier, il présente des caractéristiques qui rendent l'interprétation très complexe. En effet, ce signal est très redondant (il véhicule beaucoup d'informations ce qui par ailleurs le rend très résistant aux bruits) il est aussi variable d'un locuteur à un autre (variabilité inter-locuteur), et pour le même locuteur (variabilité intra-locuteur). Nous ajoutons les variabilités dues aux conditions d'enregistrement et de l'environnement.

1.3.1. Variabilité intra locuteur

La variabilité intra locuteur exprime les différences dans le signal produit par une même personne. Cette variation peut résulter de l'état physique ou moral du locuteur. Une maladie des voies respiratoires peut ainsi dégrader la qualité du signal de parole de manière à ce que celui-ci devienne totalement incompréhensible, même pour un être humain. L'humeur ou l'émotion du locuteur peut également influencer son rythme d'élocution, son intonation ou sa phraséologie.

Il existe un autre type de variabilité intra locuteur lié à la phase de production de la parole ou de préparation à la production de parole, due aux phénomènes de coarticulation.

1.3.2. Variabilité interlocuteur

La variabilité interlocuteur est un phénomène majeur en reconnaissance du locuteur. Comme nous venons de le rappeler. Mais un locuteur reste identifiable par le timbre de sa voix, malgré une variabilité qui peut être parfois importante.

La cause principale des différences interlocuteurs est de nature physiologique. La parole est produite par les vibrations des cordes vocales, qui déterminent l'importance et la forme du flux d'air s'échappant des poumons et amplifiées par les organes respiratoires, cette opération génère un son à une fréquence de base, le fondamental. Cette fréquence de base est différente d'un individu à l'autre et plus généralement d'un genre à l'autre ; une voix d'homme est plus grave qu'une voix de femme, la fréquence du fondamental étant plus faible. Ce son est ensuite transformé par l'intermédiaire du conduit vocal, délimité à ses extrémités par le larynx et les lèvres. Cette transformation, par convolution, permet de générer des sons différents. Or le conduit vocal est de forme et de longueur variable selon les individus et, plus généralement, selon le genre et l'âge. Ainsi, le conduit vocal féminin

adulte est, en moyenne, d'une longueur inférieure de 15% à celui d'un conduit vocal masculin adulte. Le conduit vocal d'un enfant en bas âge est bien sûr inférieur en longueur à celui d'un adulte. Les convolutions possibles seront donc différentes et, le fondamental n'étant pas constant, un même phonème pourra avoir des réalisations acoustiques très différentes.

La variabilité interlocuteur trouve également son origine dans les différences de prononciation qui existent au sein d'une même langue et qui constituent les accents régionaux.

1.3.3. Variabilité due à l'environnement

La variabilité liée à l'environnement peut, parfois, être considérée comme une variabilité intra locuteur mais les distorsions provoquées dans le signal de parole sont communes à toute personne soumise à des conditions particulières. La variabilité due à l'environnement peut également provoquer une dégradation du signal de parole sans que le locuteur ait modifié son mode d'élocution. Cette variation, peut être considérée comme du bruit.

La variabilité environnementale due au locuteur peut tout d'abord être de nature physiologique. Ainsi, un système mécanique provocant une déformation du conduit vocal provoquera immanquablement une variation dans le signal de parole produit.

1.3.4. Variabilité due aux conditions d'enregistrement

Pour appliquer dans le commerce un système de reconnaissance des locuteurs, il est important de connaître les effets de la transmission téléphonique sur un signal sonore.

La transmission de la parole par un canal téléphonique entraîne une limitation dans la gamme de fréquence, de 300 Hz à 3400 Hz de la bande passante. La caractéristique de transfert n'est pas plate mais change de forme selon la ligne sélectionnée. Les spectres fournis par les lignes téléphoniques sont donc limités par la bande passante et également multipliés par une fonction de transfert de forme inconnue. Dans un premier stade, les études ont montré que la limitation des spectres de, longue durée à la bande passante caractérisant la qualité du téléphone n'affecte pas sensiblement le taux d'identification. Cependant, la pondération des spectres par des fonctions arbitraires du transfert, détruit la fiabilité de l'identification parce que, dans certains cas, l'effet de la fonction de transfert sur les spectres est plus important que les caractéristiques des voix.

1.3.5. Autres problèmes

Les résultats des études de l'effet du codage de la parole utilisé dans le réseau téléphonique mobile GSM sur les performances de vérification du locuteur ont montré une dégradation marginale des performances. Le réseau GSM ne semble donc pas poser de problèmes liés au codage de la parole. Cependant, des problèmes d'un autre ordre se posent dans le cas de la téléphonie mobile pour les applications de reconnaissance, ils sont essentiellement dus au bruit ambiant dont les caractéristiques sont variables au cours du temps, ce qui rend difficile les techniques d'adaptation en ligne.

1.4. La reconnaissance automatique du locuteur (RAL)

La reconnaissance automatique du locuteur s'inscrit dans le domaine plus général de la communication homme-machine (CHM). Il consiste à extraire l'information contenue dans le signal acoustique de la parole et éventuellement de l'interpréter pour connaître automatiquement l'identité d'une personne prononçant une ou plusieurs phrases, à l'aide d'un ordinateur qui joue aujourd'hui un grand rôle dans ce domaine. Les applications directes de la RAL concernent les problèmes de confidentialité et d'authentification. Nous distinguerons

- les applications "sur site" : serrures vocales pour contrôle d'accès, cabines bancaires en libre service,
- les applications liées aux télécommunications : ces applications concernent l'identification du locuteur à travers le réseau téléphonique pour accéder à un service de transactions bancaires à distance ou pour interroger des bases de données en accès privé,
- les applications judiciaires (forensic applications) : recherche de suspects, orientations d'enquêtes.

La difficulté de la tâche de reconnaissance n'est pas la même d'une application à l'autre. Dans le cas des applications 'sur site', l'environnement de prononciation de la phrase ou du mot de passe est plus facilement contrôlé que dans le cas des applications via le réseau téléphonique (distorsions dues au canal, différences entre les combinés téléphoniques, bande passante limitée). Les applications judiciaires présentent quand à elles des difficultés d'un autre ordre (locuteurs non coopératifs, enregistrements de mauvaise qualité).

1.4.1. Identification et vérification du locuteur

On distingue deux tâches différentes en reconnaissance du locuteur : l'identification et la vérification.

L'identification du locuteur consiste à reconnaître le vrai locuteur parmi une population (ou base) composée de N locuteurs connus. L'entrée du système est un enregistrement de parole d'un locuteur inconnu. La sortie du système correspond à l'identité du locuteur de la base de référence qui est le plus "proche" du signal de parole inconnu. Dans cette tâche, nous supposons que le signal de parole à identifier est prononcé par un des locuteurs de la base de référence (identification en ensemble fermé).

La vérification du locuteur consiste à déterminer si le locuteur est bien celui qu'il prétend être. Le système dispose en entrée d'un échantillon de parole et d'une identité proclamée. Une mesure de ressemblance est calculée entre l'échantillon et la référence du locuteur correspondant à l'identité proclamée. Si cette mesure est en dessous d'un certain seuil, le système accepte le locuteur. Dans le cas contraire, le locuteur est considéré comme un imposteur et il est rejeté.

Enfin, l'identification en ensemble ouvert est une combinaison des deux tâches précédentes, Identification du locuteur le plus probable parmi les locuteurs de la base, vérification que l'échantillon inconnu a bien été prononcé par le locuteur choisi dans l'étape d'identification.

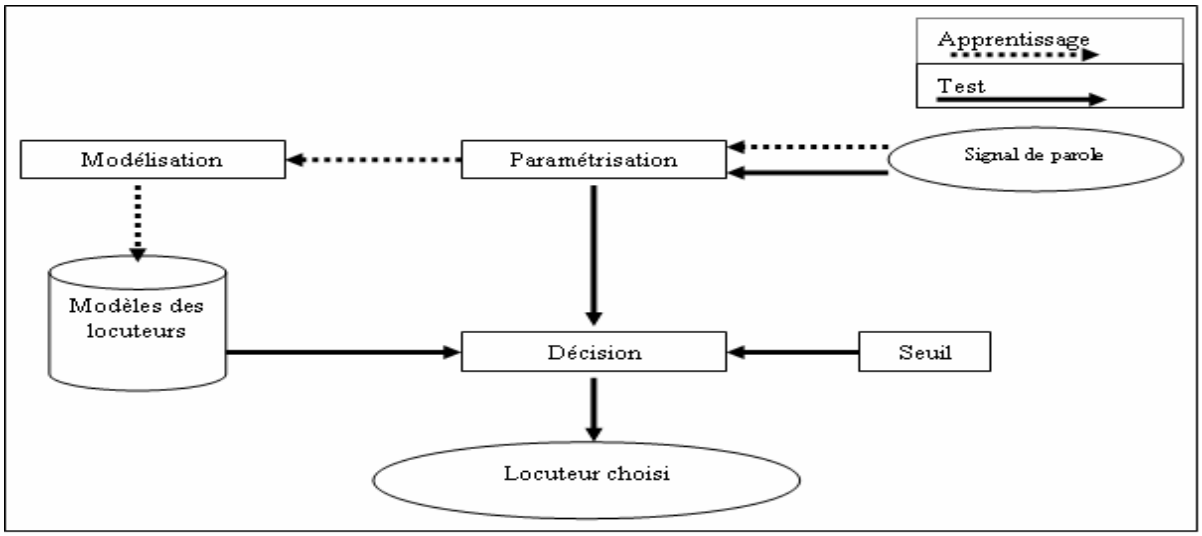


FIG. 1.1 Schéma modulaire d'un système d'IAL

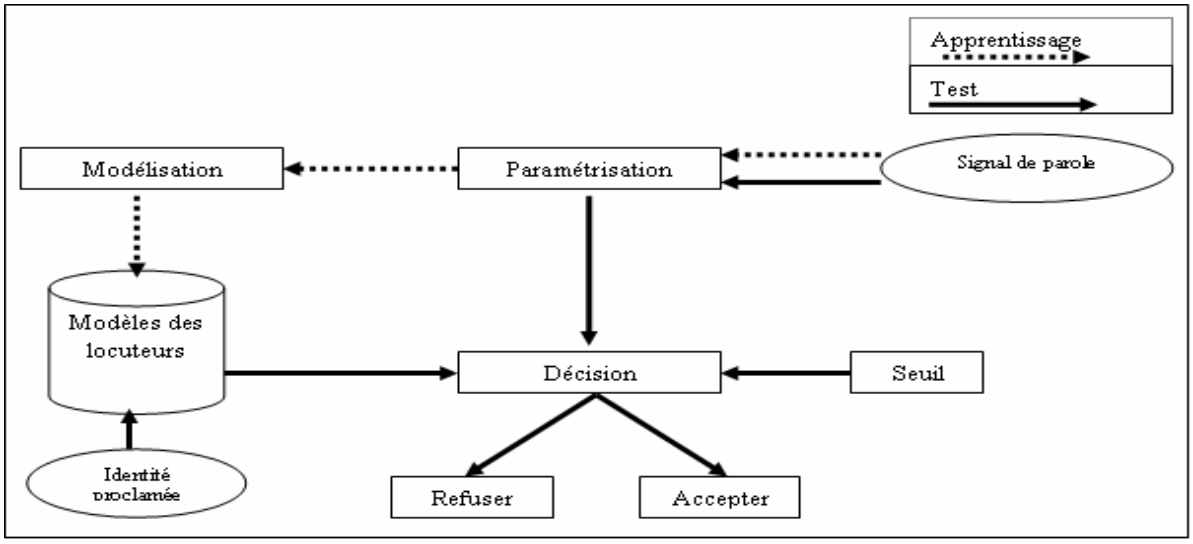


FIG. 1.2 Schéma modulaire d'un système de VAL

1.4.2. Dépendance et indépendance vis à vis du texte

La distinction est faite entre les systèmes dépendants du texte et les systèmes indépendants du texte. En mode dépendant du texte, le texte prononcé par le locuteur pour être reconnu du système doit être le même que celui qu'il a prononcé lors de l'apprentissage de sa voix. En mode indépendant du texte, le locuteur peut prononcer n'importe quelle phrase pour être reconnu.

Néanmoins, il existe plusieurs niveaux de dépendance au texte suivant les applications (listés selon le degré croissant de dépendance au texte)

- systèmes à texte libre (ou free_text) : le locuteur prononce ce qu'il veut,
- systèmes à texte suggéré (ou text_prompted) : un texte, différent à chaque session et pour chaque personne, est imposé au locuteur et affiché à l'écran par la machine. On parle également de systèmes sound_prompted dans le cas où un enregistrement du texte proposé est joué au locuteur,
- systèmes dépendants de traits phonétiques (ou speech event dependent) : certains traits phonétiques spécifiques sont imposés dans le texte que le locuteur doit prononcer,
- systèmes dépendants du vocabulaire (ou vocabulary dépendent) : le locuteur prononce une séquence de mots issus d'un vocabulaire limité (ex. : séquence de digits),

- systèmes personnalisés dépendants du texte (ou user_specific text dependent) : chaque locuteur a son propre mot de passe.

Les systèmes dépendants du texte donnent généralement de meilleures performances de reconnaissance que les systèmes indépendants du texte car la variabilité due au contenu linguistique de la phrase prononcée est alors neutralisée.

1.5. Structure des systèmes en RAL

La tâche de reconnaissance automatique du locuteur peut se subdiviser en trois étapes, qui sont la paramétrisation, la classification et la décision. Un premier module de traitement du signal réalise L'analyse acoustique du signal de parole. A l'issue de cette étape, le signal est représenté par des vecteurs de coefficients, ce qui permet de réduire l'information en quantité et en redondance. Ces vecteurs sont éventuellement représentés par un modèle mathématique, on parle alors de méthodes paramétriques. Dans la phase de classification, les vecteurs du signal de test (ou leur modèle) sont comparés aux vecteurs des locuteurs de référence (ou à leurs modèles). La phase de décision désigne le locuteur finalement reconnu.

La structure d'un système d'identification du locuteur en ensemble fermé (qui constitue le cadre de notre travail) est représentée sur la figure 1.3.

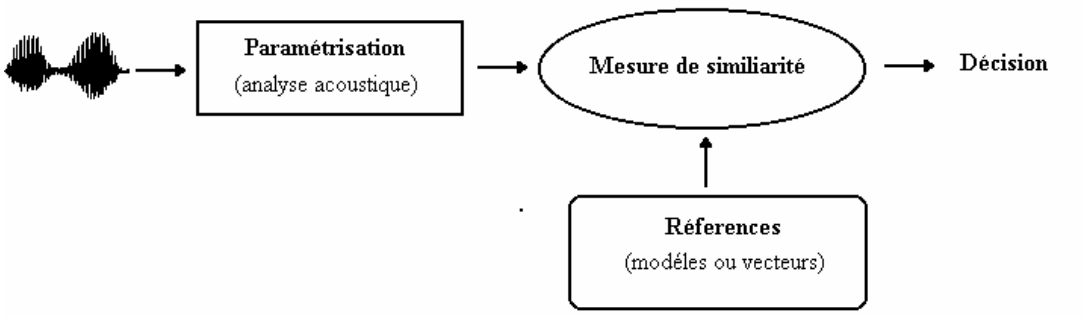


FIG .1.3 Structure d'un système RAL

1.6. Evaluation des performances en RAL

Classiquement, les performances d'identification du locuteur sont données par le taux d'erreur d'identification : pourcentage des cas où le système ne reconnaît pas le bon locuteur. Dans le cas d'un système de vérification du locuteur, on distingue le taux de fausse acceptation : pourcentage des cas où le système accepte le locuteur alors que celui-ci n'est pas la personne qu'il prétend être (locuteur imposteur) ; et le taux de faux rejet :

situation où le système rejette le locuteur alors qu'il est vraiment la personne qu'il prétend être (locuteur honnête).

L'évaluation des performances d'un système de RAL n'est cependant pas un problème trivial et on ne peut comparer deux systèmes à partir de ces seuls taux d'erreur qui dépendent de multiples facteurs. Ainsi, les éléments suivants doivent également être pris en compte :

- qualité de la parole : enregistrements en studio ou via le canal téléphonique ; environnement calme ou bruité ; type de réseau téléphonique,
- quantité de parole : durée de parole pour l'apprentissage des références de chaque locuteur ; durée de parole des sessions de test,
- variabilité intra locuteur : la voix d'un locuteur dépend de son état physique et émotionnel ; de plus, le comportement d'un locuteur se modifie lorsque celui-ci s'habitue à un système,
- population de la base de locuteurs : en identification du locuteur, la taille de la population a une influence directe sur les performances ; la qualité de la population
- (proportion hommes/femmes, bonne répartition géographique des locuteurs parlant une même langue) est également un facteur à intégrer,
- intention des locuteurs : la distinction est faite entre les locuteurs coopératifs (qui veulent être reconnus par le système) et les locuteurs non coopératifs qui modifient leur voix pour ne pas être reconnus (cas de certaines applications judiciaires par exemple).

Enfin, certains locuteurs imitent la voix d'une autre personne pour être reconnus à sa place : ce sont des imposteurs. A ce propos, lors de l'évaluation d'un système, les imposteurs sont en général d'autres locuteurs de la base de référence ce qui n'est pas très réaliste. En effet, en pratique, un imposteur réel tentera d'imiter la voix du locuteur pour lequel il voudra être reconnu.

1.7. La parole

L'information portée par le signal de parole peut être analysée de bien des façons. On en distingue généralement plusieurs niveaux de description non exclusifs : acoustique, phonétique, phonologique, morphologique, syntaxique et sémantique. Dans notre travail de l'identification en s'intéresse par les deux premiers niveaux : acoustique et phonétique.

1.7.1. Le niveau acoustique

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La phonétique acoustique étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone.

De nos jours, le signal électrique résultant est le plus souvent numérisé. L'opération de numérisation, schématisée à la figure 1.4, requiert successivement : un filtrage de garde, un échantillonnage, et une quantification.

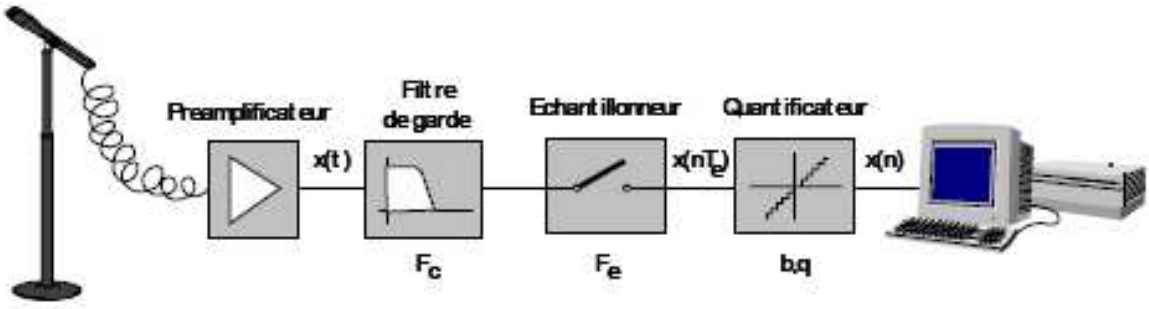


Fig. 1.4 Enregistrement numérique d'un signal acoustique.

La fréquence de coupure du filtre de garde, la fréquence d'échantillonnage, le nombre de bits et le pas de quantification sont respectivement notés fc , fe , b, et q.

Audiogramme

La figure 1.5 : représente l'évolution temporelle, ou audiogramme, du signal vocal pour les mots 'parenthèse', et 'effacer'. On y constate une alternance de zones assez périodiques et de zones bruitées, appelées zones voisées et nonvoisées.

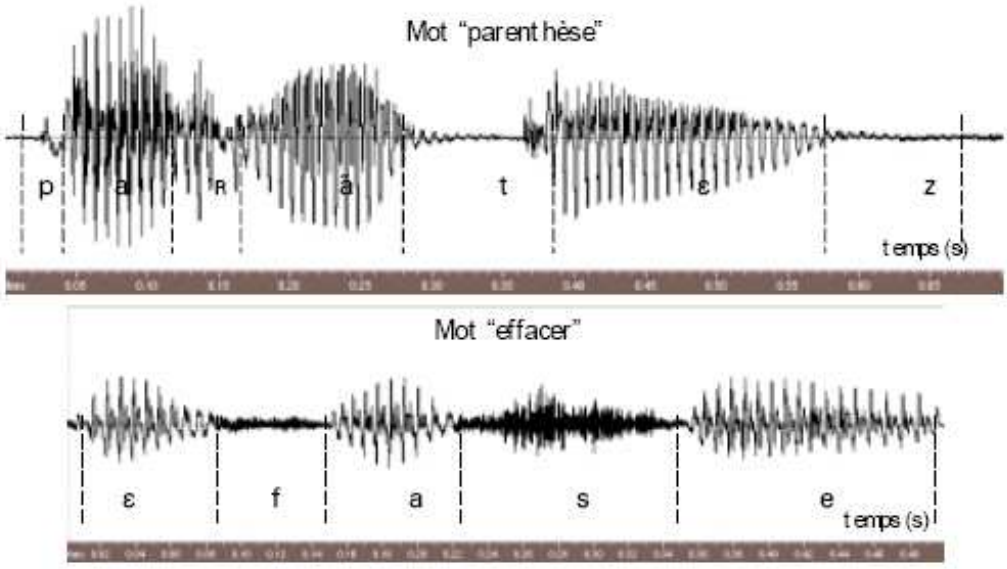


Fig.1.5 Audiogramme de signaux de parole

Transformée de Fourier à court terme

La transformée de Fourier à court terme est obtenue en extrayant de l'audiogramme une trame de 30 ms de signal vocal, en pondérant ces échantillons par une fenêtre de pondération (souvent une fenêtre de Hamming) et en effectuant un transformée de Fourier sur ces échantillons.

La figure (1.6) illustre la transformée de Fourier d'une tranche voisée et celle d'une tranche non-voisée. Les parties voisées (SV) du signal apparaissant sous la forme de successions de pics spectraux marqués, dont les fréquences centrales sont multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé (SNV) ne présente aucune structure particulière. La forme générale de ces spectres, appelée enveloppe spectrale, présente elle-même des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal et sont appelés formants et anti-formants. L'évolution temporelle de leur fréquence centrale et de leur largeur de bande détermine le timbre du son. Il apparaît en pratique que l'enveloppe spectrale des sons voisés est de type passe bas, avec environ un formant par 1kHz de bande passante,

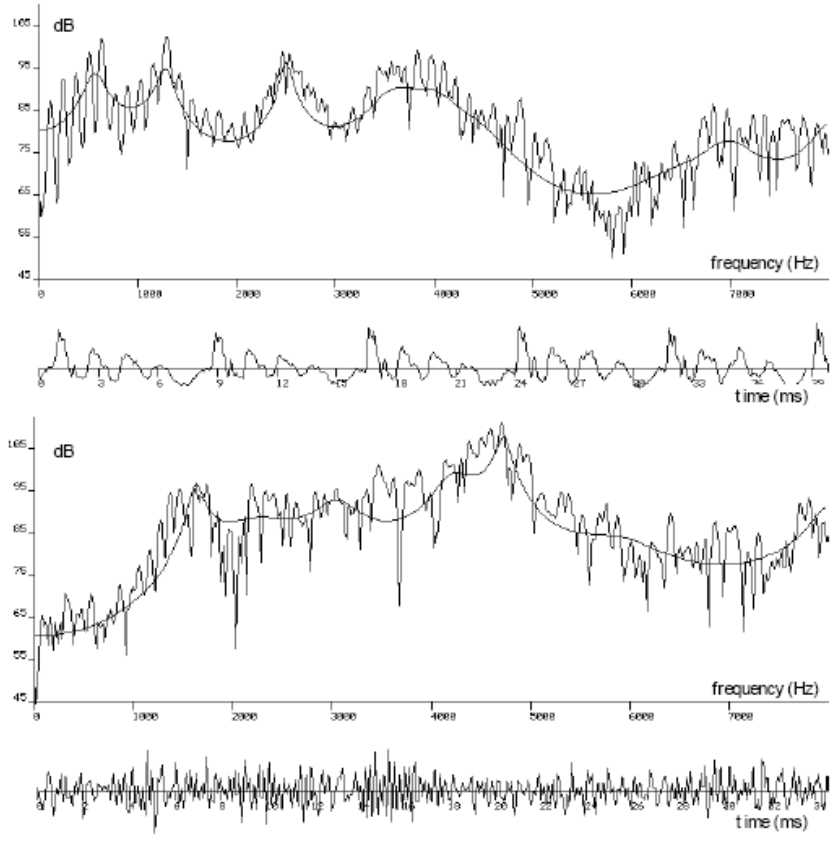


Fig.1.6 Evolution temporelle (en haut) et transformée de Fourier discrète (en bas) du [a] et du [ŝ] de 'baluchon' (signaux pondérés par une fenêtre de Hamming de 30 ms)

1.7.2. Le niveau phonétique

Au contraire des acousticiens, ce n'est pas tant le signal qui intéresse les phonéticiens que la façon dont il est produit par le système articulatoire, présenté à la figure 1.7, et perçu par le système auditif.

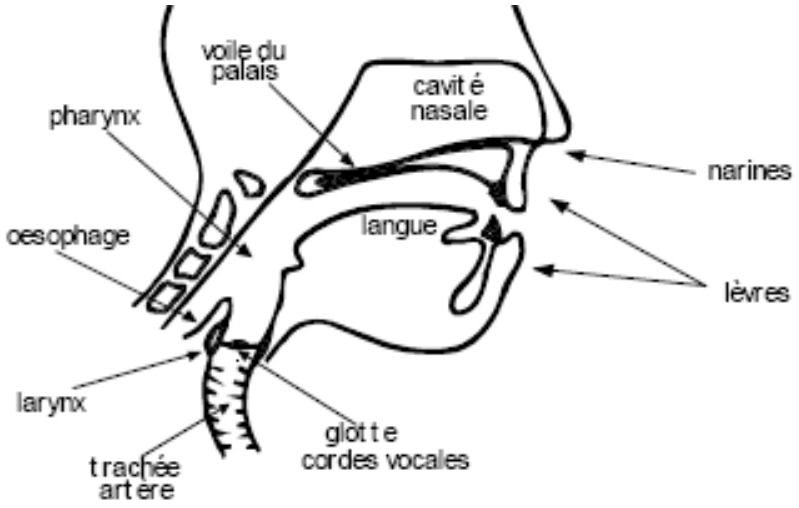


Fig.1.7 L'appareil phonatoire

Phonation

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par les sensations kinesthésiques.

L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le larynx ou la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée (Fig.1.8). Les cordes vocales sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire appelée glotte. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés. Les sons voisés résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales, et les force à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal, composé des cavités pharyngienne et buccale pour la plupart

des sons. Lorsque la luette est en position basse, la cavité nasale vient s'y ajouter en dérivation.

Notons pour terminer le rôle prépondérant de la langue dans le processus phonatoire. Sa hauteur détermine la hauteur du pharynx : plus la langue est basse, plus le pharynx est court. Elle détermine aussi le lieu d'articulation, région de rétrécissement maximal du canal buccal, ainsi que l'aperture, écartement des organes au point d'articulation.

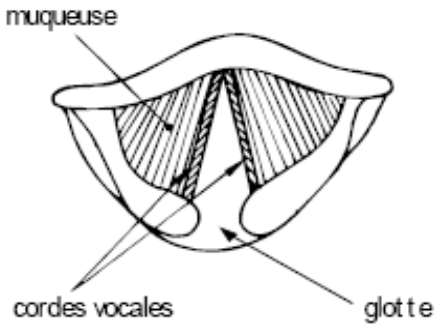


Fig. 1.8 Section du larynx, vu de haut

Audition – perception

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante que la maîtrise des mécanismes de production.

Les ondes sonores sont recueillies par l'appareil auditif, ce qui provoque les sensations auditives. Ces ondes de pression sont analysées dans l'oreille interne qui envoie au cerveau l'influx nerveux qui en résulte; le phénomène physique induit ainsi un phénomène psychique grâce à un mécanisme physiologique complexe.

L'appareil auditif comprend l'oreille externe, l'oreille moyenne, et l'oreille interne (Fig. 1.9). Le conduit auditif relie le pavillon au tympan : c'est un tube acoustique de section uniforme fermé à une extrémité ; son premier mode de résonance est situé vers 3000 Hz, ce qui accroît la sensibilité du système auditif dans cette gamme de fréquences. Le mécanisme de l'oreille interne (marteau, étrier, enclume) permet une adaptation d'impédance entre l'air et le milieu liquide de l'oreille interne. Les vibrations de l'étrier sont transmises au liquide de la cochlée. Celle-ci contient la membrane basilaire qui transforme les vibrations mécaniques en impulsions nerveuses. La membrane s'élargit et s'épaissit au fur et à mesure que l'on se rapproche de l'apex de la cochlée; elle est le support de l'organe de Corti qui est constitué par environ 25000 cellules ciliées raccordées au nerf auditif. La réponse en fréquence du conduit au droit de chaque cellule est esquissée à la figure 1.10. La fréquence de résonance dépend de la position occupée par la cellule sur

la membrane; au-delà de cette fréquence, la fonction de réponse s'atténue très vite. Les fibres nerveuses aboutissent à une région de l'écorce cérébrale appelée aire de projection auditive et située dans le lobe temporal. En cas de lésion de cette aire, on peut observer des troubles auditifs. Les fibres nerveuses auditives afférentes (de l'oreille au cerveau) et efférentes (du cerveau vers l'oreille) sont partiellement croisées : chaque moitié du cerveau est mise en relation avec les deux oreilles internes.

A l'intérieur de son domaine d'audition, l'oreille ne présente pas une sensibilité identique à toutes les fréquences. La figure 1.11 fait apparaître les courbes d'égale impression de puissance auditive (aussi appelée sonie, exprimée en sones) en fonction de la fréquence. Elles révèlent un maximum de sensibilité dans la plage [500 Hz, 10 kHz], en dehors de laquelle les sons doivent être plus intenses pour être perçus.

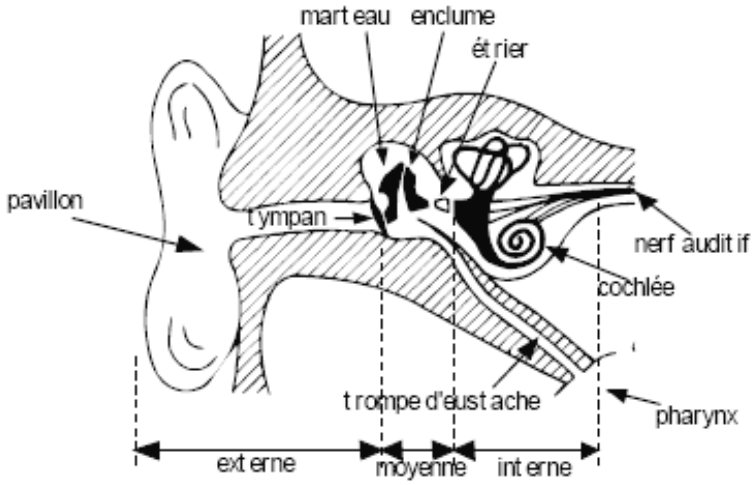


Fig.1.9 le système auditif

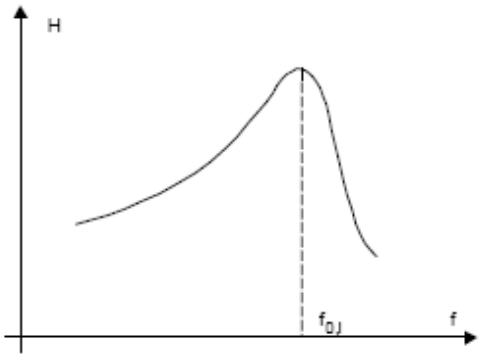


Fig. 1.10 Réponse en fréquence d'une cellule ciliée

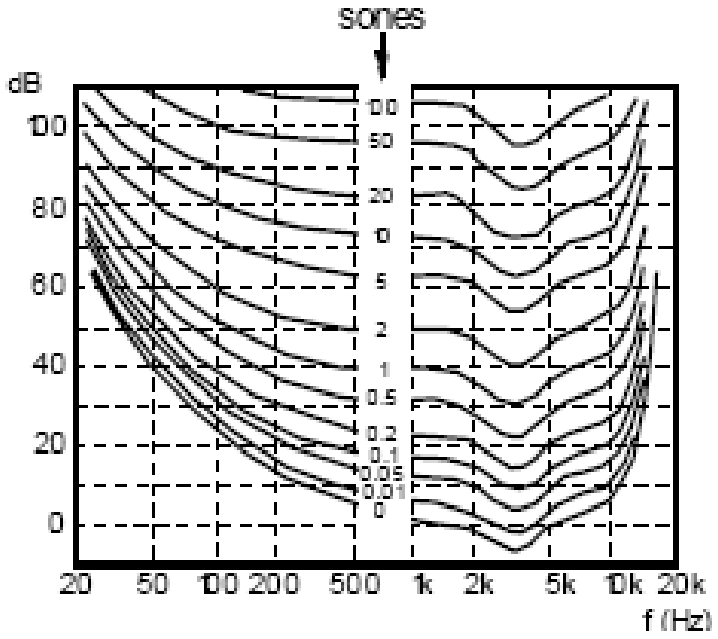


Fig. 1.11 Courbes isosoniques en champ ouvert

1.8 Conclusion

Les moyens biométriques permettent une authentification sûre car ils sont basés sur l'individu lui-même.

Il est alors indispensable de caractériser l'individu par une empreinte afin de le différencier des autres sans aucune ambiguïté, cette empreinte est une clé codant l'identité d'une personne sans redondance ni variabilité. La plupart des indices biométriques, comme les empreintes digitales ou génétiques, répondent à ces critères.

Il en est différemment pour la voix dont la disposition à varier est inscrit dans sa nature même. Si nous voulons vraiment parler d'empreinte vocale, il faut tenir compte du fait que la variabilité interlocuteur est plus importante que la variabilité intralocuteur.

La voix devient donc un indice biométrique intéressant à exploiter car pratique et disponible via le réseau téléphonique, contrairement aux autres indices.

Il est très intéressant de connaître les différents niveaux de description de la parole dans le domaine de l'identification automatique d'un locuteur. Elle s'introduit dans l'extraction de paramètres des locuteurs.

Chapitre 2

Paramétrisation du signal

2.1. Introduction

L'objectif d'un système de paramétrisation est d'extraire les informations caractéristiques du signal de parole en éliminant au maximum les parties redondantes.

Un tel système prend un signal en entrée et retourne un vecteur de paramètres (appelé indifféremment vecteur acoustique ou encore vecteur d'observations). Les vecteurs de paramètres doivent être pertinents (précis, de taille restreinte et sans redondance), discriminants (pour faciliter la reconnaissance) et robustes (aux différents bruits et/ou locuteurs).

Il existe un certain nombre d'approches pour la paramétrisation. Nous présentons ici celles utilisées le plus couramment dans la littérature.

2.2. Paramètres prosodiques

Dans plusieurs domaines d'analyse et de traitement du signal vocal, on définit par paramètres prosodiques

- la fréquence fondamentale (vibration des cordes vocales),
- l'intensité de la voix (ou énergie),
- la durée.

Ces paramètres prosodiques prennent une importance particulière pour donner aux systèmes de synthèse une meilleure intelligibilité tout en permettant aux systèmes de

reconnaissance d'effectuer une analyse ou segmentation par ordre d'unité phonétique. La variation dans le temps de ces paramètres (intonation) véhicule divers indices caractéristiques de l'individu que ce soit au niveau de son état physique (age, sexe, physiologie), de son état émotionnel ou de son accent régional.

2.3. Paramètres de l'analyse spectrale

Les principaux paramètres de l'analyse spectrale utilisés en reconnaissance vocale sont les coefficients de prédiction linéaire, et les paramètres cepstraux.

Plusieurs méthodes permettent d'obtenir des coefficients cepstraux

- grâce à une récursion depuis les coefficients LPC, ce qui donne les coefficients LPCC,
- par l'utilisation d'une FFT et d'une FFT inverse ; cette technique permet de calculer les coefficients MFCC, LFCC et PLP.

2.3.1. MFCC (Mel Frequency Cepstral Coefficient)

Les coefficients cepstraux issus d'une analyse par transformée de Fourier caractérisent bien la forme du spectre et permettent de séparer l'influence de la source glottique de celle du conduit vocal.

Le cepstre du signal de parole est défini comme étant la transformée de Fourier inverse du logarithme de la densité spectrale de puissance. Pour ce signal, la source d'excitation glottique est convoluée avec la réponse impulsionnelle du conduit vocal.

$$s(t) = e(t) * h(t) \tag{2.1}$$

Où $s(t)$ est le signal de parole, $e(t)$ est la source d'excitation glottique et $h(t)$ est la réponse impulsionnelle du conduit vocal.
L'application à l'équation (2.1) du logarithme du module de la transformée de Fourier donne :

$$\log|S(f)| = \log|E(f)| + \log|H(f)| \tag{2.2}$$

Par une transformée de Fourier inverse, on obtient :

$$s'(cef) = e'(cef) + h'(cef) \tag{2.3}$$

La dimension du nouveau domaine est homogène à un temps et s'appelle la *quéfrence* (*cef*), le nouveau domaine s'appelle donc : le domaine *quéfrentiel*. Un filtrage dans ce domaine s'appelle *liftrage.*

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'une simple fenêtre dans le domaine quéfrentiel (liftrage passe-bas) pour le conduit vocal.

Les coefficients cepstraux les plus répandus sont les MFCC (Mel Frequency Cepstral Coefficients). Il présentent l'avantage d'être faiblement corrélés entre eux, et qu'on peut donc approximer leur matrice de covariance par une matrice diagonale.
Pour simuler le fonctionnement du système auditif humain, les fréquences centrales du banc de filtres sont réparties uniformément sur une échelle perceptive. Plus la fréquence centrale d'un filtre est élevée, plus sa bande passante est large. Cela permet d'augmenter la résolution dans les basses fréquences, zone qui contient le plus d'information utile dans le signal de parole. Les échelles perceptives les plus utilisées sont l'échelle Mel et l'échelle Bark.

Echelle Mel

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{2.4}$$

Echelle Bark

$$Bark(f) = 6 \operatorname{Arc\,sinh}\left(\frac{f}{1000}\right) \tag{2.5}$$

f représente la fréquence [Hz].

La procédure de calcul des coefficients MFCC est illustrée sur la figure .2.1

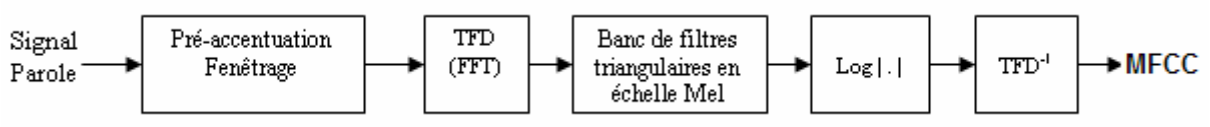


FIG. 2.1 Calcul des coefficients MFCC

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

$$S(k) = \sum_{n=0}^{N-1} s(n)\, exp\left(\frac{-j\,2\pi\, n\, k}{N}\right),\ 0 \le k \le N-1 \qquad (2.6)$$

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 :

$$B_m = B_b + m\,\frac{B_h - B_b}{M+1}\ ,\ 0 \le m \le M+1 \qquad (2.7)$$

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

$$f_m = B^{-1}\left(B_b + m\,\frac{B_h - B_b}{M+1}\right) \qquad (2.8)$$

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

$$B^{-1}(x) = 700\left(10^{\frac{x}{2595}} - 1\right) \qquad (2.9)$$

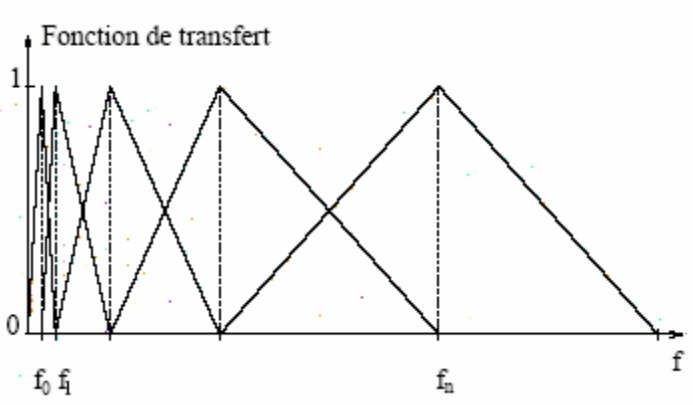


FIG. 2.2 Banc de filtres sur l'échelle Mel

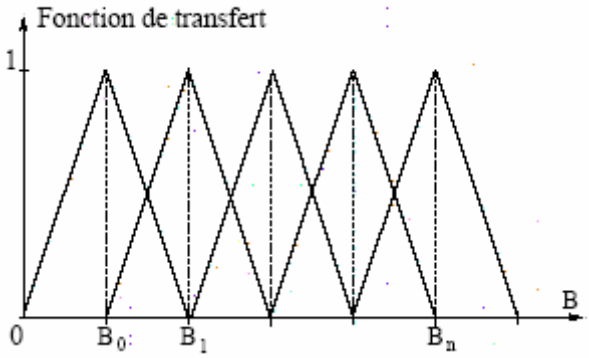


FIG. 2.3 Banc de filtres sur l'échelle linéaire

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

$$c_k = \sum_{i=1}^{M} log\, E_i \; cos\left[\frac{\pi k}{M}\left(i - \frac{1}{2} \right) \right] \quad , 1 \le k \le d \qquad (2.10)$$

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 Paramétrisation par la méthode de prédiction linéaire

Le signal de parole n'étant pas complètement aléatoire, les échantillons successifs sont corrélés. Peut-on utiliser cette corrélation pour réduire la quantité de données ?A la différence des méthodes précédentes, les modèles de production ne cherchent à reproduire que le schéma de principe du mécanisme phonatoire, par le biais de son équivalent électrique. On y décrit la parole comme le signal produit par un assemblage de générateurs et de filtres numériques. Les paramètres de ces modèles sont ceux des générateurs et filtres qui les constituent.

Un modèle électrique de la phonation : le modèle AutoRégressif (AR)

Fant a proposé en 1960 un modèle de production dont nous résumons ici la version numérique.

Un signal voisé peut être modélisé par le passage d'un train d'impulsions *u(n)* à travers un filtre numérique récursif de type tout pôles. On montre que cette modélisation reste valable dans le cas de sons non-voisés, à condition que *u(n)* soit cette fois un bruit blanc. Le modèle final est illustré à la figure 2.4. Il est souvent appelé modèle auto-régressif, parce qu'il correspond dans le domaine temporel à une régression linéaire des *p* cœfficients précédents.

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad (2.11)$$

Ce qui exprime que chaque échantillon est obtenu en ajoutant un terme d'excitation à une prédiction obtenue par combinaison linéaire de *p* échantillons précédents. Les coefficients du filtre sont d'ailleurs appelés coefficients de prédiction et le modèle AR est souvent appelé modèle de prédiction linéaire.

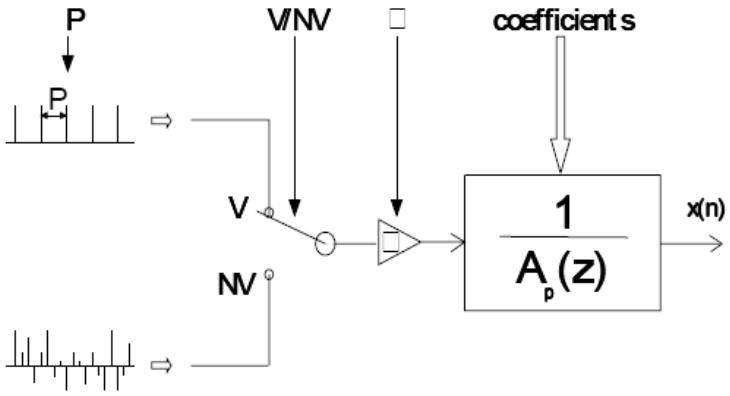


Fig.2.4 le modèle auto-régressif

Les paramètres du modèle AR sont : la période du train d'impulsions (sons voisés uniquement), la décision Voisé/NonVoisé (V/NV), le gain □, et les coefficients du filtre 1/A(z), appelé filtre de synthèse.

Le problème de l'estimation d'un modèle AR, souvent appelée analyse LPC revient à déterminer les coefficients d'un filtre tout pôles dont on connaît le signal de sortie, mais pas l'entrée. Il est par conséquent nécessaire d'adopter un critère, afin de faire un choix parmi l'infinité de solutions possibles.

La fonction de transfert du modèle de la production de la parole est décrite par la relation (2.11). Ainsi, chaque échantillon de la parole *s(n)* est constitué par une combinaison linéaire de *p* échantillons passés de la parole. Le prédicteur est défini comme un système dont la sortie est:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \qquad (2.12)$$

l'erreur de la prédiction est donnée par :

$$e(n)=s(n)-\hat{s}(n)=s(n)-\sum_{k=1}^{p} a_k s(n-k) \qquad (2.12)$$

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 :

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

$$\partial E/\partial a_i=0 \qquad \text{pour } i=1,\ldots,p \qquad (2.14)$$

alors :

$$\frac{\partial E}{\partial a_i} = -2\sum_n \left\{\left[s(n)-\sum_{k=1}^{p} a_k s(n-k)\right] s(n-i)\right\} = 0 \qquad (2.15)$$

cette dernière équation nous conduit à écrire :

$$\sum_n s(n)s(n-i) = \sum_n \sum_{k=1}^{p} a_k s(n-i)s(n-k) \qquad (2.16)$$

on définit:

$$\phi(i,k)=\sum_n s(n-i)s(n-k) \qquad (2.17)$$

alors:

$$\sum_{k=1}^{p} a_k \, \phi(i,k) = \phi(i,0) \qquad i=1,\ldots,p \tag{2.18}$$

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 $0 < n < L_a\text{-}1$, 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 :

$$\phi(i,k) = \sum_{n=0}^{La+P-1} s(n-i)s(n-k) \qquad i=1,\ldots,p$$

$$k=0,\ldots,p \tag{2.19}$$

$$\text{on pose } m=(n-i), \qquad \phi(i,k) = \sum_{m=0}^{La-1-(i-k)} s(m)s(m+i-k) \tag{2.20}$$

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

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

$$donc: \sum_{k=1}^{p} a_k R(|i-k|)=R(i) \tag{2.21}$$

on obtient :

$$\begin{pmatrix} R(0) & R(1) & R(2) & \ldots & R(p-1) \\ R(1) & R(0) & R(0) & \ldots & R(p-2) \\ R(2) & R(1) & R(0) & \ldots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \ldots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(p) \end{pmatrix} \tag{2.22}$$

La matrice 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].

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

$\forall\, i=1,2,\ldots,p$

$$a_j = a_j^{(p)} \quad \forall 1 \le j \le p \tag{2.23}$$

H(z) peut se mettre sous la forme :

$$H(z)=\sigma/A(z) \text{ Avec : } A(z)=1+\sum_{i=1}^{p} a_i z^{-i} \tag{2.24}$$

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

Les paramètres LSP (Line Spectral Pair) ont été présentés dans 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:

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_P z^{-P} \tag{2.25}$$

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 :

$$A(z)=0.5[P(z)+Q(z)]$$

$$P(z)=A(z)+z^{-(p+1)} A(z^{-1}) \tag{2.26}$$

$$Q(z)=A(z)-z^{-(p+1)} A(z^{-1})$$

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^{jwi} 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 :

$$0= w_0^{(Q)}< w_1^{(P)}< w_2^{(Q)}<\ldots< w_P^{(Q)}< w_{P+1}^{(P)}= \pi \qquad (2.27)$$

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

2.4 Conclusion

Dans ce chapitre on a définit les principales méthodes de paramétrisations telle que celles utilisées pour déterminer les cœfficients MFCC qui repose sur la séparation de la source avec le conduit vocal. L'utilisation de ces paramètres pour le traitement se fait avec des matrices de covariance diagonale. Etant donné qu'elles fournissent des paramètres discriminants étendus sur tout le spectre de la bande de perception de l'oreille humaine. Ainsi que les cœfficients LSP qui nous fournit des coefficients qui peuvent caractériser le conduit vocal, ils reposent sur le fait que les échantillons successifs sont corrélés.

Chapitre 3

Classification des vecteurs acoustiques

3.1 Introduction

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

- 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 la 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 mélanges de gaussiennes et par 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 référence dont les modèles sont bien appris.

3.2. Quantification vectorielle (QV)

3.2.1. Introduction

La numérisation demeure la méthode la plus simple de codage d'un signal. Cette technique consiste à coder le signal échantillon par échantillon, à l'aide d'un quantificateur scalaire. L'idée de la quantification vectorielle est de segmenter le signal en blocs d'échantillons. Ces blocs que l'on nomme vecteurs, seront codés par d'autres vecteurs prédéfinis, choisis dans un catalogue couramment appelé dictionnaire ou répertoire.

3.2.2. Définition

La quantification vectorielle (notée QV) consiste à représenter tout vecteur *x* de dimension *k* par un autre vecteur y_i de même dimension mais appartenant à un ensemble fini *D* de *L* vecteurs. Les y_i sont appelés les vecteurs représentants, les vecteurs de reproduction ou les codes vecteurs. *D est* appelé le dictionnaire où catalogue des formes.

La quantification vectorielle permet d'avoir une constellation qui minimise l'erreur quadratique moyenne pour un dictionnaire de taille *k* donnée. Elle permet de tirer partie de la corrélation qui existe souvent entre les composantes d'un vecteur.

La quantification vectorielle peut fournir un décodage rapide en utilisant une table simple d'identification. La figure 3.1 illustre ce principe.

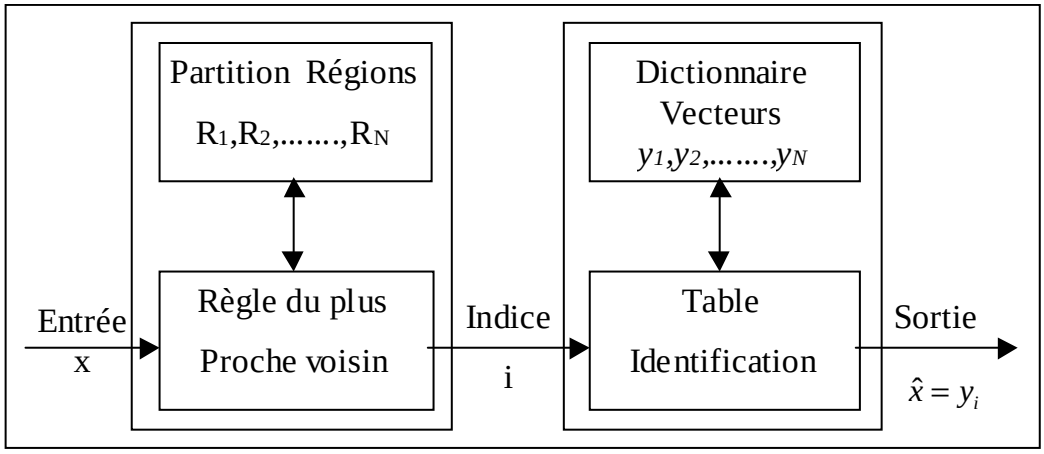


FIG.3.1 Modèle d'un quantificateur vectoriel

Un quantificateur vectoriel de dimension *k* et de taille *L* peut être défini mathématiquement comme une application *Q* de R^k vers *D* :

$$\begin{aligned} Q: \quad & R^k \rightarrow D \\ & x \qquad Q(x) = y_i \\ & D = \{ y_i \in R^k / i = 1,2,...,L \} \end{aligned} \tag{3.1}$$

Cette application *Q* détermine implicitement une partition de l'espace source R^k en *L* régions C_i. Ces régions, encore appelées **classes** ou **régions de Voronoï,** sont déterminées par :

$$C_i = \{ x \in R^k / Q(x) = y_i \} \tag{3.2}$$

En supposant que la grandeur d'entrée est un vecteur aléatoire distribué selon une loi *p(x)*, les performances du quantificateur peuvent être mesurées par la distorsion moyenne D_Q introduite, c'est à dire par l'espérance mathématique de la distance *d* :

$$D_Q = E\left[d(x,Q(x))\right] = \int d(x,Q(x)).\ p(x).dx \qquad (3.3)$$

Dans la pratique, la distribution des points d'entrée étant généralement inconnue, on approximera D_Q par une distorsion moyenne calculée sur un large nombre d'échantillons $\{x_1, x_2, .., x_N\}$ de vecteurs d'entrée :

$$D_Q \cong \frac{1}{N}\sum_{j=1}^{N} d(x_j, Q(x_j)) \qquad (3.4)$$

La distance introduit implicitement une partition de l'ensemble des vecteurs d'entrée en k classes $\{C_i, i = 0, 1,..., k-1\}$. La classe C_i étant l'ensemble des vecteurs associés à y_i par le quantificateur :

$$C_i = Q^{-1}(y_i) = \{x;\ Q(x) = y_i\} \qquad (3.5)$$

Nous appellerons centroïde de la classe C_i le vecteur c_i tel que sa distance moyenne à tous les éléments de la classe soit minimale (en géométrie euclidienne, le centroïde est le centre de gravité) :

$$E[\,d(x,c_i)\ ;\ x \in C_i] = \inf_{x_i}\{\ E[d(x,x_i)\ ; x \in C_i]\} \qquad (3.6)$$

Etant donné une distance et une taille de dictionnaire, il existe un quantificateur qui minimise la distorsion moyenne : c'est le quantificateur optimal.

3.2.3. Quantificateur Vectoriel Optimal

Le quantificateur optimal répartit les vecteurs de production en tenant compte de la distribution (densité de probabilité multidimensionnelle) des vecteurs à coder dans l'espace.

Un quantificateur se décompose en deux applications : un codeur et un décodeur.
Le quantificateur optimal est alors celui réunissant le codeur optimal et le décodeur optimal.

- Le codeur optimal : étant donné un dictionnaire, $\{\hat{s}_1 .... \hat{s}_L\}$ la meilleure partition est celle qui vérifie : $R_i = \{s : (s-\hat{s}_i)^2 \le (s-\hat{s}_j)^2 \quad \forall j \in \{1,....,L\}\}$ C'est la règle dite du plus proche voisin.
- Le décodeur optimal : étant donné une partition $\{R_1 .....R_L\}$, les meilleurs représentants sont obtenus par la condition dite du «centroïde» (centre de gravité de la partie de la densité de probabilité placée dans la région R_i).

$$\hat{s}_i = \frac{\int_{x \in R_i} x \; P_s(x) \; dx}{\int_{x \in R_i} P_s(x) \; dx} = E\{ s / \; s \in \; R_i \} \tag{3.7}$$

- Une troisième condition est nécessaire : il faut que la probabilité qu'un vecteur à coder se trouve à la même distance de deux représentants soit nulle, sinon ce vecteur source est affecté à l'un des deux représentants, et dans ce cas, la partition de l'espace n'est plus optimale. Si les vecteurs source sont à amplitude continue, cette troisième condition est toujours vérifiée.

Ces trois conditions conduisent à la conception d'un algorithme qui réalise, à partir d'une séquence d'apprentissage représentative de la statistique de la source à coder, la construction d'un dictionnaire optimal. Cet algorithme de classification, encore appelé algorithme des k-moyens (*k-means*) est l'extension au cas vectoriel de l'algorithme de LLOYD-MAX du cas scalaire.

3.2.4. Algorithme de Lloyd Généralisé (LBG)

Le principe de LLOYD-MAX généralisé au cas vectoriel reste identique à celui du cas scalaire, il faut :

- Initialiser le dictionnaire.
- Appliquer successivement la règle du plus proche voisin et la condition du centroïde.
- Itérer l'étape précédente tant que la décroissance de la distorsion moyenne reste importante.

Le choix du dictionnaire initial est essentiel car il conditionne les résultats finaux de l'algorithme. Plusieurs méthodes ont été proposées pour le déterminer on peut citer

- ***Initialisation aléatoire***

Le dictionnaire le plus simple est celui qui contient les *L* premiers vecteurs de la suite d'apprentissage ou *L* vecteurs extraits aléatoirement de cette suite. Ces vecteurs peuvent bien sûr ne pas être du tout représentatifs de la suite d'apprentissage et on aboutit à des résultats très médiocres.

- ***L'algorithme à seuil***

Au lieu de prendre *L* vecteurs aléatoirement, on fixe une distance minimale entre les éléments du dictionnaire initial. Cette méthode permet d'obtenir une meilleure représentativité que dans le cas précédent.

- ***Méthode des vecteurs produits***

Cette méthode nécessite de quantifier scalairement les *k* composants des vecteurs de la séquence d'apprentissage et d'effectuer un produit cartésien entre les dictionnaires de base pour obtenir les *L* représentants initiaux.

- ***Méthode par dichotomie vectorielle***

Cette méthode introduit, dans l'algorithme LBG, une technique de «Splitting » à l'itération de LLOYD. Celle-ci consiste à découper chaque vecteur représentant y_i en deux nouveaux vecteurs $y_i + \varepsilon$ et $y_i - \varepsilon$; (ε étant un vecteur de perturbation), avant d'appliquer au nouveau dictionnaire obtenu les itérations de LLOYD. L'algorithme génère ensuite une succession de dictionnaire.

La figure suivante donne le schéma de fonctionnement de l'algorithme LBG.

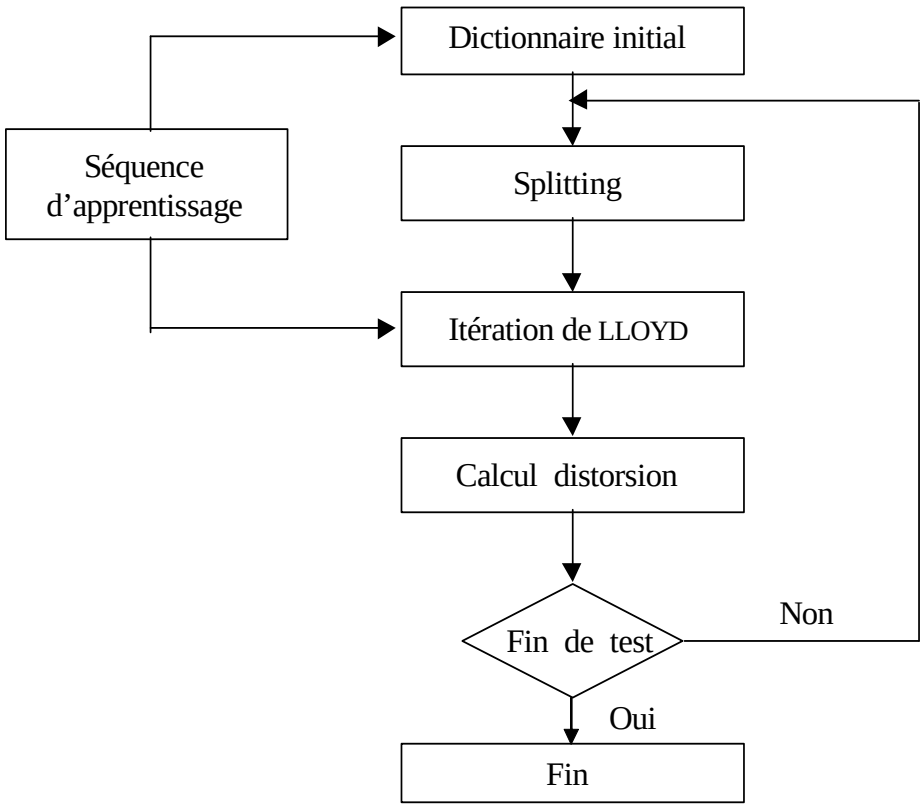


FIG.3.2 Schéma de fonctionnement de l'algorithme LBG

3.3. Les Modèles de Markov Cachés HMM (Hidden Markov Models)

Le modèle HMM est introduit dans un cadre purement statistique, il s'est ensuite imposé en reconnaissance de la parole avant d'être appliqué en reconnaissance automatique du locuteur. Le modèle HMM présente différents avantages : clarté, rigueur, efficacité et généralité. Un modèle HMM se caractérise par un système à états comportant deux processus.

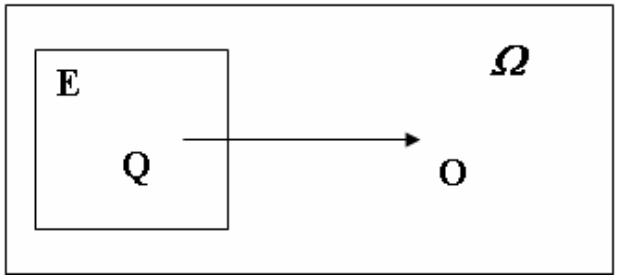


FIG. 3.3 Constituants d'un HMM

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

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

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

Soit λ un modèle Markovien de N états et $Q=(q_1,q_2,\ldots q_T)$ une séquence d'états correspondant à l'observation $O=(o_1,o_2,\ldots o_T)$ où q_t est le numéro de l'état atteint par le processus à l'instant t. L'état du modèle de Markov λ qui correspond à o_t n'étant pas directement observable, on dit qu'il est caché. D'où le nom de modèle de Markov caché. La figure 3.4 représente un exemple de modèle de Markov. Un tel modèle est définit par :

un ensemble d'états cachés $\{S_1, S_2, \ldots, S_N\}$.

un ensemble d'observations $\{V_1, V_2, \ldots, V_M\}$..

probabilités de transition $a_{ij}=P(q_{t+1}=S_j / q_t=S_i)$.

probabilités d'observation $b_{ij}=P(o_t=v_k / q_t=S_j)$, qui sont en général des mélanges de gaussiennes.

un ensemble de probabilités initiales de se trouver dans chaque état $\pi=\{\pi_i / \pi_i=P(q_1=S_j) \; i=1...N\}$

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

3.3.1. Problèmes des modèles HMM

Trois problèmes se posent avec les modèles de Markov cachés :

L'évaluation

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

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

Estimation des états cachés

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

Apprentissage

C'est le problème principal d'un modèle HMM. En effet, la qualité d'un système utilisant une modélisation HMM dépend principalement de la qualité de ses modèles. C'est

pourquoi l'étape d'apprentissage qui consiste à estimer les paramètres des modèles HMM est très importante. Il existe plusieurs méthodes pour résoudre ce problème, nous présentons les plus utilisées :

L'algorithme de Viterbi associé à des estimateurs empiriques : l'algorithme de Viterbi sert à déterminer la séquence d'états cachés la plus vraisemblable, correspondant aux données d'apprentissage. Les paramètres des densités de probabilité de chaque état peuvent être alors ré-estimés en utilisant des estimateurs empiriques et les observations associées à chaque état le long du chemin de Viterbi.

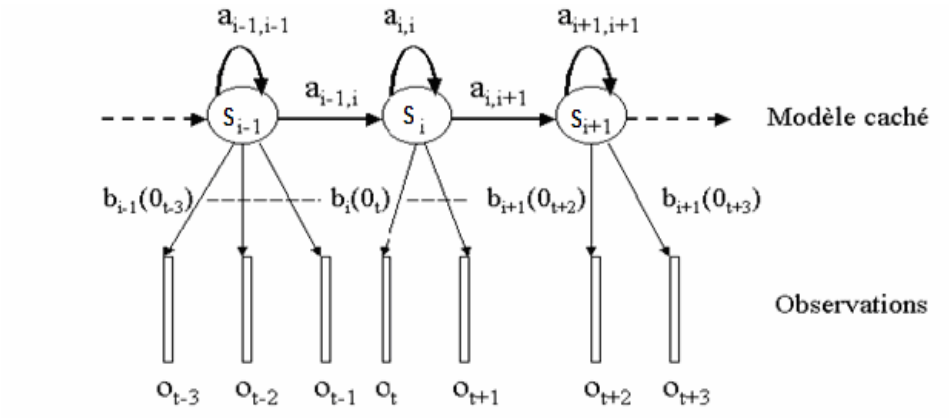
L'algorithme EM (Expectation-Maximisation) : cet algorithme permet de résoudre le problème d'apprentissage en estimant de manière itérative les paramètres d'un modèle au sens du maximum de vraisemblance.

3.3.2. La phase de reconnaissance

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

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

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



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

FIG. 3.4 Exemple d'une machine Markovienne

3.4. 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 :

$$P(x|\lambda_s)=\sum_{m=1}^{M} \pi_m^s b_m^s \tag{3.8}$$

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

$$b_m^s = \frac{1}{(2\pi)^{D/2}\left|\sum_m^s\right|^{1\backslash 2}} .\exp\left[-\frac{1}{2} (x-\mu_m^s)' (\sum_m^s)^{-1} (x-\mu_m^s) \right] \tag{3.9}$$

et les représentent les poids du mélange, avec :

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

un locuteur est donc modélisé par un ensemble de paramètres notés

$$\lambda_s = \left\{\pi_m^s, \mu_m^s, \sum_m^s\right\} m = 1,...,M$$

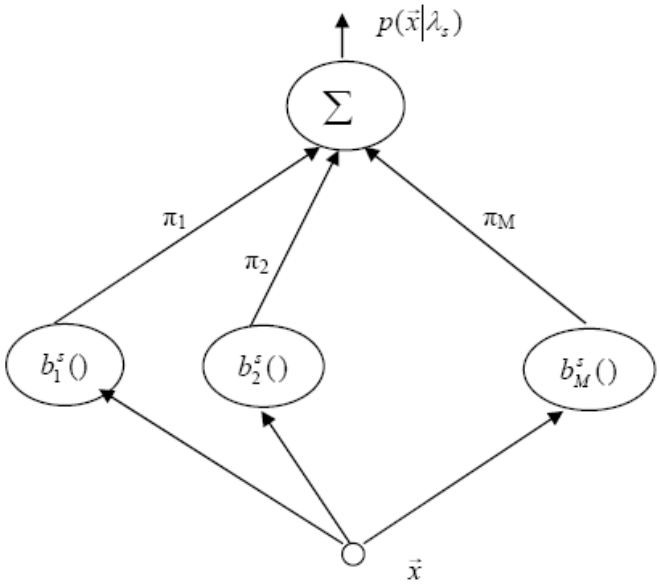


Fig.3.5 modèle de mélange de gaussiennes

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).

3.4.1. Apprentissage du modèle

Il s'agit, lors de la phase d'apprentissage, d'estimer l'ensemble λ des paramètres d'un modèle GMM du 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 $X=\{x_1,x_2,\ldots,x_N\}$ (suffisamment indépendants), la vraisemblance du modèle GMM est :

$$P(X\backslash\lambda)=\prod_{n=1}^{N} P(x_n\backslash\lambda) = \prod_{n=1}^{N} \sum_{m=1}^{M} P(x_n\backslash\pi_m,\mu_m, \sum_m) \qquad (3.10)$$

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 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.

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 :

$$Q(\theta,\theta^{(t)})=\sum\sum p(m\backslash x_n,\theta^{(t)}).log\ p(x_n,m\backslash\ \theta) \quad (3.11)$$

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

$$Q(\theta, \theta^{(t)})=\sum\sum \gamma_{n,m}^{(t)} \left[\log \pi_m - \frac{D}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma_m| \right] - \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\frac{1}{2} (x_n - \mu_m) \right] \quad (3.12)$$

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

$$\gamma_{n,m}^{(t)} = \frac{\pi_m^{(t)}\, p(x_n \backslash \mu_m^{(t)}, \Sigma_m^{(t)})}{\sum_{k=1}^{M} \pi_k^{(t)}\, p(x_n \backslash \mu_k^{(t)}, \Sigma_k^{(t)})} \quad (3.13)$$

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

$$Q(\theta, \theta^{(t)})= \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \log \pi_m - \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\text{Cste} + \log \sigma_m^2 + \frac{(x_n - \mu_m)^2}{\sigma_m^2} \right] \quad (3.14)$$

Où σ^2_m 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. Cela dit, il faut tenir compte de la contrainte qui existe sur ces paramètres

La maximisation sous contrainte se résout simplement en introduisant un multiplicateur de Lagrange associé à cette contrainte et l'on obtient :

$$\pi_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,m}^{(t+1)} \quad (3.15)$$

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

$$\mu_m^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} x_n}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{3.16}$$

Et pour les variances

$$\sigma_m^{2(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)} (x_n - \mu_m^{(t)})^2}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{3.17}$$

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. 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 [15] :

- Les poids des gaussiennes

$$\pi_m = \frac{n_m^0 + n_m}{\sum_{k=1}^{M} (n_k^0 + n_k)} \tag{3.18}$$

- Les vecteurs des moyennes

$$\mu_m = \frac{n_m^0 \overline{X_m^0} + n_m \overline{X_m}}{n_m^0 + n_m} \tag{3.19}$$

- Les variances

$$\sigma_m^2 = \frac{n_m^0 \overline{X_m^0 X_m^{0\prime}} + n_m \overline{X_m X_m'}}{n_m^0 + n_m} - \mu_m \mu_m' \tag{3.20}$$

Où n (respectivement n^0) représente le poids, $\overline{X}$ (respectivement $\overline{X^0}$) le moment d'ordre 1 et $\overline{XX'}$ (respectivement $\overline{X^0 X^{0'}}$) 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.

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, les modèles initiaux correspondent au modèle du monde UBM (Universal Background Model).

3.4.2. 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 ici la phase de décision d'un système d'identification.

Soit un groupe de S locuteurs, représentés par les modèles GMM :

$\lambda_1,\lambda_2 ,\ldots,\lambda_s$. 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 :

$$\hat{s} = arg\ max\ p(\lambda_s \backslash X) \quad \text{avec} \quad 0 \le s \le S \qquad (3.21)$$

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

$$\hat{s} = \arg\max \frac{p(X \setminus \lambda_s)}{p(X)} p(\lambda_s) \quad \text{avec} \quad 0 \leq s \leq S \qquad (3.22)$$

En supposant l'équiprobabilité d'apparition des locuteurs $P(\lambda_s) = 1/S$, la loi de classification devient :

$$\hat{s} = arg\ max\ p(X \setminus \lambda_s) \quad \text{avec} \quad 0 \leq s \leq S \qquad (3.23)$$

En utilisant le logarithme et l'indépendance entre les observations, le système d'identification calcule le score *0* suivant :

$$\hat{s} = arg\ max \sum_{n=1}^{N} log\ p(x_n \setminus \lambda_s) \quad \text{avec} \quad \leq s \leq S \qquad (3.24)$$

3.5. Identification par mélanges de gaussiennes orthogonales (OGMM)

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

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

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

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

L'idée de base est d'effectuer une analyse en composantes principales (ACP) [11].

Pour réduire la corrélation entre les coefficients acoustiques, une transformation linéaire est opérée sur les vecteurs acoustiques. La matrice de transformation dans ce cas est propre à chaque locuteur, et est composée des vecteurs propres de la matrice de covariance initiale du même locuteur. Cette étape terminée, on applique sur les vecteurs acoustiques résultants la modélisation GMM standard.

3.6. Utilisation du Pitch

3.6.1 Introduction

La fréquence de vibration des cordes vocales appelée fréquence fondamentale (ou pitch), est un paramètre très important pour caractériser le locuteur. L'oreille est en effet très sensible à ces 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.

3.6.2 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).

Le pitch joue un rôle très important quand la dépendance de la source et le conduit vocal est maintenue [1]. 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.

3.6.3 La reconnaissance

Pendant l'apprentissage on estime un modèle par locuteur, le signal parole est découpé en trames de 20ms chacune, avec décalage de fenêtre de 10ms. Pour la reconnaissance on utilise la méthode basée sur l'estimation de la probabilité a posteriori.

Reconnaissance basée sur l'estimation de la probabilité a posteriori

Des mesures statistiques montrent que 90% des valeurs du pitch interviennent à l'intervalle [100Hz, 250Hz]. Nous avons divisé l'espace en quatre sous espaces : I1 = [100Hz, 150Hz] Hz, I_2 = [150,200] Hz, I_3 = [200,250] Hz, 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 du locuteur se fait sur bande étroite, ou ce qu'on appelle l'effet de loupe. 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. En conséquence chaque locuteur est représenté par quatre mixtures gaussiennes (GMM). [1]

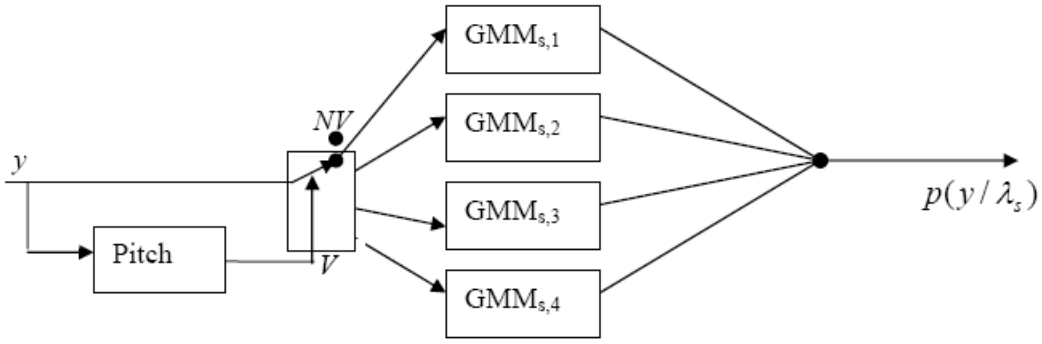


Fig. 3.6 Modèle de Reconnaissance basée sur l'estimation de la probabilité a posteriori

3.7. Conclusion

Dans la 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 centroïde. Ainsi, on va représenter un locuteur par son dictionnaire de quantification.

La modélisation GMM utilise des matrices de covariance diagonales pour gagner en temps de calcul. Néanmoins, cette approximation ne permet pas de prendre en compte la corrélation qui existe entre les coefficients acoustiques.

L’orthogonalisation de l’espace acoustique permet de réduire la corrélation entre les cœfficients acoustiques et la considération des matrices acoustiques devient dans ce cas justifiée.

Le pitch joue un rôle très important quand la dépendance de la source et le conduit vocal est maintenue. 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.

Chapitre 4

Évaluations expérimentales

4.1 Introduction

Ce chapitre présente l'évaluation expérimentale de quatre approches de l'identification du locuteur : la quantification vectorielle (QV), la modélisation par mixture de gaussiennes (GMM), OGMM et GMMpitch.

Avant de commencer la simulation il faut toujours préparer une base de données bien organisée, et c'est sur cette base que le système d'identification va être testé. Dans la première partie on donne la description des bases de données utilisée dans cette simulation, et dans les autres parties on donne les différentes étapes pratiques pour entraîner et tester un système d'identification du locuteur, ainsi que les résultats et les interprétations de la simulation.

4.2. Description des bases de données utilisées

4.2.1. La base de données TIMIT

La base de données TIMIT est une base de données acoustique et phonétique dédiée à la reconnaissance automatique de la parole ainsi qu'au développement et à l'évaluation des systèmes de reconnaissance automatique de la parole. Elle contient les enregistrements de 630 locuteurs américains prononçant chacun 10 phrases. Les données sont échantillonnées sur 16 KHz sur 16 bits.

- Le texte est lu dans de bonnes conditions d'enregistrement.
- Le type de parole : phrases continues en anglais,
- Microphone utilisé : large bande.

Dans le cadre de ce travail, on a utilisé une base de données composée de 32 locuteurs 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 5 phrases (de 6 à 10) pour l'apprentissage et les 5 autres phrases sont utilisées pour le test.

4.2.2. Notre base de données

Notre base de données est une base acoustique dédiée seulement pour l'identification du locuteur. Elle est constituée de 45 locuteurs (37 hommes et 8 femmes) algériens. Pour chaque locuteur, on dispose de 10 phrases indépendamment du texte, chacune de 3 secondes en moyenne. On a concaténé 5 phrases (6 à 10) pour l'apprentissage et les 5 autres phrases (1 à 5) sont utilisées pour le test. Les données sont échantillonnées sur 16 KHz sur 16 bits.

- La chambre d'enregistrement est un peu bruyante
- le type de parole : phrases continues en Arabe
- le microphone utilisé : bidirectionnelle.

L'enregistrement se fait dans le laboratoire du signal et communication de l'Ecole Nationale Polytechnique.

4.3 Analyse acoustique et paramétrisation du signal vocal

4.3.1 Extraction des paramètres

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

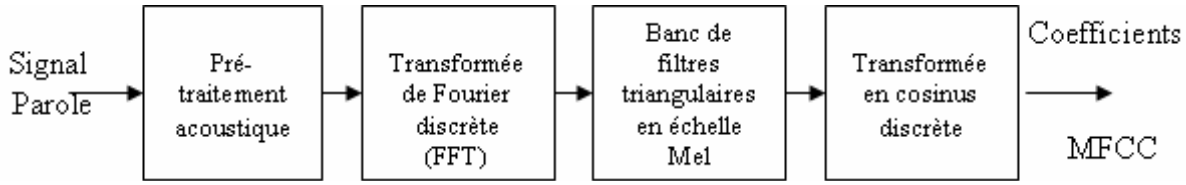


Fig. 4.1 Extraction des coefficients MFCC

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

Voici un schéma expliquant l'effet de pré-accentuation (la fréquence d'échantillonnage est sur 16 kHz).

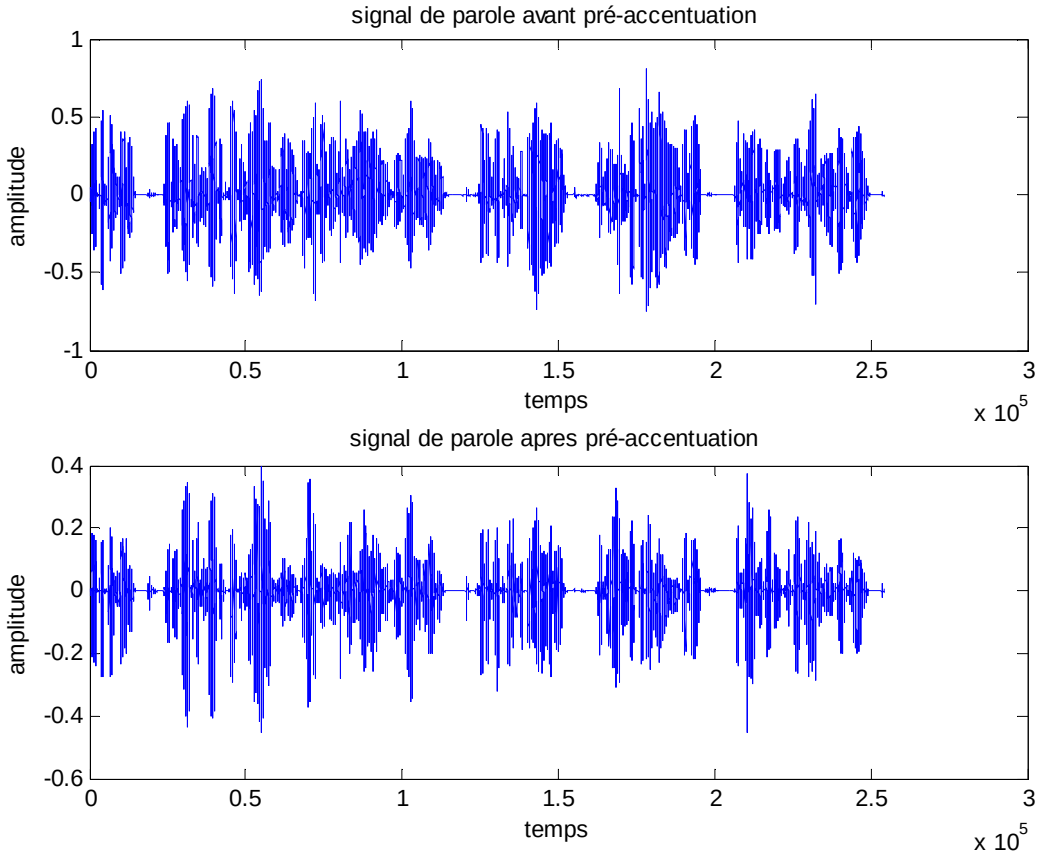


Fig. 4.2 l'effet de pré-accentuation

On constate une grande condensation dans le système avant la phase de pré-accentuation.

2. L'étape de fenêtrage qui consiste à multiplier le signal vocal par une fenêtre de pondération glissante. Nous avons utilisé une fenêtre de Hamming glissante de durée de 20 ms avec déplacement de 10 ms.

La figure 4.3 illustre une fenêtre de pondération de Hamming sur 512 échantillons, et qui est définie par :

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

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

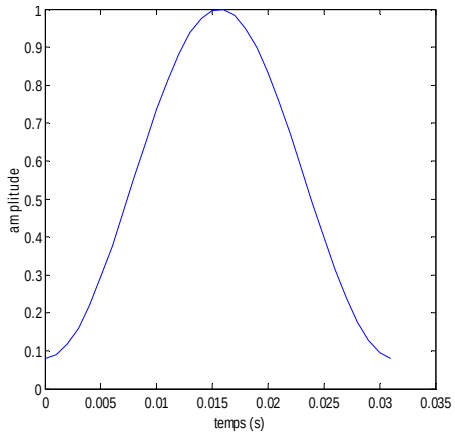


Fig. 4.3 Fenêtre de pondération de Hamming

4.3.2 Logiciel utilisé

On utilise MATLAB version 6.5 qui possède des instructions spéciales au traitement numérique de signal.

4.3.3 Détection et élimination du silence

Les périodes du silence 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, 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 du silence. Voici un exemple d'élimination du silence pour un pourcentage d'énergie inférieur à 0.008.

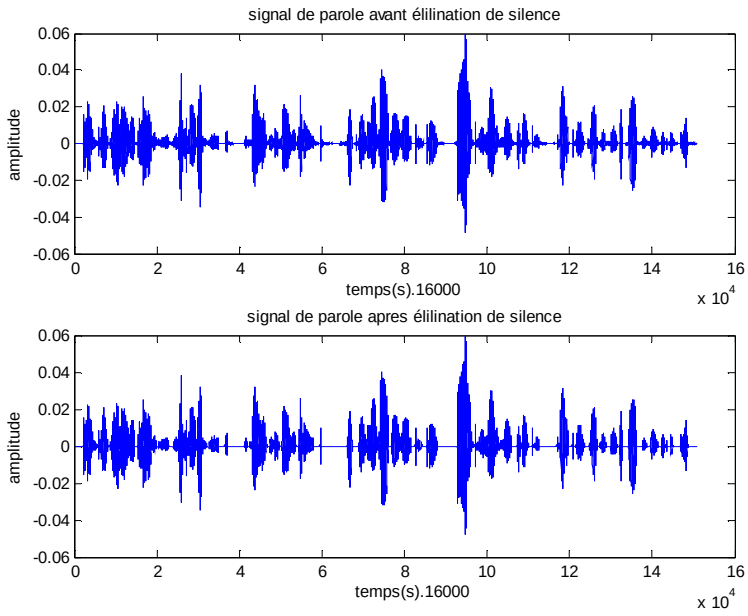


Fig. 4.4 Elimination du silence

4.3.4 Filtrage dans la bande téléphonique et ré-echantillonnage

La base de données est échantillonnée sur 16 kHz et les données sont sur 16 bits .On introduit une adaptation sur les données à utiliser afin d'obtenir une qualité réseau téléphonique commuté (RTC).

Tout d'abord on a procédé par un filtrage passe –bande 300-3400 Hz, ensuite on ré-échantillonne à 8 kHz .Enfin on ajoute un bruit blanc gaussien avec rapport signal sur bruit de 50 dB.

4.4. Protocole d'évaluation

Le taux d'identification correcte est défini par :

$$\text{Ic (\%)} = \frac{\text{Nombre de segments de test correctement identifiés}}{\text{Nombre total de segments de test}} \times 100$$

4.5 Evaluations expérimentales

a. L'ordre du modèle

On varie l'ordre du modèle (nombre de centroïdes pour la quantification vectorielle) ou le nombre de composantes gaussiennes de 1 jusqu'à 64 gaussiennes pour voir l'effet de cette variation sur le taux d'identification.

b. La qualité des paramètres d'identification

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

c. Le nombre de locuteurs

On utilise 32 locuteurs

d. L'algorithme d'apprentissage

On va évaluer les résultats obtenus avec les trois algorithmes d'apprentissage VQ, GMM, OGMM et GMMpitch.

4.5.1 Influence du nombre de cœfficients

On fixe le nombre des classes VQ (avec les paramètres MFCC). La fréquence d'échantillonnage est 16 KHz. On fait changer le nombre de coefficients et les résultats obtenus sont dans le tableau ci-dessous :

Nombre de coefs	1	5	10	15	20	25	40
Taux (%)	19.37	81.87	92.5	95.62	96.87	97.5	98.12

Tableau 4.1 influence du nombre de coefficients

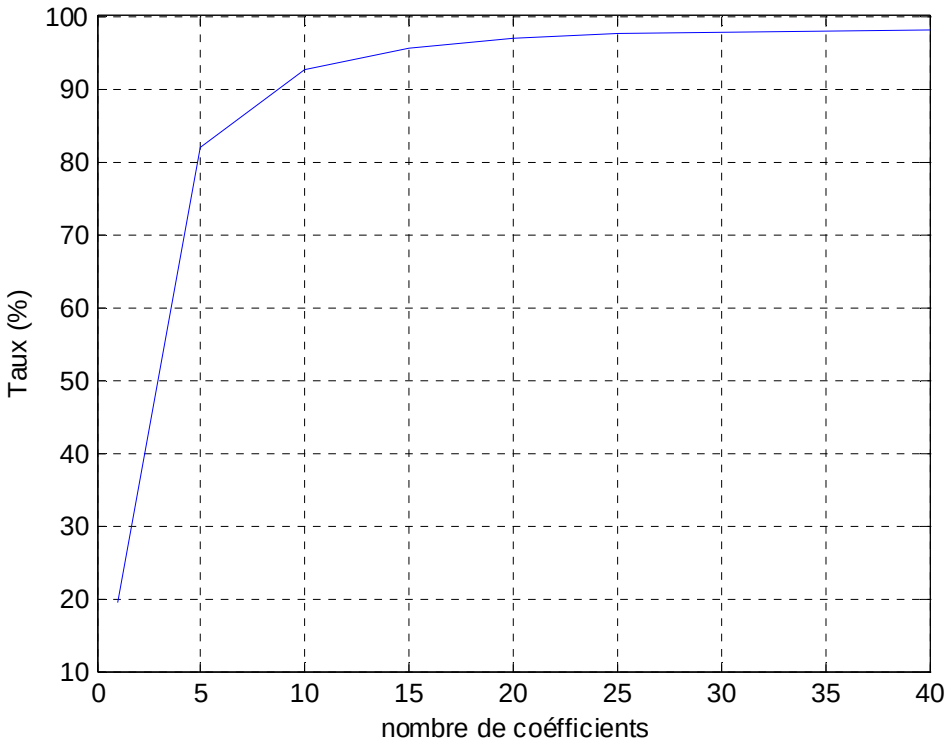


Fig. 4.5 influence du nombre de cœfficients

Commentaires et conclusion

- on constate que la quasi-totalité d'énergie de signal de la parole est contenue dans les 15 premiers coefficients MFCC.
- les coefficients MFCC d'ordre supérieur n'apportent pratiquement pas un plus d'informations sur l'identité du locuteur.
- pour une fréquence d'échantillonnage de 16 kHz, on constate qu'il faut utiliser un nombre de coefficients supérieur à 10 pour avoir de bonnes performances.

4.5.2 Influence de la quantité de données

Les résultats obtenus sur notre base de données avec la fréquence de 16 kHz avec VQ et les paramètres MFCC sont dans le tableau ci-dessous :

nbr de classes	1	2	4	8	16	32	64
I_C (MFCC (1))	18.12	56.87	70	76.25	82.5	82.5	78.12
I_C (MFCC (5))	71.25	68.75	84.37	91.25	93.12	93.25	94.37

Tableau 4.2 influence de la quantité de données

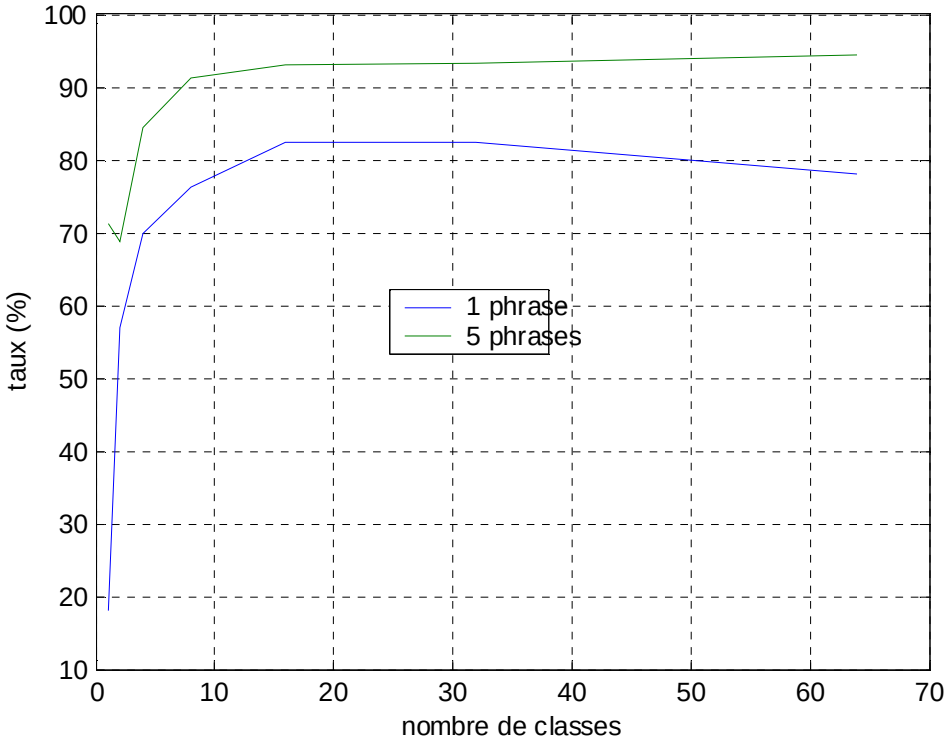


Fig. 4.6 influence de la quantité de données

Commentaires

- On observe que pour les deux courbes le taux d'identification augmente avec la quantité de données d'apprentissage. On remarque encore que les deux courbes ont pratiquement la même allure.
- On constate que la quantité de données d'entraînement est un facteur déterminant pour les performances d'un système d'identification.

4.5.3 Etude comparative entre notre base de données et la base TIMIT

On fait la comparaison dans le cas où le nombre de locuteurs utilisés est 32 et le nombre de coefficients est de 12. La fréquence d'échantillonnage est 16KHz.

a. Cœfficients MFCC avec une classification VQ (quantification vectorielle)

les résultats obtenus sont dans le tableau ci-dessous :

nbr de classes	1	2	4	8	16	32	64
I_C (TIMIT)	38.75	77.5	79.37	85.62	92.5	97.5	98.75
I_C (Notre base)	71.25	68.75	84.37	91.25	93.12	96.25	94.37

Tableau 4.3 VQ-MFCC: comparaison entre notre base et la base TIMIT

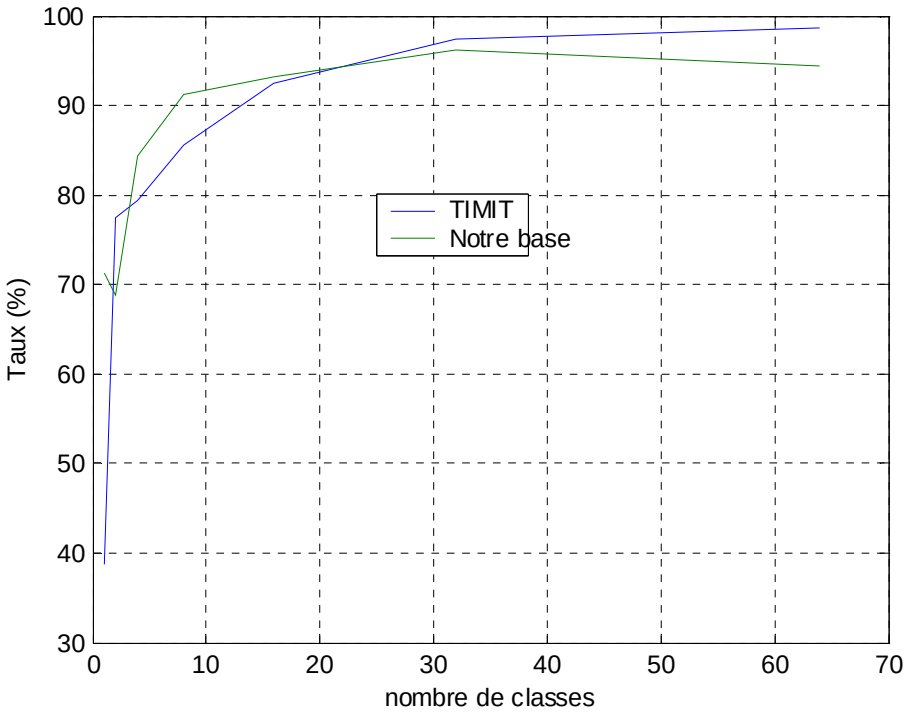


Fig. 4.8 VQ-MFCC: comparaison entre notre base et la base TIMIT

Au début où le nombre de classes est inférieur à 10 notre base de données donne des taux d'identification supérieurs à ceux de TIMIT .Dès que le nombre de cœfficients dépasse 10, la base TIMIT donne des résultats meilleurs que les nôtres. On a un maximum de taux d'identification pour un nombre de classes égale 32 pour les deux bases.

b. Cœfficients LSP avec classification VQ (quantification vectorielle)

nbr de classes	1	2	4	8	16	32	64
I_C (TIMIT)	45.62	81.25	90	93	95.62	97.5	96.87
I_C (Notre base)	62.5	77.5	88.12	92.5	95	94.37	95.62

Tableau 4.4 VQ-LSP : comparaison entre notre base et la base TIMIT

Ic : taux d'identification corrects

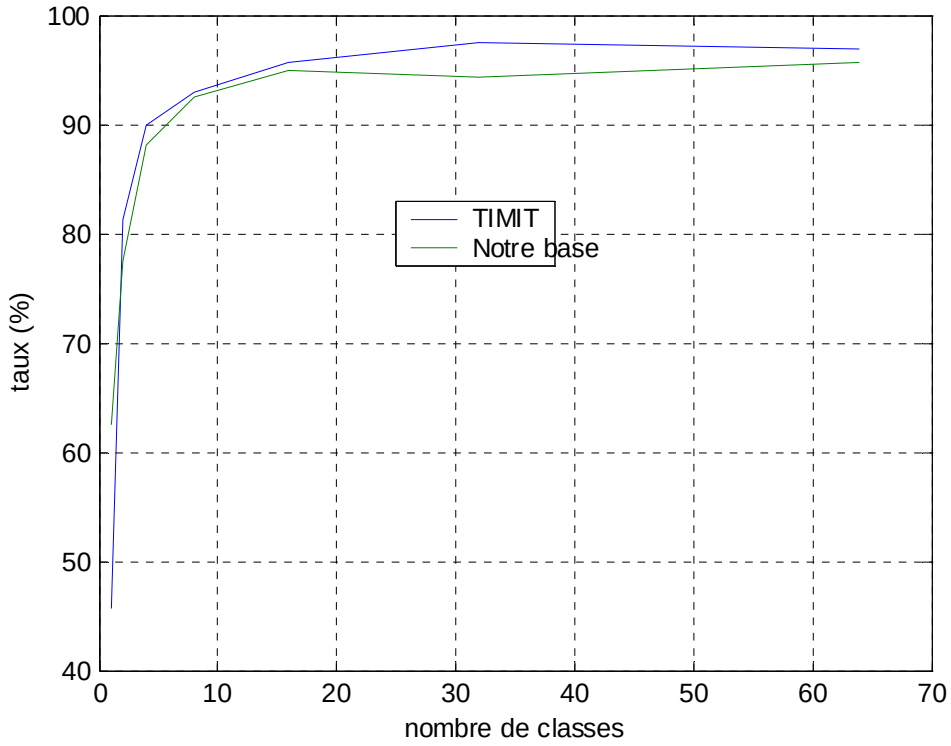


Fig. 4.8 VQ-LSP: comparaison entre notre base et la base TIMIT

On observe comme précédemment, que la base TIMIT donne un meilleur taux d'identification dans le cas où le nombre de classes est supérieur à 5. Elle se stabilise vers la valeur 97(%).Par contre notre base se stabilise vers la valeur 95(%).

Donc pour la quantification vectorielle et pour les nombres de classes inférieures notre base de données les taux sont meilleurs, par contre pour les nombres de classes supérieurs la base TIMIT donnent des taux d'identifications meilleurs que la nôtre.

c. Cœfficient MFCC avec une classification GMM

nbr de classes	1	2	4	8	16	32	64
I_C (Notre base)	40	56.25	73.75	85	89.37	85.62	81.87
I_C (TIMIT)	56.87	66.87	83.75	87.5	92.5	91.25	89.37

Tableau 4.5 GMM-MFCC: comparaison entre notre base et la base TIMIT

Ic : taux d'identification corrects

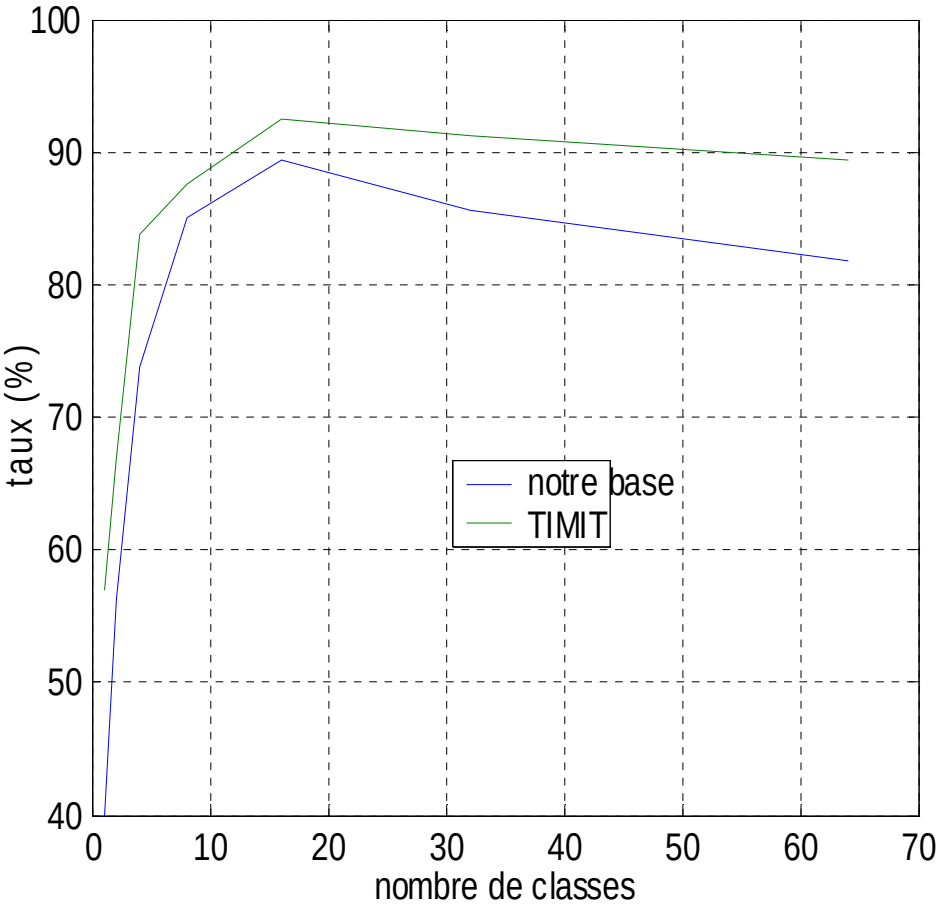


Fig. 4.9 GMM-MFCC: comparaison entre notre base et la base TIMIT

Pour toutes les valeurs du nombre de classes les résultats obtenus avec TIMIT sont meilleures que ceux obtenues avec notre base. On a un maximum de taux d'identification avec un nombre de classes égales à16 dans les deux cas.

d. Cœfficients LSP avec une classification GMM

Nbr de classes	1	2	4	8	16	32	64
I_C (TIMIT)	81.25	93.75	96.25	96.25	98.75	98.75	94.37
I_C (Notre base)	78.75	85.62	90	93.75	94.37	95	92.5

Tableau 4.6 GMM-LSP: comparaison entre notre base et la base TIMIT

Ic : taux d'identification corrects

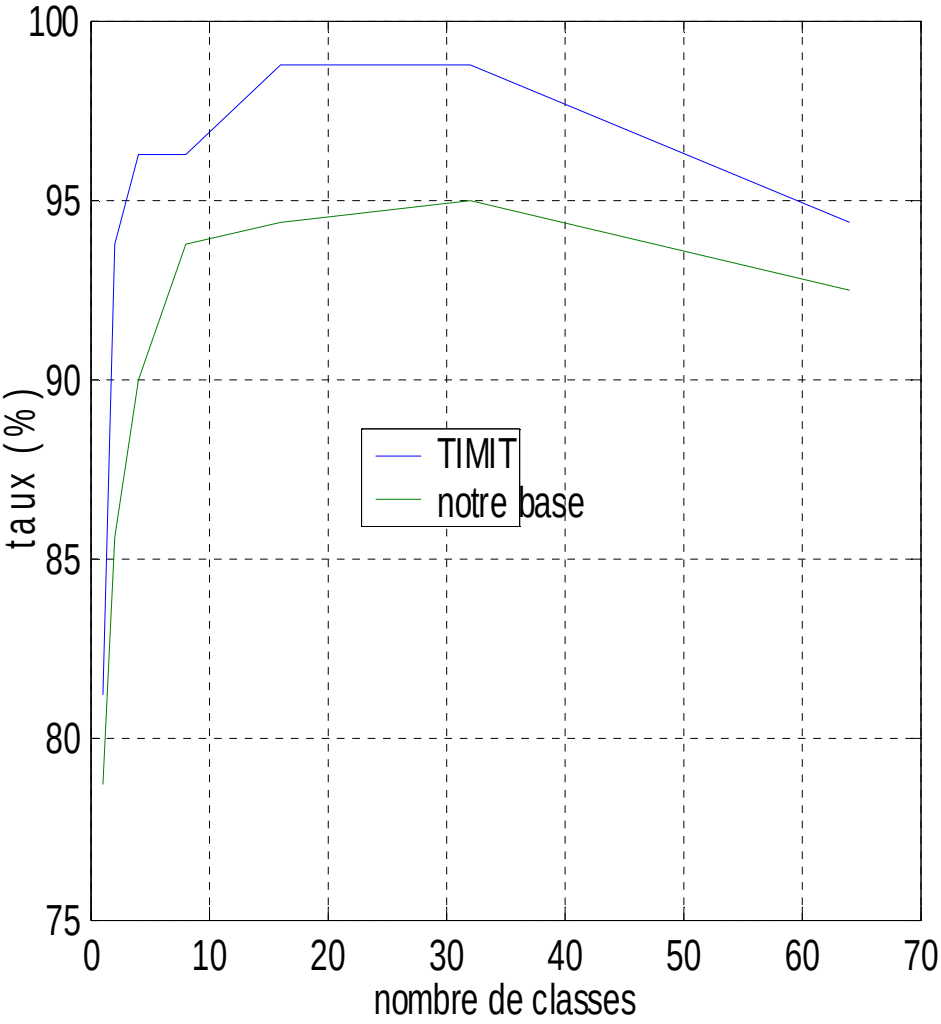


Fig. 4.10 GMM-LSP: comparaison entre notre base et la base TIMIT

Pour toutes les valeurs du nombre de classes les résultats obtenus avec la base TIMIT sont meilleures que ceux obtenues avec notre base. On a un maximum de taux d'identification avec un nombre de classes de 16 à 32 en utilisant la base TIMIT et un nombre de classes de 32 sur notre base.

Conclusion

Pour tous les cas, les résultats obtenus avec la base TIMIT sont meilleurs que ceux obtenus avec notre base, et cela est dû

- au bruit ambiant,
- à l'utilisation du microphone bidirectionnel,
- au problème d'adaptation de l'homme à la machine.

4.5.4 Etude de variation du taux d'identification en fonction de paramètres utilisés et la fréquence d'échantillonnage

a. la quantification vectorielle

Les résultats obtenus avec la fréquence 16 kHz sont dans le tableau ci-dessous :

nbr de classes	1	2	4	8	16	32	64
I_C (MFCC)	71.25	68.75	84.37	91.25	93.12	93.25	94.17
I_C (LSP)	62.5	77.5	88.12	92.5	95	94.37	95.62

Tableau 4.7 La Quantification Vectorielle ; Fe= 16 kHz
Ic : taux d'identification corrects

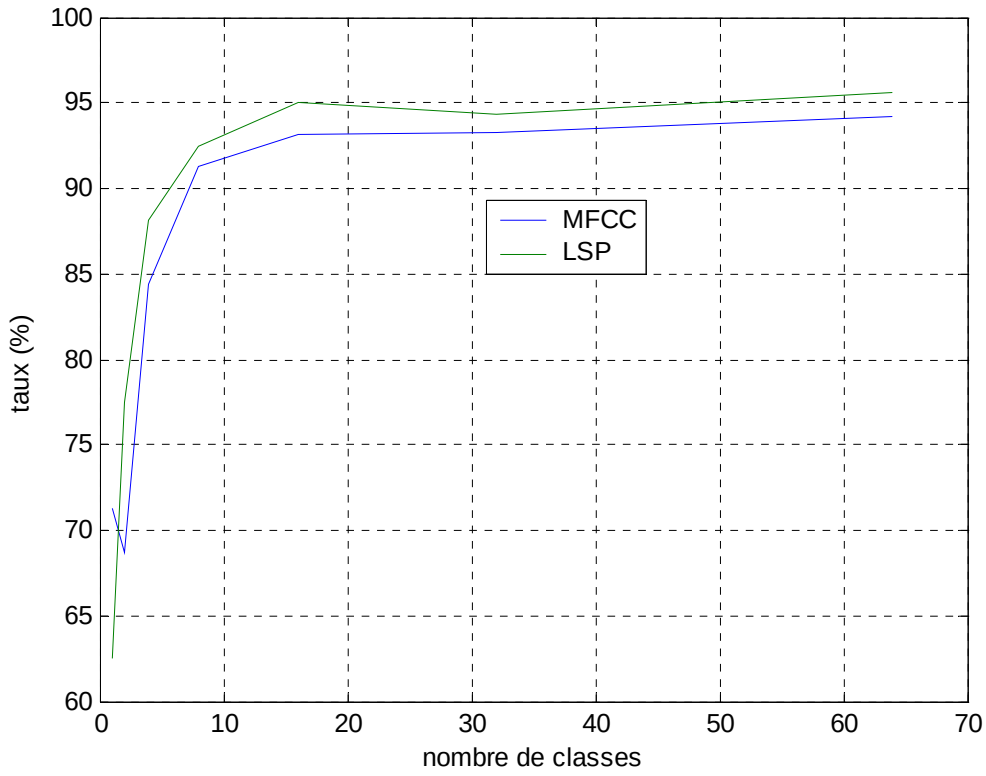


Fig. 4.11 La Quantification Vectorielle ; Fe= 16 kHz

Les résultats obtenus avec la fréquence 8 kHz sont dans les tableaux ci-dessous :

nbr de classes	1	2	4	8	16	32	64
I_C (MFCC)	55.62	55	66.25	78.75	87.50	91.88	92.50
I_C (LSP)	34.38	65.00	66.25	74.38	87.75	92.50	93.12

Tableau 4.8 La Quantification Vectorielle ; Fe= 8 kHz

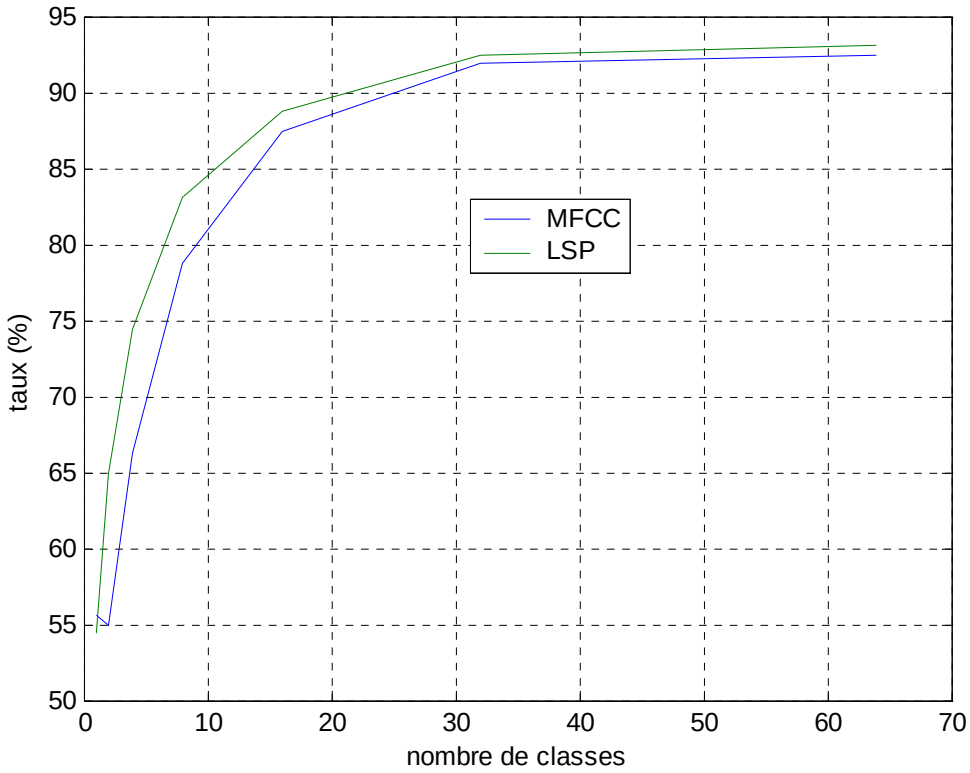


Fig. 4.12 La Quantification Vectorielle ; Fe= 8 kHz

La figure 4.11 donne le taux d'identification en fonction du nombre de centroides. Pour le système à 16 kHz le taux d'identification est croissant avec le nombre de centroides. Pour les LSP, il devient presque stable au delà de 16 classes et atteint un maximum de 95(%). Pour les MFCC on atteint un maximum correct de 94.17 (%) pour un nombre de classes égal à 64. Les coefficients LSP donnent toujours de bons résultats par rapport aux cœfficients MFCC.

Sur la figure 4.12 on a tracé le graphe avec une fréquence de 8 kHz. Le taux d'identification est croissant avec le nombre de centroides ; il devient presque stable au delà de 30 classes et atteint un maximum de 93(%) pour les LSP et 92(%) pour les MFCC. Les coefficients LSP donnent toujours de bons résultats par rapport aux cœfficients MFCC. Les performances des coefficients LSP se dégradent plus facilement que les coefficients MFCC.

Conclusion

Les performances du système LSP sont meilleures que celles de MFCC mais elles sont plus sensibles à la variation de fréquence.

b. Le mélange de gaussiennes (GMM)

Les résultats obtenus avec le 16KHz sont dans les tableaux ci-dessous :

nbr de classes	1	2	4	8	16	32	64
I_C (MFCC)	56.87	66.87	83.75	87.5	92.5	91.25	89.37
I_C (LSP)	78.75	85.62	90	93.75	94.37	95	92.5

Tableau 4.9 GMM ; Fe= 16 kHz
Ic : taux d'identification corrects

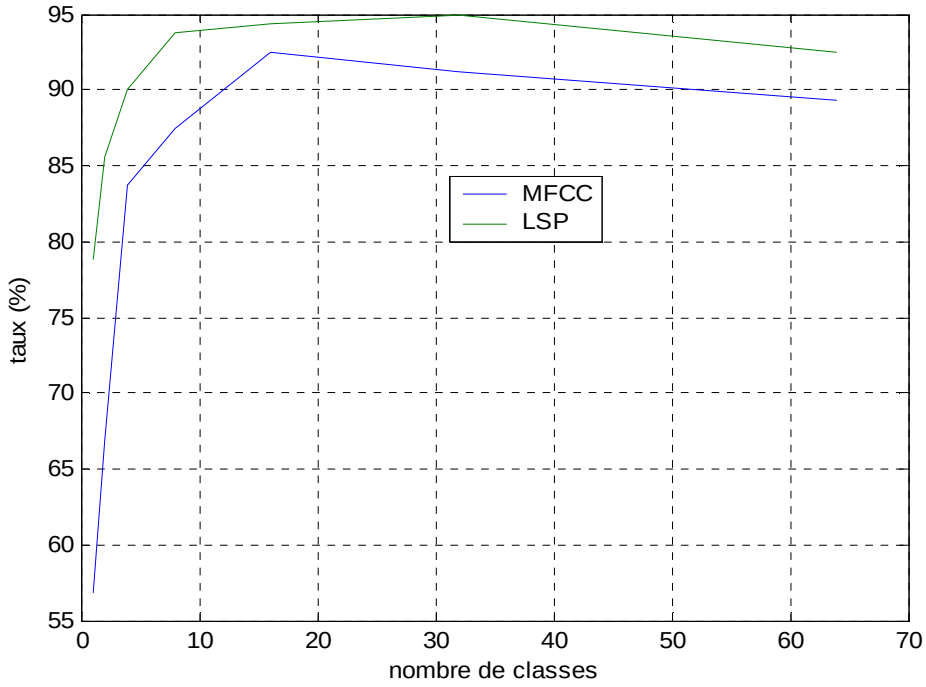


Fig. 4.13 GMM ; Fe= 16 kHz

La figure 4.13 donne le taux d'identification en fonction du nombre de centroides. Pour le système à 16 kHz le taux d'identification est croissant avec le nombre de centroides . Pour les MFCC il atteint la valeur maximale pour un nombre de centroides égal à 16 avec un taux de 92.5 (%). Pour les LSP elle atteint le taux maximal 94.37 (%) pour un nombre de gaussiennes de 32. ensuite le taux descend, pour atteindre la valeur 89.37 (%) pour les MFCC et 92.5 (%) pour les LPS.

Conclusion

- Le système LSP donne des taux d'identification meilleurs que ceux des MFCC
- Avec 16 kHz les coefficients MFCC donnent de bons résultats.

Les résultats obtenus avec le 8 KHz sont dans les tableaux ci-dessous :

nbr de classes	1	2	4	8	16	32	64
I_C (MFCC)	44.38	55.62	67.50	76.87	85.62	81.25	88.12
I_C (LSP)	71.25	75.62	83.12	94.37	93.12	90.25	88.12

Tableau 4.10 GMM ; Fe= 8 kHz
Ic : taux d'identification corrects

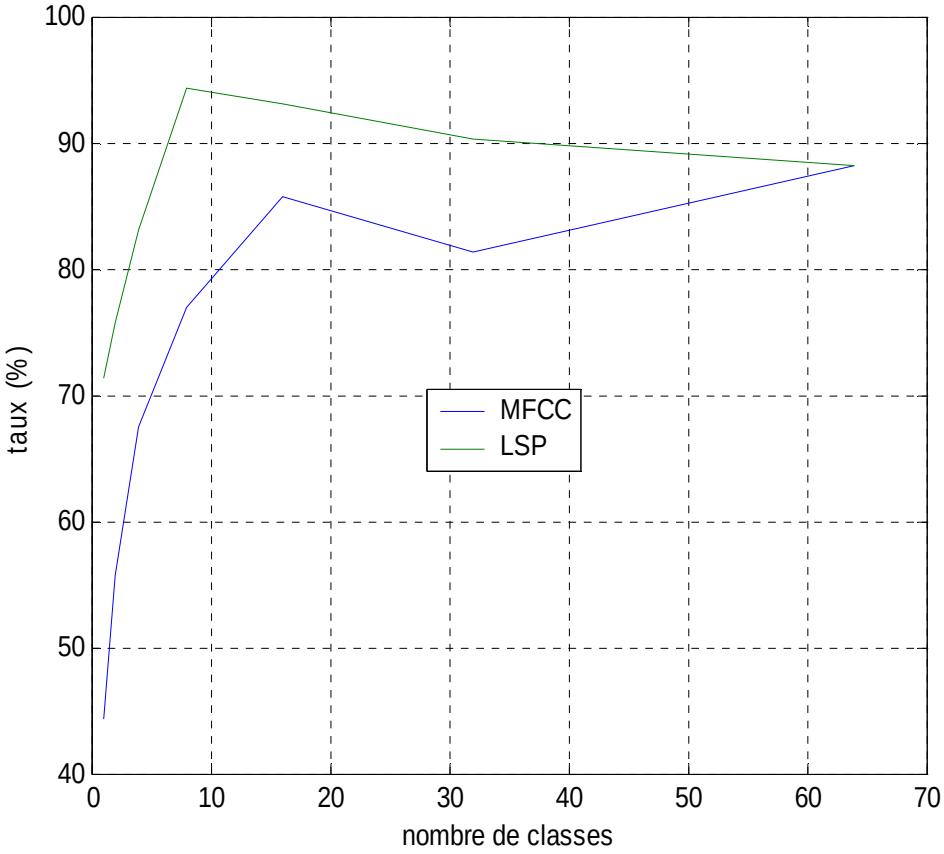


Fig. 4.14 GMM ; Fe= 8 kHz

Le système d'identification à base de GMM est sensible à la réduction de la fréquence. Le taux d'identification se dégrade jusqu'à 85.62 (%) avec les vecteurs MFCC. Une légère amélioration de l'identification est obtenue en utilisant les vecteurs LSP avec lequel le taux d'identification atteint 94.37 (%) en utilisant seulement 8 gaussiennes.

c. Le mélange de gaussiennes orthogonal (OGMM)

À la différence avec la GMM classique, les vecteurs d'entraînements subissent une orthogonalisation avant d'être utilisés pour générer les gaussiennes. Les résultats obtenus

avec 16KHz sont dans le tableau ci-dessous :

nbr de classes	1	2	4	8	16	32	64
I_C (MFCC)	50.62	56.25	62.5	75	80	81.25	77.5
I_C (LSP)	55	66.25	96.25	96.87	96.87	96.87	96.25

Tableau 4.11 OGMM ; Fe= 16 kHz

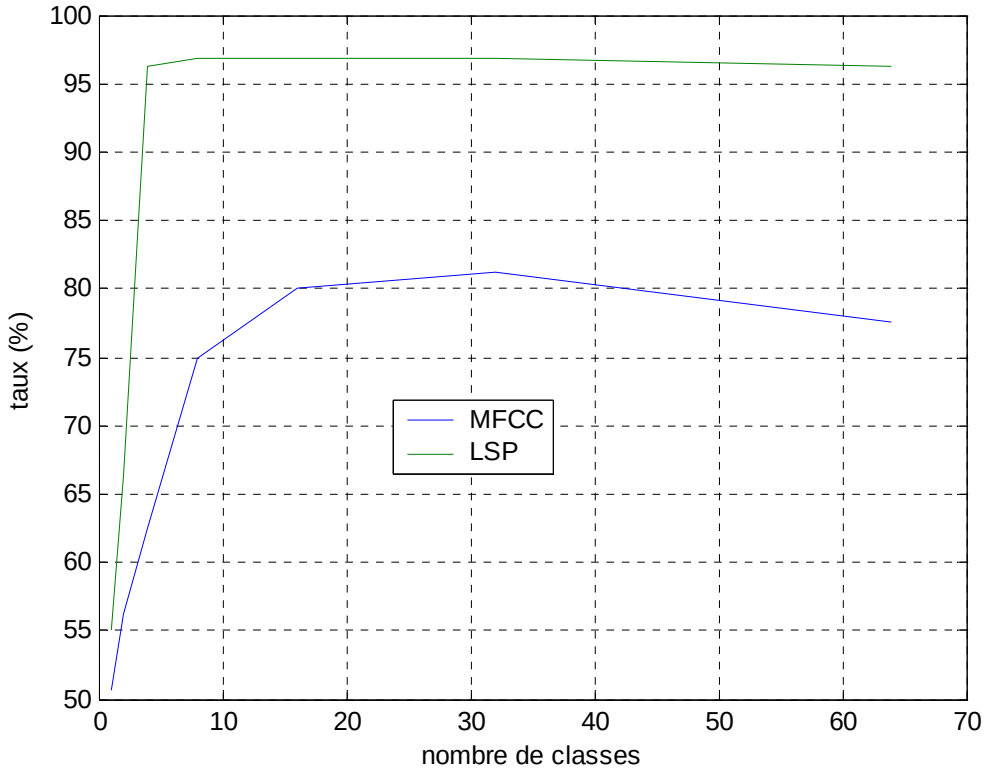


Fig. 4.15 OGMM ; Fe= 16 kHz

En utilisant les vecteurs MFCC, le taux d'identification est toujours inférieur à 81.25 (%), il est en moyenne de 81 (%) pour un nombre de classes de 16 à 32 et il se dégrade au delà de 32 vers une taux de 77.5 (%) avec une nombre de classes de 64.

Les meilleurs taux d'identification sont obtenus avec les cœfficients LSP avec 8 gaussiennes, il est suffisant pour donner un maximum local de 96.87 (%).

Les résultats obtenus avec le 8 KHz sont dans le tableau ci-dessous :

nbr de classes	1	2	4	8	16	32	64
I_C (MFCC)	14.37	25	34.37	40.62	40.62	53.12	55.62
I_C (LSP)	56.62	56.87	81.87	83.75	81.87	85.62	83.75

Tableau 4.12 OGMM ; Fe= 8 kHz

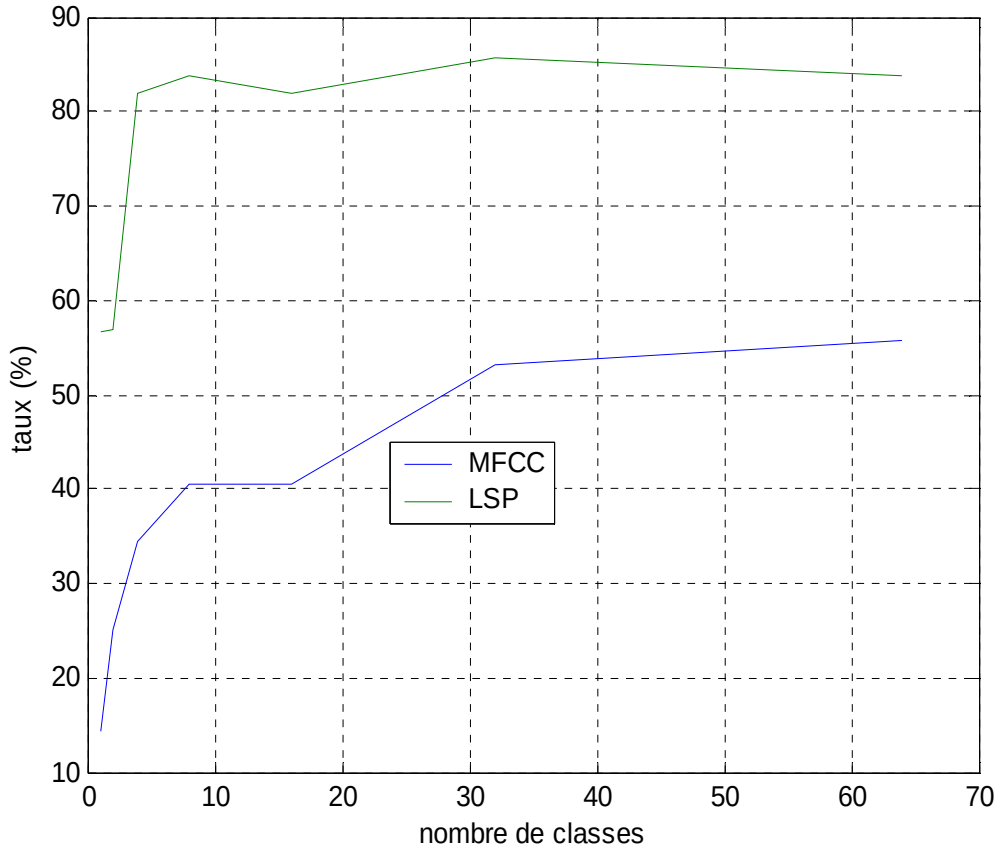


Fig. 4.16 OGMM ; Fe= 8 kHz

Pour Fe=8 kHz on obtient pour les LSP une taux de 85.62 (%) en utilisant 32 gaussiennes qui est un nombre relativement grand, nécessitant un espace mémoire important pour le stockage des références de chaque locuteur.

En utilisant les vecteurs MFCC, le taux d'identification est toujours inférieure à 56 (%), il est en moyenne de 55 (%) pour une nombre de classes de 32 à 64.

Conclusion

- Avec l'OGMM dans le cas où la fréquence est de 16khz , le temps de calcul est moins important par rapport à la GMM ; le nombre de gaussiennes nécessaires est égal à 8 en utilisant les coefficients LSP avec un taux d'identification de 96.87 (%).
- l'OGMM est plus sensible à la réduction de la fréquence d'échantillonnage par rapport à la GMM.

4.5.5 GMM pitch basé sur l'estimation de probabilité à postoriori

a. MFCC

nbr de classes	1	2	4	8	16	32	64
I_C (GMMpitch)	66.87	83.75	83.75	92.5	96.87	96.7	91.25
I_C (GMM)	40	56.25	73.75	85	89.37	85.62	81.87

Tableau 4.13 Comparaison entre GMM(MFCC) et GMM-Pitch(MFCC) **;**Fe=16 kHz

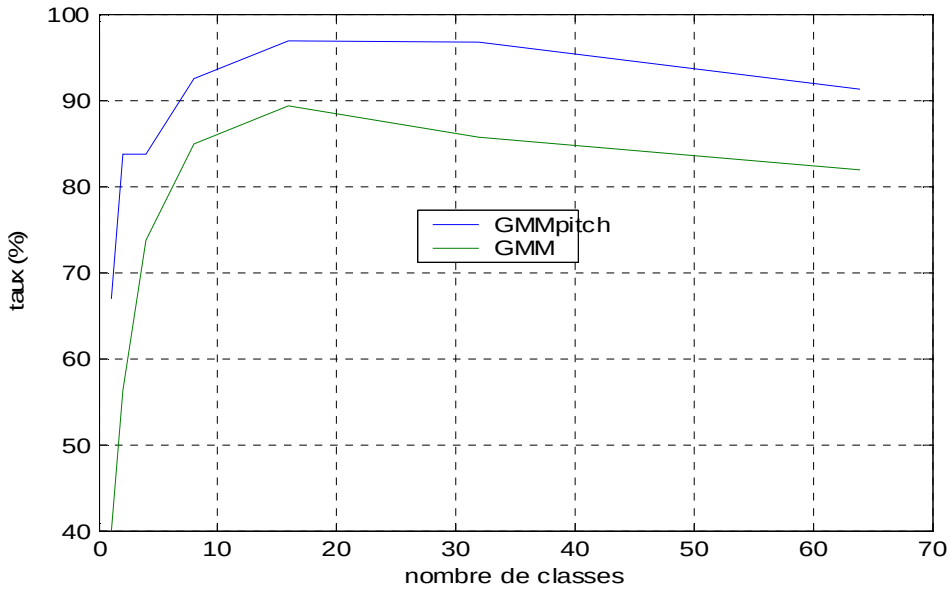


Fig. 4.17 Comparaison entre la GMM (MFCC) et la GMM-Pitch(MFCC) ;Fe=16 KHz

Pour les coefficients MFCC, la GMMpitch donne un taux d'identification meilleur que celui de GMM pour toutes les valeurs de nombre de classes. Les deux modèles atteignent le maximum pour un nombre de coefficients égal à 16 ; au delà duquel le taux d'identification diminue. La GMMpitch donne un taux d'identification maximum égal à 96,87 (%), par contre, la GMM donne un maximum de 89,37 (%).

b. LSP

nbr de classes	1	2	4	8	16	32	64
I_C (GMM)	78.75	85.62	90	93.75	94.37	95	92.5
I_C (GMMpitch)	80	85.62	91.25	96.25	96.25	95.62	95.62

Tableau 4.14 Comparaison entre la GMM(LSP) et la GMM-Pitch(LSP) ;Fe=16 KHz

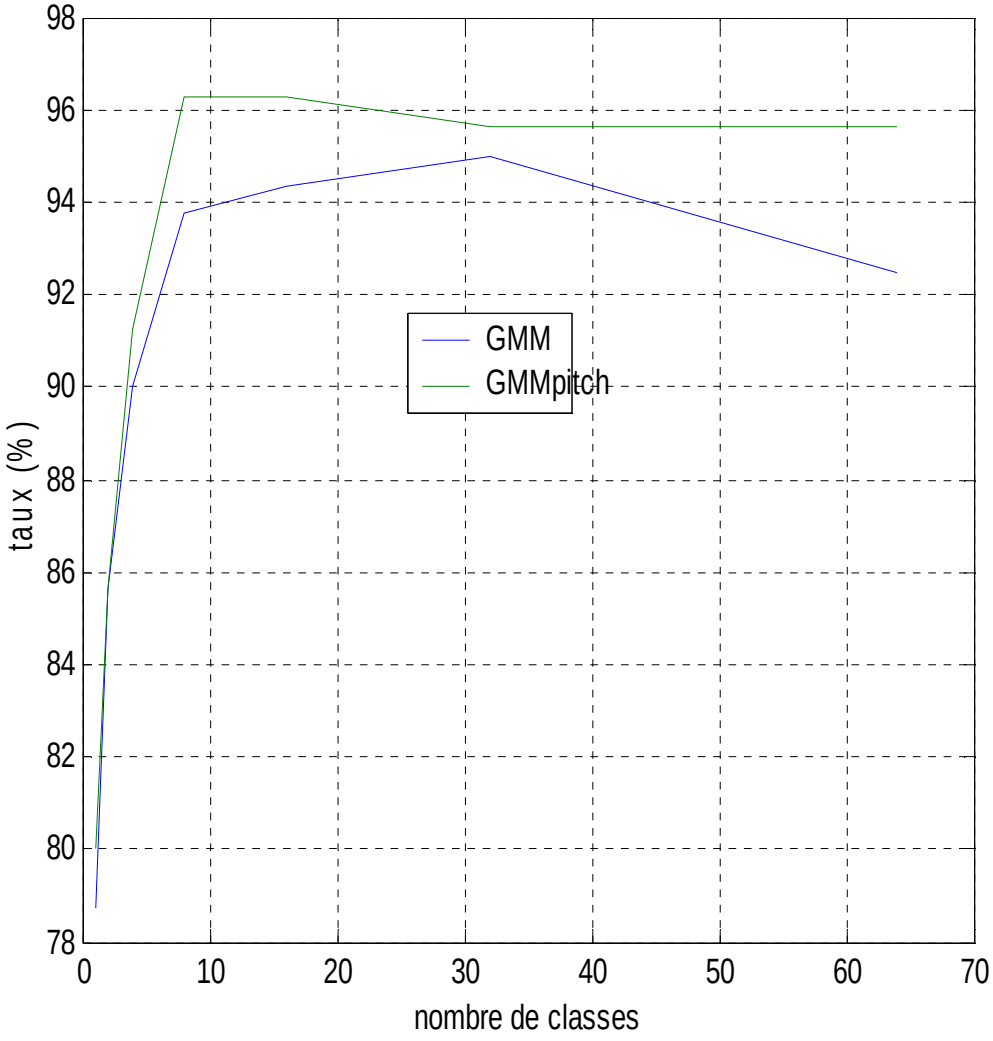


Fig. 4.18 Comparaison entre la GMM(MFCC) et la GMM-Pitch(MFCC) ;Fe=16 KHz

Pour les coefficients LSP la GMMpitch donne un taux d'identification meilleur que celui de GMM pour toutes les valeurs du nombre de classes.

Pour la GMMpitch le maximum est atteint à 96.25 (%) et cela pour un nombre de classes égal à 8, par contre pour la GMM le maximum est atteint à 95 (%) et cela pour un nombre de classes égal à 32.

Conclusion

Pour la GMMpitch, 8 classes suffisent pour modéliser un locuteur, par contre pour la GMM il faut 32 classes pour modéliser un locuteur. Le pitch améliore les performances d'identification.

d. comparaison entre les MFCC et les LSP pour la GMMpitch

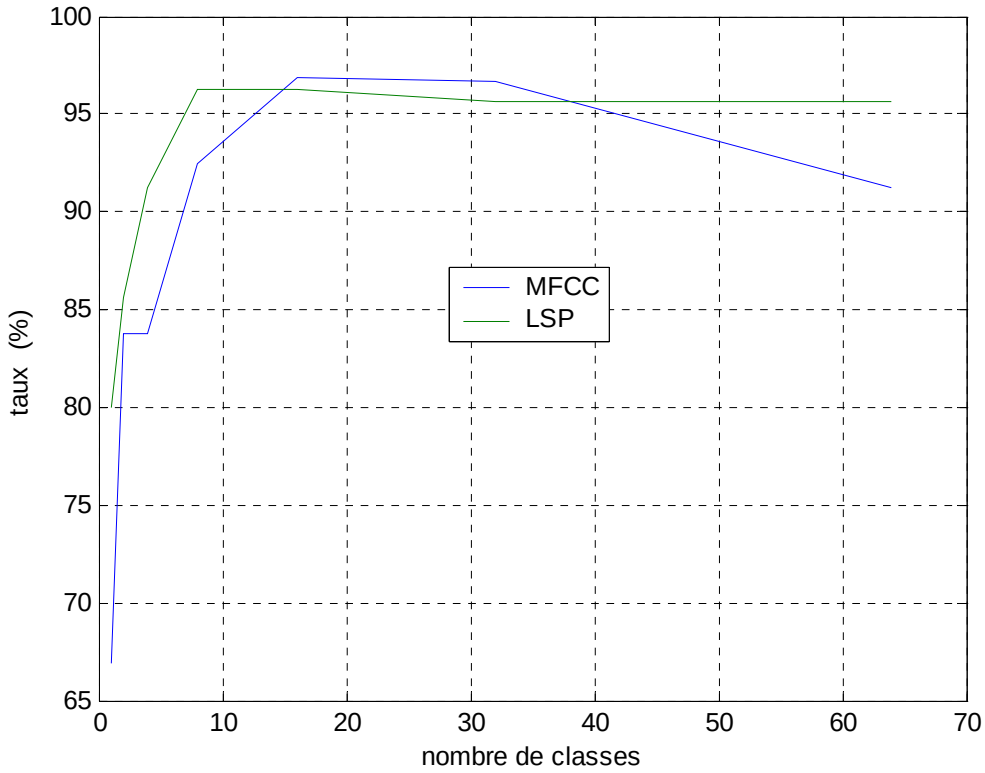


Fig. 4.19 GMMpitch ; Fe= 16 kHz

La figure 4.19 nous montre le taux d'identification en fonction du nombre de gaussiennes utilisées. On observe que le taux d'identification se dégrade au delà de 32 et qu'un nombre de 8 à 16 gaussiennes suffit pour représenter les locuteurs. Cette fois, l'espace de paramètres est partitionné en sous espaces dans lesquels la valeur du pitch est considérée constante. Donc le nombre de gaussiennes qui modélisent ces petits espaces est forcément plus petit que le nombre de gaussiennes nécessaire pour modéliser l'espace global.

4.5.6 Influence du nombre de locuteurs avec une classification VQ - MFCC - 16KHz

nbr de classes	1	2	4	8	16	32	64
I_C (16 locuteurs)	73.75	62.5	92.5	95	97.5	97.5	95
I_C (24 locuteurs)	75	70	90	95	96.67	96.67	95
I_C (32 locuteurs)	71.25	68.75	84.37	91.25	93.12	96.25	94.37

Tableau 4.15 influence du nombre de locuteurs

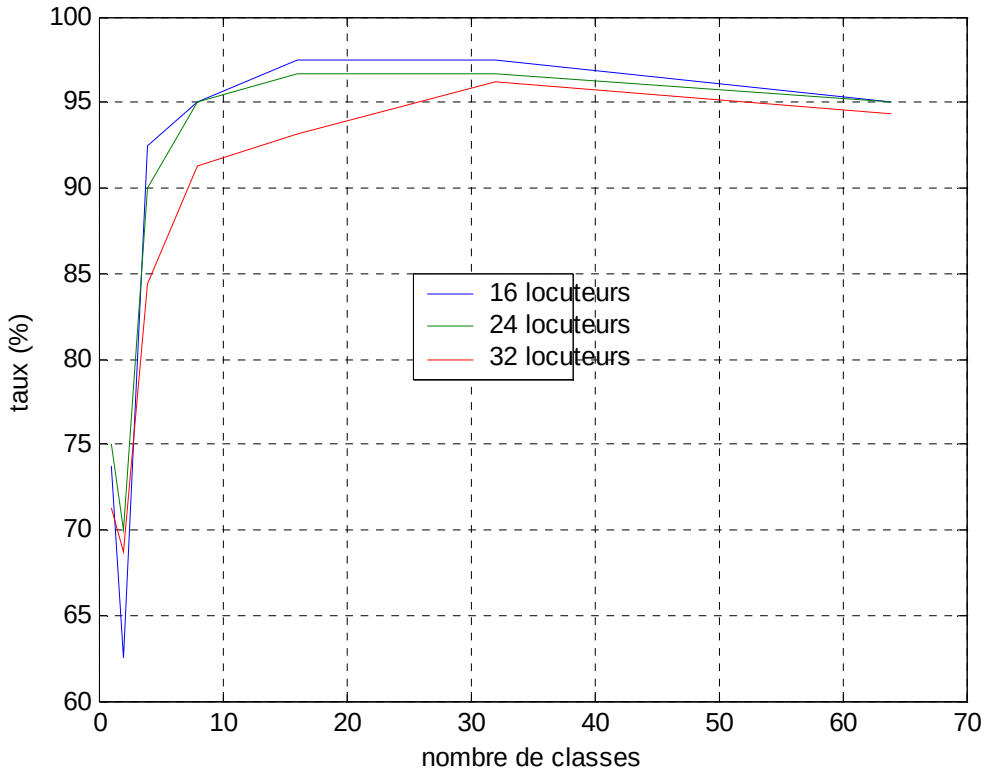


Fig. 4.20 influence du nombre de locuteurs

Commentaires

Sur les courbes de la figure 4.20, on remarque que l'ordre du modèle améliore le taux d'identification correct. Pour 16 et 24 locuteurs le maximum est atteint pour 16 coefficients (97.5, 96.67 (%) respectivement). Pour 32 locuteurs le maximum est atteint pour 32 coefficients (96.25 (%)). On observe encore que si le nombre des locuteurs augmente, le taux d'identification diminue parce que les similitudes entre locuteurs augmentent, ce qui accroît la probabilité de fausse identification. Ces accroissements se traduisent par une dégradation des performances.

4.5.7. Conclusion

- pour l'identification d'un locuteur, la capacité de discrimination des LSP est plus importante par rapport aux MFCC.
- la quantification vectorielle donne de bonnes performances en termes de taux d'identification correcte, mais elle est coûteuse en temps de calcul et espace mémoire nécessaire pour le stockage des références de locuteurs.
- le taux d'identification augmente avec la quantité de données d'entraînement. Il faut utiliser un nombre de coefficients supérieur à 10 pour avoir de bonnes

performances. Egalement lorsque le nombre de locuteurs augmente, il faut augmenter l'ordre du modèle.

- pour remédier aux problèmes de la GMM, on introduit l'OGMM (Orthogonal GMM) qui réduit considérablement le temps de calcul et l'espace mémoire requis. Il donne de bonnes performances avec les coefficients LSP, mais il est plus sensible à la réduction de la fréquence d'échantillonnage par rapport à la GMM.
- l'utilisation du pitch basée sur l'estimation de la probabilité à postoriori permet d'augmenter le taux d'identification.
- les résultats obtenus avec la base TIMIT sont meilleurs que ceux obtenus avec notre base et cela est dû principalement aux conditions d'expérimentations.

Conclusion générale

Au cours de ce travail, nous avons traité le problème de l'identification du locuteur en mode indépendant du texte utilisant notre base de données. Elle consiste à extraire des vecteurs de paramètres à partir des signaux de paroles prononcés par les locuteurs concernés, qui servent à l'entraînement des modèles mathématiques caractérisants la voix de chaque locuteur. Egalement nous avons entrepris une comparaison entre les différents paramètres qui sont les MFCC et LSP, suivant la modélisation par GMM, OGMM, QV et GMMpitch.

En ce qui concerne la VQ et à partir des expériences effectuées nous avons abouti aux conclusions suivantes

- elle donne de bons taux d'identification surtout avec les cœfficients LSP,
- elle est coûteuse en terme de temps de calcul et d'espace mémoire nécessaire.

En ce qui concerne la GMM nous avons abouti aux conclusions suivantes

- elle donne de bons taux d'identification avec les cœfficients MFCC,
- elle est très sensible à la variation de fréquence.

En ce qui concerne l'OGMM nous avons abouti aux conclusions suivantes

- elle donne de bons taux d'identification surtout avec les cœfficients LSP (un nombre de classes réduit permet de modéliser le système),
- le temps du calcul est moins important par rapport à la GMM,
- l'OGMM donne des performances meilleures que la GMM avec les coefficients LSP,
- l'OGMM est plus sensible à la réduction de la fréquence d'échantillonnage par rapport à la GMM.

En ce qui concerne la GMMpitch nous avons abouti aux conclusions suivantes

- la GMMpitch donne des performances meilleures que la GMM,
- le temps du calcul est moins important par rapport à la GMM.

Nous avons effectué encore un certain nombre d'expériences où nous avons examiné l'influence d'un certains nombres de paramètres sur le taux d'identification correct, et à partir duquel nous avons abouti aux conclusions suivantes

- le taux d'identification augmente avec la quantité de données d'entraînement,
- il faut augmenter l'ordre du modèle lorsque le nombre de locuteurs augmente pour avoir des bonnes performances. Egalement il faut utiliser un nombre de coefficients supérieur à 10.

En ce qui concerne la comparaison entre notre base et la base TIMIT, dans tous les cas les résultats obtenus avec la base TIMIT sont meilleurs que ceux obtenus avec notre base et cela est dû principalement aux conditions d'expérimentation.

Annexe

Résultats des évaluations expérimentales

Cette annexe expose les résultats intermédiaires de toutes les expériences effectuées le long de ce travail. 5 segments de tests sont utilisés pour chaque locuteur et les tableaux suivants donnent le nombre de segments correctement identifiés sur l'ensemble des segments de tests.

1. La base TIMIT

Nombre de locuteurs 32

Nombre de cœfficients 12

1.1 Fe=16KHz - MFCC - VQ

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	3	4	3	5	5	5	5
2	1	5	5	5	5	5	5
3	5	5	5	5	5	5	5
4	0	4	5	4	4	5	5
5	2	4	5	4	4	5	5
6	0	2	3	2	3	5	5
7	4	5	4	4	5	5	5
8	1	4	4	3	5	5	5
9	2	5	3	5	5	5	5
10	2	5	4	5	5	5	5
11	1	3	4	4	4	4	5
12	2	5	3	5	5	5	5
13	3	5	5	5	5	5	5
14	1	4	2	5	5	5	5
15	1	4	5	5	5	5	5
16	0	3	4	4	5	5	5

17	0	4	3	3	3	4	5
18	1	5	5	5	5	5	5
19	1	2	4	5	4	5	5
20	5	4	5	5	5	5	5
21	2	3	4	3	5	5	5
22	3	4	4	4	4	5	5
23	0	2	5	5	5	5	5
24	2	4	3	5	5	5	5
25	0	1	4	4	4	5	5
26	2	5	5	5	5	5	5
27	4	4	4	3	4	5	5
28	2	5	5	5	5	5	5
29	5	2	4	3	5	5	5
30	3	4	4	5	4	4	5
31	4	5	3	4	5	5	5
32	0	3	1	2	4	4	3
I_C (%)	38.75	77.5	79.37	85.62	92.5	97.5	98.75

1.2 Fe=16KHz -LSP -VQ

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	4	5	5	5	5	5	5
2	1	5	5	5	5	5	5
3	5	4	5	5	5	5	5
4	1	4	4	5	5	5	5
5	1	5	4	3	4	5	5
6	0	3	2	4	4	4	5
7	4	5	5	5	5	5	5
8	1	5	5	5	5	5	5
9	2	3	5	5	5	5	5
10	2	5	5	5	5	5	5
11	0	3	3	4	4	4	5
12	3	3	5	5	5	4	5
13	3	5	5	5	5	5	5
14	3	5	5	5	5	5	5
15	4	5	5	5	5	5	5
16	1	4	5	5	5	5	5
17	3	4	4	4	5	5	5
18	0	4	5	5	5	5	5
19	1	4	4	4	5	5	5
20	5	5	4	5	5	5	5
21	4	4	5	5	5	5	5
22	4	5	5	5	5	5	5
23	1	3	5	5	5	5	5
24	2	3	5	5	5	5	5
25	0	5	4	5	5	5	5
26	3	5	5	5	5	5	5
27	3	3	5	3	4	5	5
28	3	5	5	5	5	5	5

29	0	2	3	3	4	5	5
30	1	3	5	4	4	5	5
31	4	4	5	5	5	5	5
32	3	2	2	3	4	4	0
I_C (%)	45.62	81.25	90	93	95.62	97.5	96.87

1.3 Fe=16KHz - MFCC - GMM

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	2	3	3	5	5	5	5
2	4	4	5	5	5	5	5
3	0	0	1	4	4	5	5
4	2	5	5	5	5	4	5
5	0	1	3	2	2	2	2
6	4	4	5	5	5	5	5
7	4	4	5	4	5	4	4
8	5	5	5	5	5	5	5
9	4	5	5	5	5	5	5
10	5	4	4	2	3	3	3
11	2	4	4	4	4	5	4
12	0	1	3	4	5	4	5
13	0	0	3	4	4	5	3
14	2	4	4	5	5	5	5
15	0	1	3	5	5	5	5
16	2	4	4	4	4	4	4
17	2	4	4	5	5	5	5
18	0	2	4	5	5	4	4
19	5	5	5	5	5	5	5
20	4	4	4	4	4	2	3
21	5	5	5	5	5	5	5
22	5	5	5	5	5	5	5
23	5	5	5	5	5	5	5
24	1	2	3	2	4	4	3
25	5	5	5	5	5	5	5
26	4	3	4	5	5	5	4
27	4	3	5	4	5	5	5
28	5	5	5	5	5	5	5
29	1	0	5	5	5	5	5
30	5	5	5	5	5	5	5
31	4	5	5	5	5	5	5
32	0	0	3	2	4	5	4
I_C (%)	56.87	66.87	83.75	87.5	92.5	91.25	89.37

1.4 F=16KHz - LSP -GMM

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

2	5	5	5	5	5	5	5
3	5	5	5	5	5	5	5
4	4	5	5	5	5	5	5
5	4	4	4	5	4	5	5
6	4	2	5	5	5	5	3
7	5	5	5	5	5	5	5
8	5	5	5	5	5	5	5
9	3	5	5	5	5	5	5
10	5	5	5	5	5	5	5
11	3	4	4	4	5	5	5
12	4	5	5	5	5	5	4
13	4	5	5	5	5	5	5
14	4	5	5	5	5	5	5
15	5	5	5	5	5	5	5
16	4	5	5	5	5	5	5
17	5	5	4	4	5	5	5
18	3	5	5	5	4	5	4
19	3	4	5	5	5	5	5
20	3	5	5	4	5	4	3
21	4	5	5	5	5	5	5
22	5	5	5	5	5	5	5
23	2	3	5	5	5	5	5
24	3	5	5	5	5	5	5
25	3	4	5	5	5	4	4
26	5	5	5	5	5	5	5
27	5	5	5	5	5	5	5
28	2	5	5	5	5	5	5
29	5	5	5	5	5	5	5
30	5	5	4	4	5	5	4
31	5	5	5	5	5	5	5
32	3	4	3	3	5	5	4
I_C (%)	81.25	93.75	96.25	96.25	98.75	98.75	94.37

2. Notre base de données

Nombre de locuteurs 32

Nombre de cœfficients 12

2.1 Fe=16KHz - MFCC -VQ

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	5	0	5	5	5	5	5
2	3	5	5	4	4	5	5
3	4	0	4	5	5	5	5
4	3	5	5	5	5	5	5
5	3	2	4	4	5	4	4
6	1	1	3	5	5	5	4
7	4	2	4	4	4	5	5

8	2	0	5	5	5	5	5
9	4	5	5	5	5	5	5
10	4	2	4	5	5	5	3
11	4	4	4	4	4	5	4
12	2	4	4	5	5	5	5
13	4	4	5	5	5	5	5
14	3	5	5	5	5	5	5
15	5	5	4	4	5	5	5
16	1	3	5	4	5	5	5
17	4	4	4	5	5	4	5
18	5	5	5	5	5	5	5
19	5	4	3	5	4	4	4
20	4	4	4	4	4	4	4
21	4	1	0	0	0	3	5
22	5	5	5	5	5	5	5
23	3	4	3	5	5	5	5
24	5	5	5	5	5	5	5
25	5	4	5	5	5	5	5
26	4	5	5	5	5	5	5
27	3	4	5	5	5	5	5
28	4	4	5	5	5	5	5
29	3	3	3	5	5	5	5
30	5	5	5	5	5	5	5
31	0	4	5	5	5	5	5
32	3	2	2	3	4	5	3
I_C (%)	71.25	68.75	84.37	91.25	93.12	96.25	94.37

2.2 Fe=16KHz - LSP -VQ

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	5	2	4	4	4	5	5
2	4	5	4	5	5	4	4
3	2	3	4	4	5	5	5
4	2	2	4	5	5	5	5
5	2	5	4	4	4	4	5
6	2	3	3	5	5	5	5
7	5	5	5	5	5	5	5
8	3	4	5	5	5	5	5
9	0	5	5	5	5	5	5
10	3	3	3	4	4	4	5
11	0	1	4	5	4	4	4
12	3	5	4	5	5	5	5
13	5	4	5	5	5	5	5
14	4	5	5	5	5	5	5
15	0	4	5	5	5	5	5
16	5	2	2	3	4	4	5
17	5	5	5	5	5	5	5
18	4	3	5	5	5	5	5
19	4	4	5	3	4	4	3

20	3	4	4	3	4	4	4
21	1	4	5	5	5	4	4
22	3	5	5	5	5	5	5
23	4	4	5	5	5	5	5
24	5	5	5	5	5	5	5
25	3	5	5	5	5	5	5
26	2	4	5	4	5	5	5
27	5	4	5	5	5	5	5
28	1	5	5	5	5	5	5
29	5	2	4	5	5	5	5
30	4	5	5	5	5	5	5
31	2	4	4	5	5	5	5
32	4	3	3	4	4	4	4
I_C (%)	62.5	77.5	88.12	92.5	95	94.37	95.62

2.3 Fe=16KHz - MFCC -GMM

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	5	2	2	5	5	5	5
2	1	5	5	5	5	5	5
3	3	0	5	5	5	5	5
4	5	5	5	4	5	5	4
5	2	3	5	5	4	5	5
6	3	4	5	5	5	5	5
7	5	5	4	4	5	5	5
8	4	5	5	5	5	5	4
9	3	4	5	5	5	5	5
10	5	5	5	5	5	5	5
11	0	0	3	1	3	4	2
12	4	4	3	3	4	3	4
13	0	0	3	5	4	3	4
14	2	1	0	5	5	5	5
15	1	5	4	5	5	5	5
16	0	3	5	5	5	5	4
17	0	2	1	3	2	4	2
18	4	1	5	5	5	4	5
19	1	5	4	5	5	4	5
20	1	4	5	5	5	5	4
21	0	2	1	5	5	5	5
22	0	5	5	5	5	5	5
23	5	2	4	4	4	3	4
24	4	4	4	3	5	3	4
25	0	1	3	3	3	2	2
26	0	0	1	2	4	3	2
27	1	0	5	5	5	5	5
28	2	4	3	4	4	5	4
29	0	2	3	3	4	1	2
30	0	2	3	4	4	3	3

31	2	5	5	5	5	5	5
32	1	0	2	3	3	5	2
I_C (%)	40	56.25	73.75	85	89.37	85.62	81.87

2.4 Fe=16KHz - LSP -GMM

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	5	5	5	5	5	5	5
2	3	3	4	4	4	4	4
3	4	4	4	5	5	5	4
4	5	5	5	5	5	5	5
5	4	2	5	5	5	5	5
6	3	5	5	5	5	5	5
7	4	5	4	5	5	5	5
8	5	5	5	5	5	5	5
9	4	2	5	5	5	5	5
10	2	1	0	0	0	2	1
11	4	4	4	4	4	4	4
12	4	5	5	5	5	5	5
13	4	5	5	5	5	5	5
14	4	4	4	5	5	5	5
15	4	5	4	5	5	5	5
16	3	2	2	5	5	5	4
17	4	4	5	5	5	5	5
18	5	5	5	4	5	5	5
19	5	5	5	5	5	5	5
20	4	4	4	4	4	4	4
21	4	4	5	5	5	5	5
22	4	5	5	5	5	5	5
23	5	5	5	5	5	5	5
24	5	5	5	5	5	5	5
25	5	5	5	5	5	5	5
26	4	5	5	5	5	5	5
27	3	4	5	5	5	5	5
28	4	4	4	5	5	5	5
29	2	5	5	5	5	4	3
30	5	5	5	5	5	5	5
31	3	5	5	5	5	5	5
32	2	5	5	4	4	4	4
I_C (%)	78.75	85.62	90	93.75	94.37	95	92.5

2.5 Fe=16KHz -MFCC -OGMM

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

3	3	4	1	3	4	5	4
4	0	5	3	4	5	4	4
5	2	3	1	2	2	3	3
6	1	2	3	3	4	4	4
7	3	1	5	1	5	5	5
8	1	1	3	4	5	5	5
9	3	2	0	4	1	2	1
10	5	5	3	2	3	4	3
11	3	3	0	4	5	5	5
12	0	1	4	5	1	2	1
13	3	0	4	2	4	4	4
14	4	2	4	4	4	4	4
15	5	5	0	5	3	4	3
16	3	5	5	5	4	3	4
17	4	5	5	5	5	5	5
18	3	5	4	4	5	5	5
19	1	0	4	5	5	4	3
20	2	3	3	1	3	3	4
21	5	0	1	5	5	4	4
22	3	5	3	5	3	4	4
23	4	3	4	5	5	5	5
24	4	4	4	5	5	5	5
25	0	5	4	5	5	5	5
26	3	1	2	4	4	4	4
27	4	2	5	5	4	3	3
28	4	4	1	5	4	4	4
29	2	1	5	3	3	3	3
30	2	2	5	4	4	4	4
31	2	3	3	3	5	3	4
32	0	1	2	2	2	2	1
I_C (%)	50.62	56.25	62.5	75	80	81.25	77.5

2.6 Fe=16KHz - LSP - OGMM

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	5	5	5	4	5	5	4
2	5	3	5	5	5	5	5
3	4	2	5	5	4	5	5
4	4	4	5	5	5	5	5
5	2	3	4	4	5	4	5
6	2	3	5	5	5	5	5
7	1	3	5	5	5	5	5
8	1	4	5	5	5	5	5
9	2	4	5	5	5	5	5
10	5	3	5	5	5	5	5
11	3	5	5	5	5	5	5
12	1	5	5	5	5	5	5
13	0	1	5	5	5	5	5
14	2	3	5	5	5	5	5

15	5	2	5	5	5	5	5
16	5	5	5	5	5	5	5
17	5	3	4	4	4	4	5
18	5	3	5	5	5	5	5
19	0	4	5	5	5	5	5
20	2	1	4	4	4	4	5
21	1	5	5	5	5	5	5
22	4	5	5	5	5	5	5
23	4	5	5	5	5	5	5
24	4	2	4	4	4	4	4
25	5	5	5	5	5	5	5
26	1	4	5	5	5	5	5
27	2	5	5	5	5	5	5
28	4	3	5	5	5	5	5
29	1	1	5	5	5	5	5
30	2	4	5	5	5	5	5
31	4	4	5	5	5	5	5
32	0	1	4	4	4	4	4
I_C (%)	55	66.25	96.87	96.25	96.87	96.87	96.25

2.7 Fe=8KHz – MFCC-VQ

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	4	3	5	4	5	5	5
2	4	4	3	4	5	5	5
3	4	4	2	2	4	4	4
4	4	3	4	5	3	4	4
5	1	2	2	2	2	2	2
6	1	0	3	5	3	5	5
7	1	5	2	2	4	5	5
8	1	1	4	5	4	5	5
9	2	2	4	1	4	5	4
10	5	0	3	3	5	5	4
11	3	1	4	4	4	4	5
12	3	4	4	5	4	5	5
13	2	1	1	4	5	5	5
14	0	5	3	5	5	5	5
15	2	3	2	2	5	5	4
16	4	3	4	4	5	4	5
17	5	4	4	4	4	5	5
18	5	1	3	4	4	5	5
19	5	4	4	5	4	5	4
20	1	2	1	3	5	5	5
21	2	1	5	5	5	3	5
22	4	5	5	5	5	5	5
23	3	5	5	5	5	5	3
24	4	3	2	3	4	5	5
25	5	3	5	5	5	4	5
26	1	1	4	5	5	5	5

27	2	5	5	5	5	5	5
28	4	3	1	5	5	5	5
29	1	0	2	4	5	5	5
30	2	5	4	5	5	5	5
31	3	5	5	5	5	5	5
32	1	0	1	1	2	2	4
I_C (%)	55.62	55.00	66.25	78.75	87.50	91.88	92.50

2.8 Fe=8KHz - LSP -VQ

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	5	2	5	5	5	5	5
2	1	4	3	1	3	5	5
3	4	4	2	3	5	3	4
4	4	4	4	4	3	5	5
5	2	2	3	3	5	4	4
6	2	2	3	4	5	5	5
7	3	3	4	4	5	5	5
8	1	3	3	5	5	5	5
9	2	5	4	5	4	5	5
10	4	0	3	3	4	4	3
11	5	3	4	4	5	4	4
12	5	5	5	5	5	5	5
13	2	5	4	5	5	5	5
14	0	2	4	5	2	5	5
15	3	3	2	1	4	5	5
16	3	1	4	4	5	4	4
17	5	5	5	5	4	5	5
18	2	4	3	4	4	4	4
19	5	4	4	4	5	5	4
20	0	1	2	4	4	4	4
21	2	3	5	5	5	5	4
22	5	4	4	5	5	5	5
23	3	5	5	5	5	5	5
24	3	2	3	3	4	4	4
25	5	5	4	5	5	5	5
26	1	1	5	4	4	5	5
27	3	5	5	5	5	5	5
28	4	5	3	5	5	5	5
29	2	2	5	5	4	4	5
30	2	4	5	4	5	5	5
31	2	5	2	5	5	5	5
32	2	1	2	4	3	3	5
I_C (%)	54.38	65.00	74.38	83.12	88.75	92.50	93.12

2.9 Fe=8KHz – MFCC-GMM

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	0	1	2	3	3	3	3
2	0	3	4	5	5	4	5
3	0	0	0	0	3	3	2
4	3	4	4	4	5	3	4
5	0	0	0	0	0	1	2
6	5	5	3	5	5	5	5
7	0	2	4	5	5	3	5
8	4	5	5	5	5	5	5
9	2	3	5	5	5	5	4
10	4	3	3	3	5	3	3
11	3	2	4	4	4	4	4
12	0	1	2	3	3	4	5
13	0	0	1	2	3	3	5
14	1	4	4	4	5	5	5
15	2	2	2	3	4	4	4
16	1	4	4	4	4	5	5
17	3	3	4	5	5	4	5
18	0	2	2	5	5	4	5
19	4	5	5	5	5	5	5
20	0	1	0	2	2	3	3
21	5	5	5	5	5	5	5
22	5	5	5	5	5	5	5
23	2	4	5	5	5	5	5
24	0	0	2	1	4	2	2
25	3	5	5	5	5	5	5
26	4	2	5	5	5	5	5
27	5	5	5	5	5	5	5
28	3	2	4	5	5	5	5
29	2	1	5	4	5	5	5
30	5	5	4	5	5	5	5
31	5	5	5	5	5	5	5
32	0	0	0	1	2	2	5
I_C (%)	44.38	55.62	67.50	76.87	85.62	81.25	88.12

2.10 Fe=8KHz - LSP - GMM

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	5	5	5	5	5	5	5
2	4	3	3	5	5	5	5
3	4	3	5	5	5	1	1
4	4	4	4	4	5	5	5
5	2	2	2	5	5	5	5
6	2	2	1	5	5	5	5
7	2	2	4	5	5	5	5
8	4	4	5	5	5	5	5

9	3	4	4	5	5	5	5
10	4	2	4	5	2	4	0
11	1	4	4	4	4	4	3
12	4	5	5	4	5	4	5
13	5	2	2	5	5	4	5
14	4	4	5	5	5	4	5
15	4	5	4	5	5	4	5
16	3	4	4	5	4	5	4
17	2	5	4	4	5	5	5
18	4	4	4	5	5	5	4
19	5	5	5	5	5	5	5
20	5	1	2	5	2	3	2
21	2	5	5	2	5	5	5
22	4	5	5	5	5	5	5
23	5	5	5	5	5	5	5
24	5	5	4	5	4	3	3
25	5	5	5	4	5	5	5
26	2	4	5	5	5	5	5
27	4	5	5	5	5	5	5
28	4	5	5	5	5	5	5
29	3	3	4	5	4	4	4
30	4	4	5	4	5	5	5
31	3	3	5	5	5	5	5
32	2	2	4	5	4	5	5
I_C (%)	71.25	75.62	83.12	94.37	93.12	90.25	88.12

2.11 Fe=8KHz - MFCC - OGMM

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	1	1	1	1	1	2	2
2	0	1	1	2	2	2	2
3	0	0	0	0	0	0	0
4	1	2	1	3	3	2	3
5	2	2	3	2	2	0	1
6	3	5	5	5	5	5	5
7	0	1	2	5	5	2	1
8	1	1	3	5	5	5	5
9	2	3	5	4	4	4	4
10	1	2	4	2	2	2	3
11	0	0	0	1	1	2	2
12	0	0	0	2	2	0	1
13	0	0	0	0	0	0	0
14	0	2	4	2	2	4	5
15	0	0	0	1	1	1	1
16	0	0	2	0	0	4	4
17	1	0	1	0	0	2	3
18	0	0	0	2	3	3	2
19	0	0	0	3	2	5	4
20	1	1	2	1	1	1	2

21	1	2	1	2	3	5	5
22	0	2	1	2	1	5	5
23	3	4	5	5	4	3	3
24	1	2	1	4	4	0	1
25	4	5	5	5	5	5	5
26	0	0	0	0	0	2	2
27	0	0	1	1	1	4	5
28	0	1	2	2	1	3	2
29	0	0	0	0	0	2	2
30	0	1	0	0	1	4	4
31	1	2	2	2	3	2	4
32	0	0	1	1	1	1	1
I_C (%)	14.37	25	34.37	40.62	40.62	53.12	55.62

2.12 Fe=8KHz -LSP - OGMM

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	2	3	4	5	5	5	5
2	4	3	4	4	5	0	5
3	0	0	5	5	4	5	5
4	1	2	3	5	5	5	5
5	1	0	3	4	3	3	3
6	4	5	4	4	5	5	4
7	4	3	4	5	4	5	5
8	5	5	3	5	5	5	5
9	4	4	4	5	5	2	5
10	5	5	5	5	5	5	5
11	2	1	5	3	4	4	4
12	0	1	4	4	4	5	5
13	0	0	5	5	5	5	5
14	2	1	5	5	5	4	5
15	0	1	5	5	5	5	5
16	3	3	1	3	5	1	4
17	1	1	5	5	5	5	5
18	1	2	4	4	5	5	5
19	4	3	3	5	2	5	5
20	4	4	5	3	4	5	2
21	5	5	5	5	5	2	5
22	5	5	5	5	5	5	5
23	5	4	5	5	5	5	5
24	1	2	5	3	5	5	4
25	5	5	5	5	5	5	5
26	4	4	5	5	4	5	4
27	4	4	5	5	5	4	5
28	5	4	1	5	5	4	2
29	1	0	5	5	5	4	5
30	5	5	1	2	3	5	1
31	3	4	4	5	4	4	5

32	1	2	4	4	4	5	3
I_C (%)	56.62	56.87	81.87	83.75	81.87	85.62	83.75

3. Influence du nombre de cœfficients

Nous utilisons notre base de données

Nombre de classes 16

Fe=16KHz -MFCC - VQ

	Nombre de cœfficients						
locuteur	1	5	10	15	20	25	40
1	0	3	5	5	5	5	5
2	0	5	4	5	5	5	5
3	0	4	5	5	5	5	5
4	1	4	5	5	5	5	5
5	0	3	4	5	5	5	5
6	0	5	5	5	5	5	5
7	1	3	3	4	4	4	4
8	0	3	5	5	5	5	5
9	3	4	5	5	5	5	5
10	0	5	4	4	4	5	5
11	1	5	4	5	4	4	4
12	0	4	5	5	5	4	5
13	0	5	5	5	5	5	5
14	0	5	5	5	5	5	5
15	1	3	5	5	5	5	5
16	1	3	5	5	5	5	5
17	2	5	4	5	5	5	5
18	1	4	5	5	5	5	5
19	1	3	4	4	4	4	5
20	0	5	4	4	4	5	4
21	4	5	5	5	5	5	5
22	2	5	5	5	5	5	5
23	0	3	5	5	5	5	5
24	0	5	5	5	5	5	5
25	0	5	5	5	5	5	5
26	2	5	5	5	5	5	5
27	5	5	5	5	5	5	5
28	2	3	5	5	5	5	5
29	0	5	5	5	5	5	5
30	1	2	5	5	5	5	5
31	1	4	4	4	5	5	5
32	2	3	3	3	5	5	5
I_C (%)	19.37	81.87	92.5	95.62	96.87	97.5	98.12

4. Fe=16 kHz - GMMpitch basé sur la probabilité à postoriori

Nous utilisons notre base de donnée

4.1 MFCC

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	3	3	3	5	5	5	5
2	3	4	5	5	5	5	5
3	0	0	0	4	4	3	4
4	2	5	5	5	4	4	4
5	0	1	3	2	3	2	3
6	4	4	5	5	5	5	5
7	4	4	5	4	5	4	5
8	5	5	5	5	5	5	5
9	4	5	5	5	5	5	5
10	5	5	5	5	5	5	5
11	2	4	4	4	4	4	4
12	0	1	3	4	5	5	5
13	0	0	3	4	4	4	4
14	2	4	4	5	5	5	5
15	0	1	3	5	5	5	5
16	2	4	4	4	4	4	4
17	2	4	4	5	5	5	5
18	0	2	4	5	5	5	5
19	5	5	5	5	5	5	5
20	4	4	4	4	4	4	4
21	5	5	5	5	5	5	5
22	5	5	5	5	5	5	5
23	5	5	5	5	5	5	5
24	1	2	3	2	4	4	2
25	5	5	5	5	5	5	5
26	4	3	4	5	4	5	4
27	4	3	5	4	5	4	5
28	5	5	5	5	5	5	5
29	1	0	5	5	5	5	5
30	5	5	5	5	5	5	5
31	4	5	5	5	5	5	5
32	0	0	3	2	4	4	2
I_C (%)	56.87	67.5	83.75	89.37	93.12	91.87	90.62

4.2 LSP

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

6	3	5	4	5	5	5	5
7	3	5	5	4	5	4	4
8	5	5	5	5	5	5	5
9	5	2	4	5	5	5	5
10	2	5	5	5	5	5	5
11	4	4	4	4	4	4	4
12	4	5	5	5	5	5	5
13	4	5	5	5	5	5	5
14	4	4	4	5	5	5	5
15	4	5	5	5	5	5	5
16	3	2	3	4	4	4	4
17	4	4	5	5	5	5	5
18	5	5	5	5	5	5	5
19	5	5	5	5	5	5	5
20	4	3	4	4	4	4	4
21	2	4	5	5	5	5	5
22	4	5	5	5	5	5	5
23	5	5	5	5	5	5	5
24	5	5	4	5	5	5	5
25	5	5	5	5	5	5	5
26	5	5	5	5	5	5	5
27	4	4	4	5	5	5	5
28	4	4	4	5	5	5	5
29	3	4	5	5	5	4	4
30	5	5	5	5	5	5	5
31	3	5	5	5	5	5	5
32	3	4	3	4	4	4	4
I_C (%)	80	85.62	91.25	96.25	96.25	95.62	95.62

5. Etude de l’effet du nombre des locuteurs

On utilise notre base de donnée.

Fe= 16KHz -MFCC- QV

5.1 Nombre des locuteurs égaux à 16

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	5	1	5	5	5	5	5
2	4	5	5	5	5	5	5
3	5	0	4	5	5	5	5
4	3	5	5	5	5	5	5
5	4	3	5	4	5	4	4
6	2	1	4	5	5	5	5
7	4	2	4	4	4	5	5
8	2	0	5	5	5	5	5
9	4	5	5	5	5	5	5
10	4	2	4	5	5	5	3
11	4	4	4	4	4	4	4
12	2	4	4	5	5	5	5

13	5	5	5	5	5	5	5
14	5	5	5	5	5	5	5
15	5	5	5	5	5	5	5
16	1	3	5	4	5	5	5
I_C (%)	73.75	62.5	92.5	95	97.5	97.5	95

5.2 Nombre des locuteurs égaux à 24

	Ordre du modèle						
locuteur	1	2	4	8	16	32	64
1	5	1	5	5	5	5	5
2	3	5	5	5	5	5	5
3	4	0	4	5	5	5	5
4	3	5	5	5	5	5	5
5	3	2	4	4	5	5	4
6	2	1	4	5	5	5	5
7	4	2	4	4	4	5	5
8	2	0	5	5	5	5	5
9	4	5	5	5	5	5	5
10	4	2	4	5	5	5	3
11	4	4	4	4	4	4	4
12	2	3	4	5	5	5	5
13	4	4	5	5	5	5	5
14	5	5	5	5	5	5	5
15	5	5	5	5	5	5	5
16	2	3	5	4	5	5	5
17	4	4	4	5	5	4	5
18	5	5	5	5	5	5	5
19	4	4	4	4	4	4	4
20	4	4	4	4	4	4	4
21	4	5	5	5	5	5	5
22	5	5	5	5	5	5	5
23	3	5	3	5	5	5	5
24	5	5	5	5	5	5	5
I_C (%)	75	70	90	95	96.67	96.67	95

BIBLIOGRAPHIE

[1] Abdelatif Mokeddem, Thèse de doctorat, « analyse de la parole : reconnaissance multilocuteurs des mots isolés pour les système miniaturisés », université de Neuchâtel, institut de microtechnique, imprimerie centrale Neuchâtel ,1985.

[2] Alexis Moinet & Maxime Tryhoen, « IMPLEMENTATION D'UN CODEUR LPC10 complet sous matlab », Faculté Polytechnique de Mons, Belgique, 9 Rue de Houdain, 7000 Mons.

[3] Anil Alexander & Plamen Prodanov, « Analyse homomorphique ». Presses polytechniques romandes, Lausanne, mai 2003.

[4] A.V.Oppenheim, R.W.Shaffer: « Digital signal processing ».prentice Hall, New Jersey.

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

[6] Boris Mailhé, ENS Cachan, « Adaptation discriminante de modèles pour la séparation de sources : application au débruitage de signaux de parole », INRIA, 2005.

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

[8] Christophe L ´EVY, Thèse de doctorat, « Modèles acoustiques compacts pour les systèmes embarqués », l'Université d'Avignon et des Pays de Vaucluse, novembre 2006.

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

[10] 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.

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

[12] Lucie BAILLY, « Etude articulatoire de la parole produite en environnement bruyant », Laboratoire d'Acoustique Musicale, Université Paris 6, juillet 2005.

[13] 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).

[14] 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).

[15] M. Djamah, M. Boudraa, B. Boudraa, M. Bouzid, «Quantification adaptative des coefficients LSF pour le codage de la parole à bas débit », Laboratoire Communication Parlée et Traitement de signal (LCPTS) Faculté d'Electrique et d'informatique, USTHB, BP32, EL-ALIA, ALGERIE.

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

[17] M.Najim, « modélisation et identification en traitement de signal ». Edition Masson, Paris, 1988.

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

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

[20] Taoufik En-najjary, Thèse de doctorat, « Conversion de voix pour la synthèse de la parole », université de Rennes 1, Avril 2005.

[21] 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.