8/04

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

Ecole Nationale Polytechnique

Département d'Electronique

Projet de fin d'étude

Thème :

Identification du locuteur indépendante du texte

Etudié par : *M'.H.HADJ-ALI*
M'. M. BOUCHAMEKH

Soutenu devant le jury composé de :

Melle M.GUERTI : PRESIDENTE

M' B.BOUSSEKSOU : RAPPORTEUR

Mlle A.MOUSSAOUI : EXAMINATRICE

Promotion Juin 2004

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

Identification du Locuteur

ملخص

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

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

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

Résumé

L'identification du locuteur est une branche de la reconnaissance vocale, elle est en pleine expansion depuis dix ans et cela à cause de l'élargissement de l'utilisation de l'électronique dans le domaine domestique et également professionnel.
Cet élargissement a nécessité la protection des droits d'accès et l'utilisation des propriétés de la voix humaine dans la caractérisation de la personne et notamment sa modélisation.

Après une brève définition de l'identification du locuteur et la présentation des différentes étapes obligatoires et algorithmes pour l'étude de la parole : modélisation, analyse et classification, nous entamerons les différents algorithmes algébriques (utilisation du principe iso data) et statistiques (HMM et GMM).

Les mots clés : identification, coefficients de prédiction linéaire, quantification vectorielle, la multi gaussienne.

Abstract

The identification of the speaker is a branch of the voice recognition; it is into full expansion since ten years and that because of widening with the use of electronics in the domestic and professional field. This widening required the protection of the rights of accés and the use of the properties of the human voice in the characterization of the person and in particular its modeling.

After a short definition of the identification of the speaker and the presentation of the various obligatory stages and algorithms for the study of the word: modeling, analyzes and classification, we will start the various algebraic algorithms (Isodata) and statistics (HMM and GMM).

Keywords : identification, lineaire prediction coefficients, vector quantization, the gaussian mixture model.

Dédicaces

Nous dédions ce travail :

A nos parents qui nous ont soutenus, orientés et encouragés tout au long de nos études.

A nos frères et sœurs.

A nos familles.

A tous nos amis (es).

Mouslem et Hacène

Remerciements

Ce projet de fin d'étude est réalisée au Département Eléctronique de l'Ecole Nationale Polytechnique.

A Monsieur le promoteur B.Bousseksou qui nous a proposé le sujet et encouragé à la réalisation de ce travail et nous a aidé à orienter correctement nos recherches en nous faisant profiter de son expérience dans ce domaine. Sans oublier qu'il était derrière toute démarche ou contact dans le cadre de notre travail. Nous tenons à lui exprimer toute notre gratitude.

Nous remercions également notre co-promoteur Monsieur M. Ben gharabi, chercheur au CDTA, qui nous a ouvert les portes et acceuilli chaleureursement et aidé efficacement dans notre travail. Qu'il soit remercié.

Nous remercions tout particulièrement Monsieur Halimi qui n'a pas hésité une seconde et nous a totalement fait confiance et nous a permis de faire un stage au CDTA.

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 soutenus et aidés dans la réalisation de ce travail.

Enfin, nous tenons à exprimer notre reconnaissance à nos mères pour l'assistance qu'elles ont su nous apportés, jour après jour.

SOMMAIRE

INTRODUCTION A LA RECONNAISSANCE DU LOCUTEUR

Introduction

L'expression vocale est une caractéristique propre du locuteur et dépend de plusieurs paramètres pertinents qu'on va évoquer au deuxième chapitre. La reconnaissance du locuteur dépend étroitement de ces paramètres issus de la phase de pré-traitement qui affectent le taux de discrimination entre locuteurs.

Dans la reconnaissance du locuteur il y a l'identification et la vérification.
La vérification du locuteur consiste à accepter ou à refuser une identité proclamée. Après s'être identifier, la distance entre l'expression vocale et sa référence personnelle sera comparée à un seuil d'acceptation préalablement déterminé.
L'identification du locuteur qui peut être dépendante ou indépendante du texte consiste à reconnaître un locuteur parmi un ensemble fini de locuteurs en comparant sa trace vocale avec des références connues.

Notre étude a pour objet l'identification du locuteur indépendante du texte en utilisant une méthode algébrique la quantification vectorielle (K-moyennes) et des méthodes statistiques comme la GMM (Modèles à mixture gaussienne).0

1. Paramètres acoustiques.

L'extraction des paramètres acoustiques est effectuée d'une manière indépendante du texte. Les caractéristiques se trouvent dans l'enveloppe spectrale (caractéristique du conduit vocale) et dans les paramètres de la source (fréquence fondamentale).
On utilisera les paramètres MFCC (Mel Frequency Cepstrum Coefficient) .

2. Identification du locuteur indépendante du texte.

Un algorithme efficace permet d'identifier le locuteur indépendamment du texte : la GMM (Gaussian Mixture Model) .
Elle consiste à modéliser le vecteur, par exemple les coefficients MFCC par une probabilité statistique qui est un mélange de plusieurs gaussiennes .

3. Application.

Son application reste jusqu'à présent dans le militaire, elle n'a pas d'application concrète dans le civile mais elle a de grand espoir dans le développement du langage et dans le diagnostic de l'expert à l'opposé de la vérification appliquée dans la téléphonie mobile et dans les domaines de protection des droits d'accès par vérification vocale.

4. Conclusion.

L'identification du locuteur indépendante du texte, dépende en fait des paramètres utilisés et leurs degrés de pertinence et de robustesse vis-à-vis du bruit.
Nous allons définir les algorithmes nécessaires pour avoir une bonne discrimination.

Chapitre 1: MODELISATION DU SIGNAL PAROLE

Introduction

Le traitement de la parole est aujourd'hui une composante fondamentale des sciences de l'ingénieur. Située au croisement du traitement du signal numérique et du traitement du langage (c'est-à-dire du traitement de données symboliques), cette discipline scientifique a connu depuis les années 60 une expansion fulgurante, liée au développement des moyens et des techniques de télécommunications. L'importance particulière du traitement de la parole dans ce cadre plus général s'explique par la position privilégiée de la parole comme vecteur d'information dans notre société humaine.

L'extraordinaire singularité de cette science, qui la différencie fondamentalement des autres composantes du traitement de l'information, tient sans aucun doute au rôle fascinant que joue le cerveau humain à la fois dans la production et dans la compréhension de la parole et à l'étendue des fonctions qu'il met, inconsciemment, en œuvre pour y parvenir de façon pratiquement instantanée. Pour mieux comprendre cette particularité, penchons-nous un instant sur d'autres vecteurs d'information. L'image, par exemple, n'existe que dans la mesure où elle est appelée à être perçue par l'œil, et, bien au-delà, interprétée par le cerveau. Les techniques de traitement de l'image pourront en tirer parti en prenant en compte, d'une part, les caractéristiques physiques de l'œil et, d'autre part, les propriétés perceptuelles que lui confère le cortex visuel. Un exemple bien connu de ce type d'influence du récepteur sur le mode de traitement des signaux associés nous est fourni par l'image vidéo, dont les 24 images/seconde découlent directement du phénomène de persistance rétinienne. A l'inverse, un signal d'origine biologique tel que l'électromyogramme, qui mesure l'état d'activité d'un muscle, n'existe que dans la mesure où il est produit par ce muscle, sous le contrôle étroit du cortex moteur. Une bonne connaissance du muscle sera par conséquent un pré requis indispensable au traitement automatique de l'électromyogramme correspondant.

1. 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, sémantique,* et *pragmatique.*
On étudiera principalement le niveau phonétique.

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, et perçu par le système auditif.

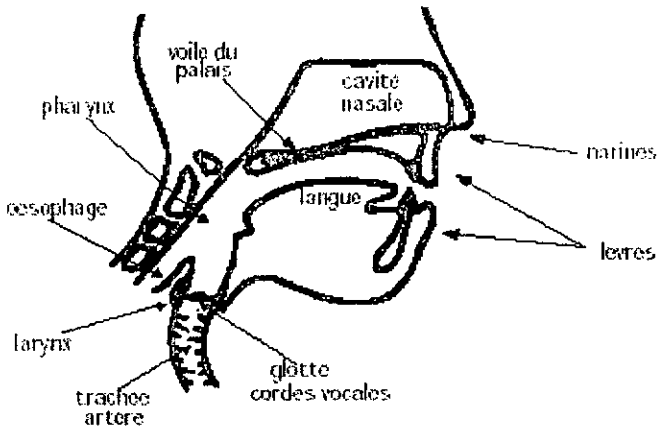


Figure 1.1 : Mécanisme de phonation.

3. Mécanisme de la 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. 1.2).

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 (ou ***sourds***). Les sons voisés (ou ***sonores***) 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.

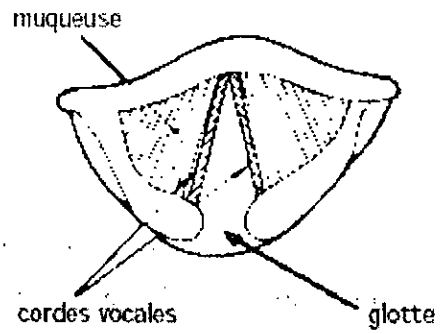


Figure 1.2 Section du larynx vu de haut

4. Concepts fondamentaux physico-acoustique et de phonétique

Nous donnons la définition des principaux termes utilisés dans le décodage acoustico-phonétique.

1. Les sons voisés

La vibration des cordes vocales produit les sons voisés (voyelles, semi voyelles, nasales).

Cette vibration est en fait leur accolement, puis leur séparation sous l'effet de la pression de l'air provenant des poumons, et de nouveau leur accolement sous l'effet des forces de Bernoulli produites par le passage de l'air.

2. Les sons non voisés

Les sons non voisés sont obtenus par divers bruits produits par le passage de l'air en un point de resserrement du canal vocal, ou par des bruits d'occlusion ou de plosive provoqués par la fermeture ou l'ouverture des lèvres, ou des chocs de la langue contre le palais.

3. Le timbre

Le timbre est caractérisé par la richesse en amplitudes relatives des harmoniques du pitch.
En effet, les cavités du résonateur de l'appareil phonatoire ont pour propriété de renforcer certaines harmoniques du son fondamentale laryngé, en atténuant les autres.

4. Spectre de la source

Les cordes vocales, vibrent périodiquement en laissant échapper l'air durant un temps qui est court devant la période du phénomène. Le spectre de la source est donc composé par une fréquence fondamentale, et un grand nombre d'harmoniques, ayant une enveloppe dont la forme est proche d'une exponentielle.

5. Spectre du canal

C'est la fonction de transfert du canal vocal en tant que tube acoustique. L'examen des spectres de signal vocal (éventuellement lissés pour éliminer l'influence de la source) montre la présence d'un certain nombre de pics dans le spectre des voyelles. Ce sont les zones correspondant aux fréquences renforcées par les différents résonateurs couplés. Ces zones sont appelées des zones formantiques, ou formants

6. La fréquence fondamentale

La vibration, qui est en fait l'accolement puis la séparation des cordes vocales portées par le larynx détermine la fréquence fondamentale appelée pitch ou F_0. Elle est comprise entre 75 et 150 Hz chez les hommes, 150 et 300 Hz chez les femmes, et est supérieure ou égale à 300 Hz chez les enfants.

7. La mélodie

La mélodie de la voix est caractérisée par les fluctuations de F_0 en fonction du temps.

8. La prosodie

La prosodie introduit dans la prononciation d'une phrase des nuances qui, dans la langue écrite, demanderaient des ponctuations ou des énoncés différents. Ce sont les caractéristiques prosodiques qui permettent à un auditeur de suivre une conversation même en milieu défavorable. Les principaux paramètres prosodiques sont l'intonation, l'intensité et la durée.

8.1 L'intonation

est un paramètre très important. Elle correspond à une hauteur donnée de F_0 et l'oreille est très sensible à ses variations.

8.2 L'intensité

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

8.3 La durée

Elle fixe le rythme de la phrase.

9. L'articulation.

Les traits distinctifs du locuteur sont liés aux phénomènes d'articulation, et concernent non plus l'activité de la source ou des cavités, mais l'activité musculaire du locuteur.

9.1 La coarticulation.

Le locuteur prononçant une phrase produit une suite de phonèmes qui enchaînés les uns aux autres de façon continue, en reliant les parties stables du signal (canal vocal en équilibre, donc signal quasi-périodique) par des zones de transition. La dynamique du conduit vocale représentée par les variations de la fonction de transfert, est donc un ensemble de traits distinctifs du locuteur, lié à la musculature.

9.2 Occlusives.

C'est la durée du silence précédent l'explosion dans les plosives / p / / t / / k /, il s'agit d'un paramètre temporel difficile à imiter car régulé par des mécanismes plutôt réflexes.

9.3 Enveloppe énergétique.

L'énergie du signal en tant qu'énergie de tout ou partie du spectre à court terme du signal est également liée dans son évolution le long d'une phrase à l'identité du locuteur. C'est une donnée assez facile à imiter ce qui explique qu'elle ne soit utilisée que conjointement à d'autres paramètres moins sensibles à l'imitation.

5. Phonèmes.

Définition :

Un phonème est la plus petite unité présente dans la parole et susceptible par sa présence de changer la signification d'un mot.
La production d'un phonème donné laisse toutefois place à une certaine variabilité sur la plan acoustique.

Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur ***mode articulatoire***. On distingue généralement trois classes principales : les ***voyelles***, les ***semi-voyelles*** et les ***consonnes***.

Les voyelles diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal (et non, comme on l'entend souvent dire, par le degré d'activité des cordes vocales, déjà mentionné sous le terme de ***voisement***).

Si le conduit vocal est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la bouche se réduit alors à une modification du timbre vocalique. Si, au contraire, le passage se rétrécit par endroit, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : une consonne est produite. La bouche est dans ce cas un organe de production à part entière.

Les voyelles se différencient principalement les unes des autres par leur *lieu d'articulation*, leur *aperture*, et leur *nasalisation*. On distingue ainsi, selon la localisation de la masse de la langue, les voyelles ***antérieures***, les voyelles ***moyennes***, et les voyelles ***postérieures***, et, selon l'écartement entre l'organe et le lieu d'articulation, les voyelles ***fermées*** et ***ouvertes***. Les voyelles ***nasales*** diffèrent des voyelles *orales* en ceci que le voile du palais est abaissé pour leur prononciation, ce qui met en parallèle les cavités nasales et buccales. Notons que, dans un contexte plus général que celui de la seule langue française, d'autres critères peuvent être nécessaires pour différencier les voyelles, comme leur *labialisation*, leur *durée*, leur *tension*, leur *stabilité*, leur *glottalisation*, voire même la *direction du mouvement de l'air*.

Les semi-voyelles, quant à elles, combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes. Enfin, les liquides sont assez difficiles à classer. L'articulation de [ä] ressemble à celle d'une voyelle, mais la position de la langue conduit à une fermeture partielle du conduit vocal. Le son [ô], quant à lui, admet plusieurs réalisations fort différentes.

On classe principalement les consonnes en fonction de leur *mode d'articulation*, de leur *lieu d'articulation*, et de leur *nasalisation*. Comme pour les voyelles, d'autres critères de différenciation peuvent être nécessaires dans un contexte plus général : l'*organe articulatoire*, la *source sonore*, l'*intensité*, l'*aspiration*, la *palatalisation*, et la *direction du mouvement de l'air*.

La distinction de mode d'articulation conduit à deux classes : les ***fricatives*** (ou ***constrictives***) et les ***occlusives*** (ou ***plosives***). Les fricatives sont créées par une constriction du conduit vocal au niveau du lieu d'articulation, lui peut être le palais, les dents ou les lèvres.

Les fricatives non-voisées sont caractérisées par un écoulement d'air turbulent à travers la glotte, tandis que les fricatives voisées combinent des composantes d'excitation périodique et turbulente : les cordes vocales s'ouvrent et se ferment périodiquement, mais la fermeture n'est jamais complète. Les occlusives correspondent quant à elles à des sons essentiellement dynamiques. Une forte pression est créée en amont d'une occlusion maintenue en un certain point du conduit vocal (qui peut ici aussi être le palais, les dents, ou les lèvres, puis relâché brusquement. La période d'occlusion est appelée la phase de tenue.

Pour les occlusives voisées un son basse fréquence est émis par vibration des cordes vocales pendant la phase de tenue; pour les occlusives non voisées, la tenue est un silence.

Enfin, les consonnes nasales font intervenir les cavités nasales par abaissement du voile du palais.

Les traits acoustiques du signal de parole sont évidemment liés à sa production. L'intensité du son est liée à la pression de l'air en amont du larynx. Sa fréquence, qui n'est rien d'autre que la fréquence du cycle d'ouverture / fermeture des cordes vocales, est déterminée par la tension de muscles qui les contrôlent. Son spectre résulte du filtrage dynamique du signal glottique (impulsions, bruit, ou combinaison des deux) par le conduit vocal, qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses. Ainsi, par exemple, on peut approximativement représenter les voyelles dans le plan des deux premiers formants (Fig. 1.3). On observe en pratique un certain recouvrement dans les zones formantiques correspondant à chaque voyelle (un affichage en trois dimensions figurant les trois premiers formants permettrait une meilleure séparation).

Classe phonétique	**abbrev**	**phonèmes**
occlusions	CLO	cl, vcl
relâchement d'occlusion (explosion)	OCC	b, d, g, k, p, t
Fricatives	FRI	S, Z, f, s, v, z
Nasales	NAS	m, n
consonnes vocaliques	DVG	R, j, l, w
voyelles	VOY	o~, A, E, U~, O, a~

Tableau : Le jeu d'unité phonétique

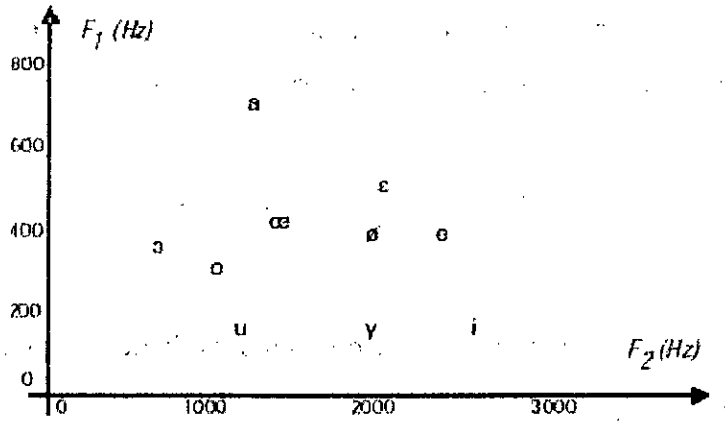


Figure 1.3 Triangle vocalique

6. Modélisation de la production de la parole

Définition.

La modélisation d'un signal *x(n)* consiste à lui associer un filtre linéaire qui, soumis à une excitation particulière reproduit ce signal le plus fidèlement possible.
L'objectif essentiel de la modélisation d'un signal est de permettre la description de son spectre par un ensemble très limité de paramètres

Fant a proposé en 1960 un modèle électrique de production de la parole [2]. 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*. Cette modélisation reste valable dans le cas de sons non-voisés, à condition que *u(n)* soit un bruit blanc. Le modèle final est illustré à la figure 1.4.
Il est souvent appelé *modèle auto-régressif*, parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme :

$$x(n)=\sigma u(n)+\sum_{i=1}^{p} a_i x(n-i) \quad i=1,\ldots,p$$

(où *u(n)* est le signal d'excitation), 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*.

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

$$X(z)=U(z)\frac{\sigma}{A(z)} \quad avec\ H(z)=\frac{\sigma}{A(z)}$$

$U(z)$: excitation (bruit blanc ou train périodique d'impulsion)

σ : gain du modèle

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

Si:

Dans le domaine temporel cela aboutit, cela revient à la récurrence suivante :

$$x(n)+\sum_{i=1}^{p} a(i)x(n-i)=\sigma u(n)$$

C'est à dire qu'un échantillon est une combinaison linéaire des échantillons précédents et du terme d'excitation. C'est cette récurrence qui défini le modèle auto régressif d'ordre p.

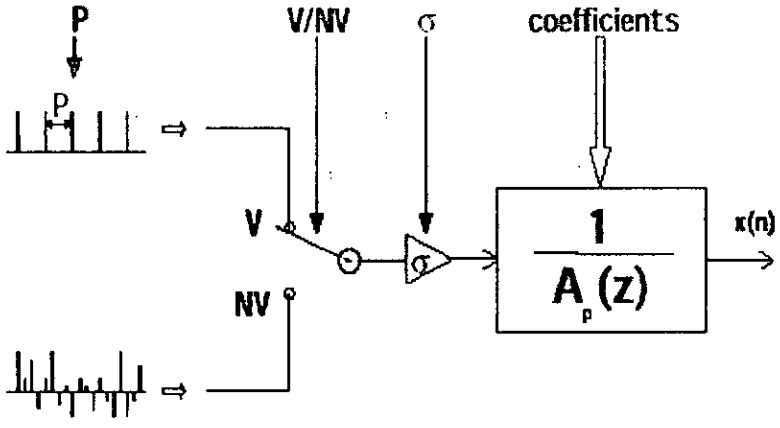


Figure 1.4 Analyse par LPC (*Linear Prediction Coefficients*).

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

Conclusion

Le mécanisme de la production de la parole qui a été décrit dans ce chapitre conduit très naturellement à une modélisation particulière appelée *modélisation autorégressive*.

Cette modélisation est un outil très commode pour procéder à l'analyse du signal vocal. Il faut préserver quelques précautions car l'hypothèse de stationnarité n'est pas valable que durant de courts intervalles de temps.
Cette modélisation fait la distinction entre deux sons de la parole, ce qui est un début pour distinguer au préalable les sons voisés et les non voisés.

Chapitre 2 : ANALYSE DE LA PAROLE

Introduction.

On peut classer les méthodes d'analyse en 3 groupes, les méthodes spectrales ou périodogramme, les méthodes temporelles comme la LPC et enfin les méthodes basées sur le modèle de la perception. Il convient d'adjoindre les méthodes de mesure de la fréquence fondamentale F0 et des formants associés au locuteur.

Au cours du chapitre précédent on a montré qu'un système linéaire tout pôle peut modéliser la production des sons par le conduit vocal.
Maintenant, il nous semble utile de déterminer les coefficients a_i du modèle :

$$A(z)=\sum_{i=0}^{p} a_i z^{-i}$$

p : ordre de la prédiction

Tout d'abord, on admet les hypothèses qui justifient ce modèle.
La stationnarité du signal vocal garantie la validité du modèle sur des intervalles de l'ordre de 30 ms. Il faut rafraîchir les valeurs des coefficients a_i plusieurs fois par seconde.

Toutefois, on voit bien qu'on a réalisé une compression de données très efficace car, sur une durée de 30 ms pour une fréquence d'échantillonnage de 10 kHz, on a 300 échantillons, et le codage par prédiction linéaire les ramène à prés de 10 (si l'ordre de la LPC est 10)

1. La paramétrisation du signal

Un système de paramétrisation du signal, appelé aussi pré-traitement acoustique, se décompose en trois étapes, un filtrage analogique, une conversion analogique/numérique et un calcul de coefficients (schéma. 1). Son rôle est de fournir et d'extraire des informations caractéristiques et pertinentes du signal pour produire une représentation moins redondante de la parole.

L'information acoustique pertinente du signal de parole se situe principalement dans la bande passante [50 Hz - 8 kHz], la fréquence d'échantillonnage devrait donc au moins être égale à 16 kHz, selon le théorème de Shannon ; mais elle peut varier en fonction du domaine d'application ou des besoins ou contraintes matériels.

Dans notre application, on va étudier la robustesse de nos algorithmes dans la bande téléphonique. N'oublions pas que la dégradation de notre signal est de 50 %.

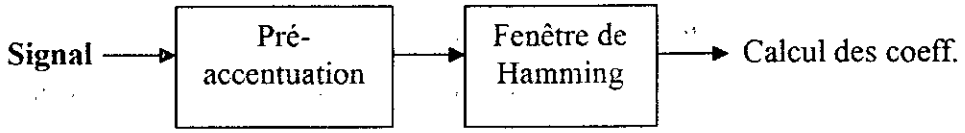


Schéma. 1 étapes de pré-traitement.

1.1 La pré-accentuation

L'onde sortante des lèvres subit une atténuation, puisqu'il y a désadaptation entre les deux milieux, et une distorsion qui peut-être assimilée à une désaccentuation de 6 dB par octave sur tout le spectre, il s'agit donc de faire une pré-accentuation des fréquences les plus hautes (filtrage analogique avant l'échantillonnage, ou filtrage numérique juste après dans notre cas) avant les traitements ainsi on peut estimer l'évolution de la forme du conduit vocal et les lieux d'articulation et en déduire ainsi la parole prononcée.

Le filtre de pré-accentuation est : $1 - 0.95\, z^{-1}$

1.2 Le filtre de Hamming

C'est la fenêtre la plus utilisée en traitement de la parole, sa forme géométrique permet de mieux segmenter notre signal sans une perte importante d'informations et sans augmenter les effets de bord.

Son expression est :

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

N : Nombre des échantillons dans chaque tranche

2. Sonagramme.

Le sonagramme ou spectrogramme sonore est une représentation tridimensionnelle du signal. Les fréquences de grande amplitude sont dessinées avec un point plus foncé et les fréquences moins importantes avec des points plus clairs.
Selon la résolution de chaque spectre, l'analyse peut être à bande étroite ou à large bande .
À bande étroite, on utilise une fréquence de 250 Hz (4 ms de fenêtre) et le résultat est l'apparition de stries dans le sens horizontal. À large bande, on utilise une fréquence de 50 Hz (20 ms de fenêtre) et le résultat est l'apparition de stries dans le sens vertical.

3. Analyse par DFT.

Elle est fondée sur une décomposition fréquentielle du signal sans connaissance a priori de sa structure fine. La seule hypothèse mise en jeu concerne le choix de fonctions sur la base desquelles le signal est décomposé : sinusoides pour la transformée de Fourier.
Elle est peut etre approchée par un vocodeur à canaux dans lequel les filtres seraient à bande extremement étroite et répartis linéairement sur l'axe fréquentiel..
La transformée de Fourier à court-terme d'un signal échantillonné est par définition :

$$X(n,\theta)=\sum_{m} x(m)w(n-m)\exp(-jm\theta),\theta=wT$$

C'est une méthode qui ne peut pas opérer sur des séquences trop courtes de signal. Pour suivre au mieux les transitions de la parole, il est nécessaire de prendre des fenètres temporelles avec recouvrement.

Remarque.
On ramène le spectre à une échelle non linéaire (Bark ou Mel) pour tenir compte de la perception humaine.
Echelle Bark : B = $6\,ArcSinh(F)$ B :Bark

Echelle Mel : M = $10^{+3}Log(1+F)$ F en Hz

4. Analyse cepstrale.

Elle résulte du modèle de production, son but est d'effectuer une déconvolution source / conduit par une transformation homomorphique :les coefficients cepstraux sont obtenus en appliquant la transformée de Fourier numérique inverse au logarithme du spectre d'amplitude. Le signal ainsi obtenu est représenté dans le domaine cepstrale ou quéfrentiel.
Les échantillons se situant en basses quéfrences correspondent à la contribution de la source et donc définir la fréquence fondamentale tandis que le conduit vocale n'apparaît qu'en hautes quéfrences.

1. Application de la FFT.

Après avoir pré-traité notre signal, on applique la transformée de Fourier rapide (FFT).
Cette application nous permet d'analyser le signal dans le domaine fréquentiel.

2. Banc de filtres.

Pour tenir compte du modèle de perception, on applique un banc de filtres sur les échantillons obtenus de la FFT. Ces filtres sont de type triangulaires qui mettent en évidence les bandes de perception de l'oreille humaine. Ils atténuent les effets de bords et favorisent les fréquences sensibles de l'oreille.

L'étalement des filtres est linéaire jusqu'à 1000 Hz, au-delà, il est logarithmique.
Le chevauchement entre filtres détermine le degré de redondance de l'information . Ce degrés est fonction du nombre de filtres utilisés dans l'analyse dans la bande de perception.

Remarque :

L'étalement des différents filtres est optimisé pour obtenir la meilleure repréresentation possible afin de mieux suivre la perception de l'oreille.

3. Transformation cosinus Le cepstre réel.

La transformation cosinus effectuée est donnée par :

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \qquad n = 1,2,...,K$$

Où $\tilde{S}_k$ sont les énergies calculées après le passage par les filtres.

5. Avantages de l'utilisation du cepstre.

1. Le cepstre rend visible les faibles densités spectrales situées dans les bandes de fréquence sensibles à la perception ; l'interprétation du contenu fréquentiel d'un spectre est facilitée.
2. La large gamme dynamique spectrale est réduite par la conversion logarithmique quantitativement. Cela revient à aplatir les données dans le domaine spectral. La conséquence de cette opération se répercute dans le domaine temporel par une concentration de l'information autour de l'origine, c'est-à-dire la plupart de l'énergie du cepstre se trouve dans la première dizaine de coefficients.
3. La conversion logarithmique soulève les formants de très faible énergie localisés fréquemment dans la bande fréquentielle perceptible.

6. Caractéristique de la parole et du locuteur.

6.1 Caractéristique de la parole

L'une des principales difficultés posées par le traitement de la parole est la variabilité intrinsèque, cette étape est liée à l'importante variabilité des productions possibles pour un même message linguistique. Cette variabilité est à la fois fréquentielle et temporelle.

Du fait des différences dans la taille des cavités articulatoires des locuteurs (liées à leur âge, à leur sexe, etc.), leurs fréquences de résonance diffèrent d'un individu à l'autre. Les filtres qui vont amplifier ou atténuer certaines fréquences du signal source sont donc différents.

Chaque locuteur génère ainsi des signaux de parole avec une hauteur de voix (fréquence fondamentale ou F0) différente ; ce phénomène induit également une répartition variable des formants sur l'échelle des fréquences. Par ailleurs, un locuteur peut parler plus ou moins rapidement en fonction des situations et changer de vitesse d'élocution à l'intérieur d'un même énoncé. Cette variabilité dans l'organisation spectrale et temporelle des sons de parole n'entrave cependant pas la stabilité perceptive qui permet à un auditeur d'entendre des phonèmes formes stables.

6.2 Caractéristique du locuteur.

Le problème essentiel pour l'extraction de l'information est lié à la grande variabilité avec laquelle sont représentés chacun des divers sons. Variabilité qui a sa cause dans la présence de trois facteurs. En effet le signal parole porte non seulement une information sur le contenu sémantique du message, mais aussi une information sur l'identité du locuteur, et une information sur l'état physique, émotif et psychique dans lequel se trouve le locuteur.

Conclusion.

La méthode d'analyse de la parole dépend de plusieurs paramètres, selon l'application envisagée, dans notre cas l'identification. On a choisit pour notre étude l'analyse cepstrale issue du modèle de perception de l'oreille étant donnée qu'elle fournit des paramètres discriminants étendus sur tout le spectre de la bande de perception de l'oreille humaine et caractérise efficacement les coordonnées du locuteur dans l'espace spectral.

Chapitre 3 : EFFET D'ENTRAINEMENT ET MODELES DE CLASSIFICATION

Introduction

La reconnaissance automatique de la parole consiste à extraire l'information lexicale contenue dans un signal de parole, et éventuellement l'interpréter.

Il y a plusieurs méthodes pour l'éxtraire, à savoir la méthode relevant de l'intelligence artificielle et la deuxième orientée vers un traitement statistique ou algébriques des données et la troisième méthode les méthodes hybrides. On va aborder ces deux dernières approches et essayer au préalable de définir l'effet d'entrainement et classification des données .

1. Classification des données

La distinction entre données d'apprentissage et de test est des plus importantes, non seulement à cause de leurs fonctionnalités différentes mais surtout à cause du fait que le choix des données d'apprentissage ainsi que de l'algorithme d'entrainement utilisé devrait garantir les performances optimales sur l'ensemble de test.

Dans notre cas, on est confronté à un très grand nombre de formes possibles, et le système de classification devrait idéalement etre capable de les reconnaître toutes. Cependant pour des raisons pratiques évidentes, les systèmes sont entrainés sur un ensemble limité d'exemples.

Avant d'aborder la classification, il faut définir la métrique et la métrique prévilégiée dans l'entrainement.

2. La métrique dans l'espace acoustique.

Def 1 : Distance entre 2 vecteurs X et Y doit satisfaire aux conditions :

1- $d(X,Y) \geq 0$;

2- $d(X,Y) = d(Y,X)$;

3- $d(X,Y) \leq d(X,U) + d(U,Y)$;

Def 2 :

Une définition particulière de la distance entre 2 spectres doit être :

- significative sur le plan acoustique.
- formalisable d'une façon efficiente sur le plan mathématique.
- définie dans un espace de paramètres judicieusement choisi.

Norme de Holder.

$$d_p(X,Y)=\left[\sum_{k=1}^{K}|x_k-y_k|^p\right]^{\frac{1}{p}}=\|X-Y\|$$

2.1 Distance cepstrale.

Soit $f(\theta)$ une fonction représentant une densité spectrale d'énergie $P_x(\theta)$. Les coefficients du cepstre réel sont donnés par :

$$\ln f(\theta)=\sum_n c(n)\exp(-jn\theta).$$

On définit la distance spectrale logarithmique :

$$d_p=\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}|V(\theta)|^p\right]^{\frac{1}{p}};V(\theta)=\ln f_1(\theta)-\ln f_2(\theta)$$: différence entre 2 spectres

Pour p=2 :

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

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

$$=\sum_{l=1}^{\infty}\left[c(l)-c'(l)\right]^2=\left[c(0)-c'(0)\right]^2+2\sum_{l=1}^{\infty}\left[c(l)-c'(l)\right]^2$$

2.2 Distance d'Itakura Saito.

La distance d'Itakura-Saito entre 2 densités spectrales $f et f'$ est définie par

$$d_{IS}\left(f,f'\right)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\left(\frac{f}{f'}\right)-\ln\left(\frac{f}{f'}\right)-1\right]d\theta$$

3. L'entraînement.

Le problème consiste à classer le mot ou les mots sous formé de valeurs caractéristiques en termes d'unités phonémiques qui pourront servir de base à la reconnaissance. A cet effet les paramètres du classificateur sont entraînés sur base d'un ensemble d'exemples de ces vecteurs caractéristiques dont on connaît la classification.

L'entraînement consiste à développer un ensemble de fonctions discriminantes tel que chaque fonction soit associée à une classe dont les paramètres seront optimisés dans le but de générer, pour chaque forme d'entrée de la base de données, la valeur la plus élevée pour la fonction correspondant à la classe correcte.
Dans l'identification indépendante du texte, l'entraînement est non supervisé.

3.1 L'entraînement non supervisé.

Le vecteur d'entrée x d'observation est connu et le modèle génératif associé à ce problème est défini en fonction de $p(x/\Theta)$ tel que Θ est le vecteur de paramètres.

Nous avons plusieurs vecteurs d'entraînement $X=\{x_n\}_{n=1}^{N}$.

Nous supposons que les x_n sont indépendants et identiquement distribués.

Nous appliquons le logarithme de la fonction vraisemblance à maximiser en fonction de Θ donc :

$$L_{\Theta}(X)=\sum_{n=1}^{N}\log p(x_n/\Theta)$$

L'algorithme utilisé pour maximiser le log-vraisemblance est l'algorithme Estimation Maximisation (EM).

3.2 Fonction discriminante.

Généralement c'est des fonctions discriminantes linéaires :

Chaque classe $w_k, \forall k=1,2,...,K$ est associée à une fonction discriminante linéaire de la forme :

$$g_k(x)=\overline{w}_k^T \overline{x} \quad , \text{T : transposée}$$

$\overline{x}=(1,x_1,...,x_d)^T$: Vecteur d'observation auquel on a ajouté une coordonnée additionnelle toujours égale à 1.

$\overline{w}_k=(w_{k0},w_{k1},...,w_{kd})^T$: les paramètres caractéristiques de la classe w_k.

w_{k0} est le biais da la classe w_k.

Remarque.

La coordonnée additionnelle est ajoutée pour ne pas contraindre l'hyperplan à contenir l'origine.

Dans notre étude, l'entraînement utilisé est non supervisé et on le discutera dans les différents algorithmes.

4. Modèles de classification.

4.1 La quantification vectorielle (QV).

La quantification vectorielle (notée **VQ**) est une généralisation de la quantification scalaire. Sa démarche consiste à quantifier des groupes (vecteurs) de K échantillons à la fois, sans tenir compte des échantillons voisins (extérieurs au vecteur courant). Le quantificateur vectoriel est défini par un ensemble fini D de niveaux de quantification, qui sont cette fois-ci des vecteurs, et par une fonction qui associe un niveau de quantification à chaque vecteur de K échantillons produit par la source. D est appelé le dictionnaire. Dans ce qui suit on donne les algorithmes d'apprentissage et comment ils sont exploité pour l'entraînement de la GMM (Gaussian Mixture Model).

4.2 Algorithme K-Moyennes ou Lloyd généralisé :

Cet algorithme est désigné originalement pour la quantification vectorielle mais il peut être exploité également pour l'entraînement des models GMM, il appartient à l'ensemble des méthodes non supervisées, chaque classe (ou partition) est représenté par sa moyenne (centroïde).

Soient l'ensemble des vecteurs d'entraînement $X = \{x_1, x_2, ..., x_T\}$, cet ensemble de vecteurs sera divisé à la fin de l'entraînement en M classes définies par leurs centroïdes $\{\mu_1, \mu_2, ..., \mu_M\}$. L'objectif est de minimiser la distorsion totale des classes.

On désignera par :

- $x_j^{(i)}$ tout vecteurs appartiennent à la classe i ;
- y_i le centroïde de la classe i ;
- $d(x_j^{(i)}, y_i)$ la distance (mesure de distorsion) entre $x_j^{(i)}$ et y_i ;
- D_i la distorsion totale de la classe i $D_i = \sum_i d(x_j^{(i)}, y_i)$;
- D la distorsion pour l'ensemble des classes $D = \sum_{i=1}^{M} D_i$;

L'algorithme se base sur les observations suivantes :

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

L'algorithme qui en résulte est le suivant :

a. Pour l'initialisation, M vecteurs aléatoires sont choisis, ils sont sélectionnés comme les M centroïdes de M classes.
b. Chaque vecteurs x_t, $1 \le t \le T$ est affecté à la classe dont le centroïde est le plus proche (distance minimale,
c. pour notre cas on utilise la distance euclidienne).

d. On recalcule la position de chaque centroïde y_i pour minimiser chaque distorsion D_i, le centroïde de la classe i donc est le vecteur moyen (fonction discriminante utilisée) de tous les vecteurs affectés à cette classe.
e. On calcule la distorsion totale D.

On itère les étapes (b) et (c) jusqu'à ce que D se fixe ou varie de moins de $\varepsilon\%$ d'une itération à la suivante.

Remarque :

Il est évident que selon le choix des classes initiales, le nombre d'itération varie jusqu'à la convergence de l'algorithme.

4.3 Algorithme LBG (*Lind-Buzo-Gray)* ou (Lloyd optimal) **:**

Dans l'algorithme k-Moyennes, l'affectation d'un vecteurs à une classe exige le calcul de M distances, ce qui rend élevé le temps de l'entraînement. Pour remédier à cette difficulté l'algorithme LBG est proposé. L'objectif et toujours la minimisation de la distorsion de l'ensemble de toutes les classes D.
Cet algorithme est appelé aussi à éclatement binaire parce que la naissance de deux classes est donnée à partir d'une classe pour chaque niveau. Le nombre de classes M est prédéfini, il ne peut être qu'une puissance de 2.

a. Tous les vecteurs d'entraînement sont supposés appartiennent à la même classe, on calcule donc le vecteur moyen de cette unique classe initiale μ_1^1.
b. Soit $\vec{\varepsilon}$ un vecteur avec une petite magnitude, donc le nombre de centroïdes est doublé par éclatement : $\mu_1^2 = \mu_1^1 + \varepsilon$ et $\mu_2^2 = \mu_1^1 - \varepsilon$.
c. Les vecteurs d'entraînement sont répartis entre ces deux nouvelles classes définies par leurs centroïdes, puis les moyennes sont réestimées.
d. Chaque centroïde obtenu est éclaté comme précédemment et les vecteurs moyens sont réestimés après chaque affectations.

La procédure est répétée jusqu'à ce que le nombre de classes soit égal à M.

Pour l'établissement des classes, diverses variantes existent, si après un éclatement et une réaffectation l'une des classes est presque vide, on peut estimer que seule l'autre classe est éclatée au stade suivant.

On peut d'ailleurs systématiser cette façon de faire et n'éclater qu'une seule classe à chaque stade :

- Soit celle qui possède le plus grand nombre d'éléments.
- Soit celle qui présente la distorsion D_i la plus grande.
- Soit encore celle qui présente la distorsion moyenne (D_i divisée par le nombre d'éléments de la classe) la plus grande.

Organigramme de LBG.

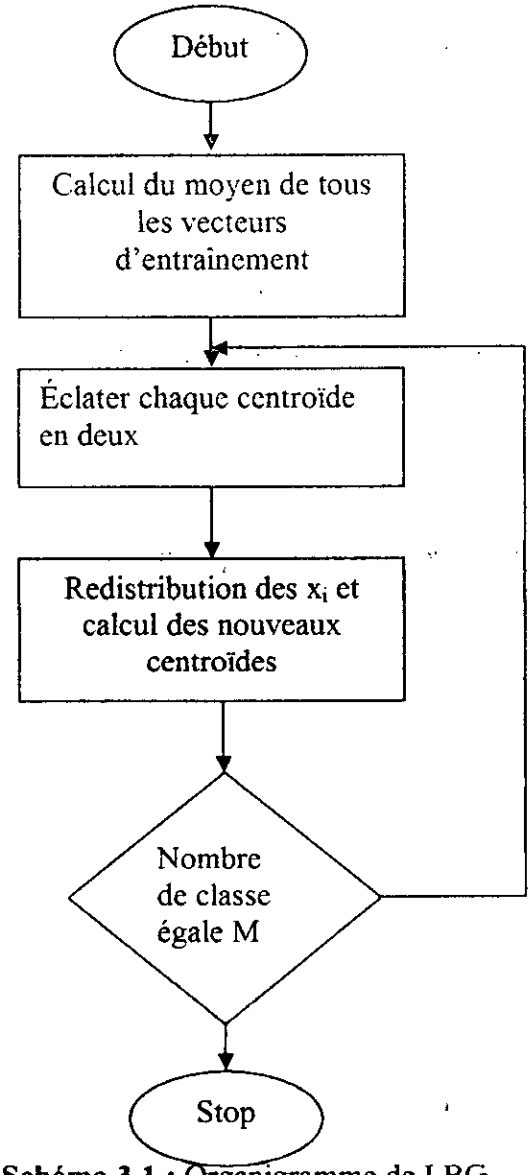


Schéma 3.1 : Organigramme de LBG

4.4 Algorithme EM (*Expectation Maximisation).*

Cet algorithme maximise de façon itérative dans Θ, la fonction de vraisemblance $p(X/\Theta)$ de l'ensemble des observations X conditionné sur Θ.

Définissons une fonction auxiliaire $Q(\Theta,\Theta^{(t)})$, un vecteur d'observations $X=\{x_n\}_{n=1}^{N}$.

Un vecteur caché $Y=\{y_n\}_{n=1}^{N}$ de la classe Ω associé à chaque X.

But.

Le but est de maximiser $p(X,Y/\Theta)$ et estimer Θ qui maximise la fonction :

$$L_\Theta(X)=\log p(X/\Theta)=\log\sum_Y p(X,Y/\Theta)$$

$$p(X,Y/\Theta)=\sum_{n=1}^{N} p(x_n,y_n/\Theta)$$

4.4.1 Étape d'estimation [2]

Nous évaluons la distribution a posteriori des variables cachées en utilisant les paramètres $\Theta^{(t)}$ à $p(Y/X,\Theta)=\dfrac{p(X,Y/\Theta^{(t)})}{\sum_Y p(X,Y/\Theta^{(t)})}$

La fonction auxiliaire est définie comme étant l'espérance mathématique du logarithme de la vraisemblance jointe sur l'ensemble complet des variables d'entraînement :

$$Q(\Theta,\Theta^{(t)})=\sum_Y p(Y/X,\Theta^{(t)})\log p(X,Y/\Theta)$$

4.4.2 Étape de maximisation

On recherche, par les méthodes habituelles d'optimisation de fonctions statistiques, l'ensemble des paramètres à utiliser à l'itération *(t+1)* tel que :

$$\Theta^{(t+1)}=\underset{\Theta}{\operatorname{argmax}}\,Q(\Theta,\Theta^{(t)})$$

4.4.3 Étude de la convergence de l'algorithme.

L'algorithme EM bénéficie d'une preuve de convergence garantissant que l'itération de l'étape d'estimation et de maximisation converge vers un maximum (local et dépendant de la valeur des paramètres $\Theta^{(0)}$ choisis pour l'initialisation.

La fonction auxiliaire peut également s'écrire :

$$Q(\Theta,\Theta^{(t)}) = \sum_Y p(Y/X,\Theta^{(t)}) \log[p(X/\Theta)p(Y/X,\Theta)]$$

$$= \log p(X/\Theta) \sum_Y p(Y/X,\Theta^{(t)}) + \sum_Y p(Y/X,\Theta^{(t)}) \log p(Y/X,\Theta)$$

$$= \log p(X/\Theta) + \sum_Y p(Y/X,\Theta^{(t)}) \log[p(Y/X,\Theta)] \quad \ldots (1)$$

étant donné que $\sum_Y p(Y/X,\Theta^{(t)}) = 1$. Si on choisit $\Theta = \Theta^{(t)}$ on a alors :

$$Q(\Theta^{(t)},\Theta^{(t)}) = \log p(X/\Theta^{(t)}) + \sum_Y p(Y/X,\Theta^{(t)}) \log(Y/X,\Theta^{(t)}) \quad \ldots\ldots (2)$$

En soustrayant (1) de (2), et en regroupant les termes, on obtient alors :

$$\log p(X/\Theta) - \log p(X/\Theta^{(t)})$$

$$= Q(\Theta,\Theta^{(t)}) - Q(\Theta^{(t)},\Theta^{(t)}) + \sum_Y p(Y/X,\Theta^{(t)}) \log \frac{p(Y/X,\Theta^{(t)})}{p(Y/X,\Theta)}$$

Commentaire.

Le dernier terme n'est autre que la divergence de Kullback-Leibler [3], et est toujours positive. Par conséquent, si un changement de paramètres Θ augmente Q ce changement accroître également $\log p(X/\Theta)$.

En d'autres termes, lorsqu'on change les paramètres Θ de façon à maximiser l'espérance mathématique du logarithme de vraisemblance de la distribution jointe des données et variables cachées, nous maximisons également le logarithme de la vraisemblance des données.

En principe, on peut donc maximiser l'espérance mathématique $Q(\Theta,\Theta^{(t)})$ pour chaque valeur de Θ, et ensuite ré-estimer les paramètres, ce qui conduira à l'augmentation maximale de $p(X/\Theta)$ pour chaque itération.

4.4.4 Propriétés de la fonction de Log-vraisemblance.

On remarque 2 propriétés de cette fonction :

1. La log-vraisemblance $\log p\left(y_{1:T};\Theta^{(n)}\right)$ croit.
2. Les points d'accumulations possibles de la suite $\left(\Theta^{(n)}\right)_{n\geq 1}$ sont les points stables de log-vraisemblance $\left\{\Theta:\nabla_{\Theta}\log p\left(Y_{1:T};\Theta^{(n)}\right)=0\right\}$

4.5 Modèle de Markov caché (MMC).

Un modèle de Markov caché (MMC) ou Hidden Markov Model (HMM) est un graphe probabilisé dans lequel chaque nœud est censé produire un ou plusieurs segments stables ou transitoires du signal vocal. A chaque état ou nœud est associée une distribution de probabilité d'émettre un vecteur spectral ou cepstral (dans notre cas)

4.5.1 MMC Modèle génératif.

Les modèles MMC sont génératifs basés sur la vraisemblance d'une séquence d'un modèle M.

Le problème posé dans la reconnaissance est l'inverse :

étant donné une séquence $Y=\{y_1,y_2,\ldots,y_n,\ldots,y_N\}$ de longueur N et un modèle de Markov caché M dont les paramètres (A,B,Π) sont connus.

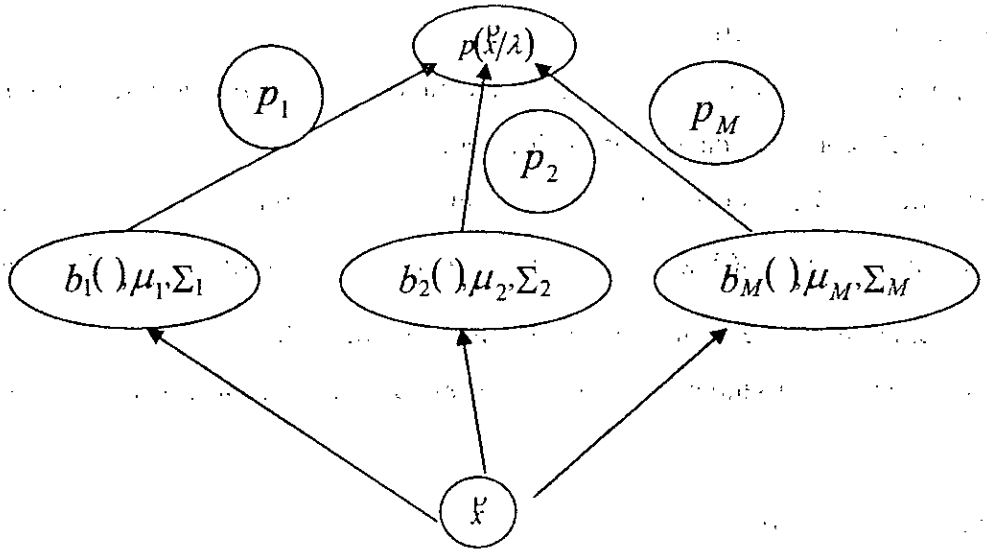
On désire connaître la séquence d'état $Q=\left(q^1,q^2,\ldots,q^n,\ldots,q^N\right)$ de M qui est la plus probable d'avoir généré Y.

Inconvénient.

- C'est une méthode qui caractérise un seul mode c'est-à-dire une seule moyenne donc elle ne prend pas exhaustivement les problèmes de coarticulation, conséquence identification médiocre.
- Elle a besoin d'une base d'entraînement très large donc un grand espace mémoire.
- Les problèmes de la segmentation et modélisation des états de la HMM en tenant compte de la variabilité du locuteur qui représente une caractéristique discriminante de ce locuteur.

5. La modélisation mixture gaussienne (GMM) pour l'identification de locuteur :

La mixture gaussienne est un ensemble de M classes, chacune d'elle représente une fonction gaussienne de moyenne $\vec{\mu}_i$, matrice de covariance Σ_i, et un poids p_i, $i = 1,...,M$.



Graphe3.1 représentation de la GMM à M modèles

Pour un vecteur aléatoire x de dimension D, la densité de x, en passant par le modèle λ est donnée par la somme : $p(\vec{x}|\lambda) = \sum_{i=1}^{M} \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)\right\}$.

Les poids p_i, $i = 1..M$,vérifient : $\sum_{i=1}^{M} p_i = 1$.

Chaque locuteur donc, est modélisé par une mixture gaussienne $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}\ i = 1..M$.

La puissance de modélisation par la mixture gaussienne est évidemment très forte, elle divise l'espace de chaque locuteur en un ensemble de sous ensembles, chaque sous ensemble est balayé par une fonction gaussienne, donc c'est une bonne modélisation car un seul locuteur possède une probabilité de points qui augmente la performance de l'identification de ce locuteur.

6. Estimation des paramètres par la maximisation de la fonction vraisemblance :

Avant d'utiliser la GMM, il faut trouver, pour chaque locuteur, les paramètres de la mixture gaussienne liée à ce locuteur, cette étape est l'entraînement du model. Donc on fait entraîner notre système sur le locuteur correspondant qui sera représenté par ce système au cours de l'identification. Pour ce faire, on suppose la séquence d'entraînement de T vecteurs : $X = \{x_1, x_2, ..., x_T\}$.

On suppose que les échantillons du vecteur sont indépendants et identiquement distribués.
La fonction de vraisemblance, qu'on cherche à maximiser est donnée donc par :

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda).$$

Mais malheureusement cette fonction non linéaire en fonction de λ, est difficile à maximiser. La solution proposée consiste donc à exploiter la technique EM (*Estimation Maximisation*) pour trouver un modèle correspondant à chaque locuteur.
L'idée principale de l'EM est de commencer par un modèle initial λ, puis on estime un nouveau modèle $\bar{\lambda}$ vérifiant $p(X|\bar{\lambda}) \geq p(X|\lambda)$, ce dernier sera initial pour l'estimation suivante et le processus se répète jusqu'à la convergence de l'algorithme.
En fonction des vecteurs de la séquence X, on donne les paramètres de la GMM, entraîné par l'algorithme EM :

- Les poids : $\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)$.
- Les moyennes : $\bar{\mu}_i = \frac{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda).\vec{x}_t}{\sum_{t=1}^{T} p(i|\vec{x}_t, \lambda)}$.
- Variances : $\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i/\vec{x}_t, \lambda)x_t^2}{\sum_{t=1}^{T} p(i/\vec{x}_t, \lambda)} - \bar{\mu}_i^2$.
- $p(i|\vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)}$. C'est la probabilité a posteriori de la classe i.

7. L'identification :

Pour l'identification de locuteur, un groupe de S locuteurs $s = \{s_1, s_2, ..., s_S\}$, représenté par S modèle GMM $\lambda_1, \lambda_2, ..., \lambda_S$. L'objectif est de trouver la personne désirée parmis les S locuteurs qui vient de prononcer la phrase qui est en cours de traitement, donc cette recherche conduit à trouver le modèle ayant la probabilité a posteriori la plus grande :
Donc on écrit $\hat{S} = \arg\max_{1 \leq k \leq S} \Pr(\lambda_k|X)$.

Et selon la loi de Bayes : $\hat{S} = \arg\max_{1 \leq k \leq S} \frac{p(X|\lambda).\Pr(\lambda_k)}{p(X)}$ [2].

Dans le cas où tous les locuteurs sont équiprobable, et $p(X)$ est la même pour tous les locuteurs : $\hat{S} = \arg\max_{1 \le k \le S} p(X|\lambda_k)$.

En utilisant la propriété de l'indépendance entre les observations, on trouve :

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\overset{\rho}{x}_t|\lambda_k)$$

Conclusion.

Nous avons décrit les différents modèles de classification, le modèle algébrique :la quantification vectorielle et modèle statistique : la mixture gaussienne (la GMM).
Cette description a pour objet l'application de ces différents algorithmes sur une base de données et les évaluer en fonction des paramètres choisis. Une étude comparative s'impose.

Chapitre 4 : Simulation & Résultats

Introduction

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

1. L'analyse :

L'analyse est l'opération qui permet l'extraction des informations utiles à partir du signal parole, et puisqu'on est intéressé par l'information relative à l'identité du locuteur, et d'après l'étude théorique, on a choisi d'utiliser les coefficients MFCC's comme paramètres d'identification.
La figure suivante illustre les étapes suivies afin d'extraire les coefficients MFCC's :

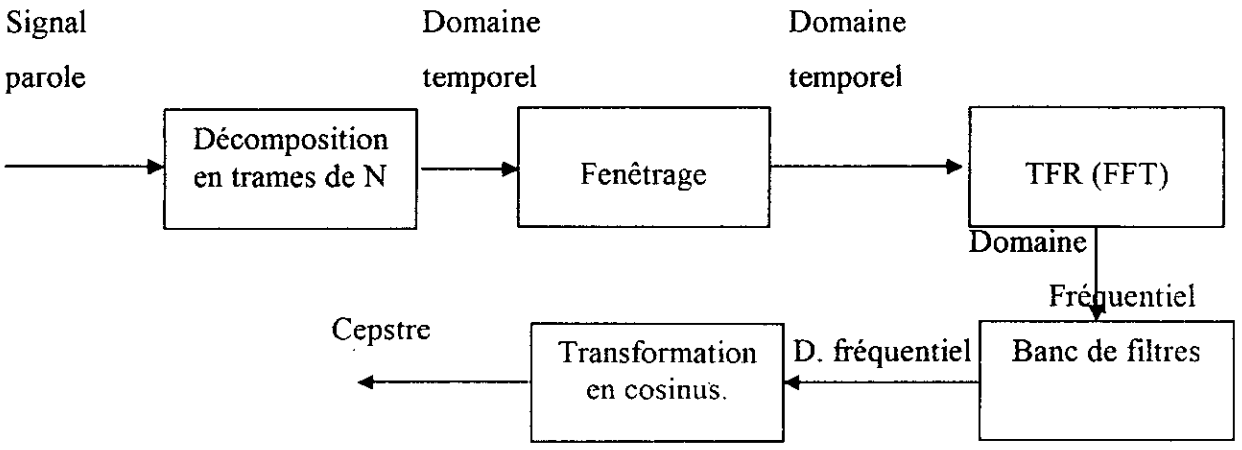


Schéma 4.1 : Analyse cepstrale

Chaque bloc est simulé avec une fonction MATLAB contenant des variables d'entrée du signal à traiter, le résultat de cette fonction est une donnée à traiter par la seconde fonction. Les entrées et les sorties sont, soit des vecteurs, soit des matrices à traiter.

1.1. La décomposition du signal en trames :

Dans cette étape le son articulé continu est stocké dans des trames de N échantillons, avec les trames adjacentes séparé par M (M < N). La première tranche comprend les premiers échantillons *de N*. La deuxième trame commence de l'échantillon M après la première, avec un chevauchements de N -M échantillons. Ce procédé continue jusqu'a ce que tout le discours soit décomposé en une ou plusieurs trames.

1.2. Fenêtrage :

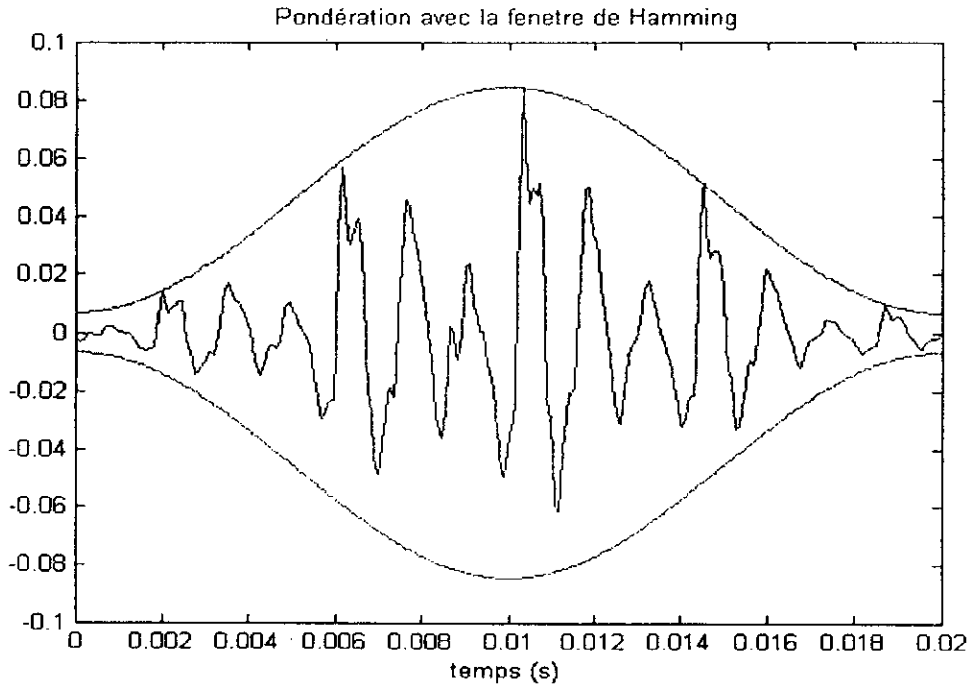
La prochaine étape dans le traitement est le fenêtrage. On utilise la fenêtre de Hamming afin de limiter le nombre d'échantillons. En effet cette fenêtre a pour propriétés de réduire les effets de bord (ou de Gibbs).

Si nous définissons la fenêtre comme$W(n)$, $0 \le n \le N-1$, où N est le nombre des échantillons dans chaque tranche, alors le résultat du fenêtrage est le signal

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1$$

La fenêtre de *Hamming* est :

$$W(n) = 0.54 + 0.46 * \cos\left(\frac{2\pi n}{N-1}\right).$$



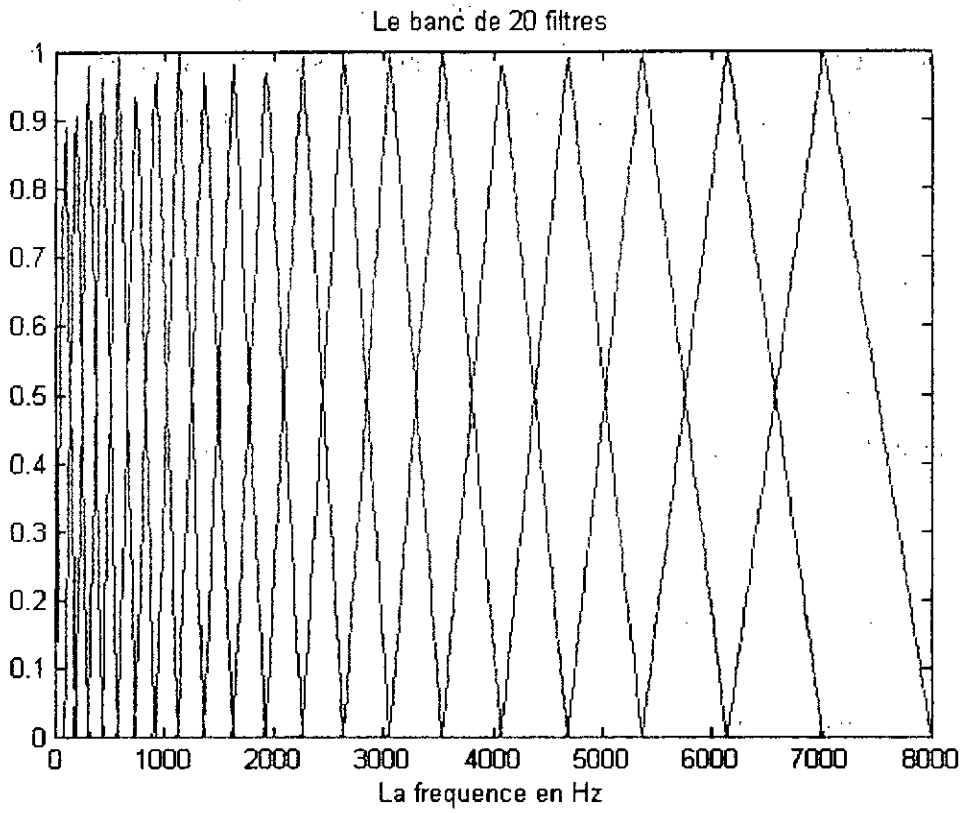
Graphe 4.1 : Fenêtre de Hamming

1.3. La transformée de Fourier :

La transformée de Fourier nous permet le passage de l'espace temps à l'espace des fréquences. La transformée de Fourier rapide (FFT) est toujours exploitée pour assurer une rapidité de simulation.

1.4. Filtrage par un banc de filtres :

On génère l'ensemble des filtres triangulaires, chaque filtre opère sur une bande de fréquence avec un chevauchement entre les filtres adjacents, l'énergie du signal est calculée à la sortie de chaque filtre.



Graphe 4.2 Banc de filtres (20 filtres)

1.5. Le cepstre réel :

Dans cette dernière étape, nous convertissons le log spectre de *mels* et nous travaillons dans l'espace quéfrentiel.

La transformation cosinus effectuée est donnée par :

$$\tilde{c}_n = \sum_{k=1}^{K} (\log \tilde{S}_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \qquad n = 1,2,...,K$$

Où $\tilde{S}_k$ sont les énergies calculées après le passage par les filtres.

1.6 Détection de silence / parole

Pour enlever les périodes de silence, on utilise un filtre qui les sélectionne d'après leurs niveaux d'énergie par rapport à l'énergie totale de la séquence. Toutes trames du niveaux énergétique inférieur à un seuil déterminé pratiquement après une étude statistique faite sur les périodes parole / silence de notre base de données.

2. Description de la simulation

L'identification est dite préalable c'est-à-dire que chaque locuteur produit un segment de parole de durée suffisante pour l'identification (Dans notre cas minimum 14 s d'enregistrement). L'entraînement se fait sur ce segment ou plusieurs qui a permis de déterminer les coordonnées du système-parole adéquate.

2.1 La base de donnée :

La base donnée utilisée pour simuler les systèmes d'identification est extraite principalement de la base TIMIT. La base TIMIT est une collection de 6300 phrases parlées par 630 locuteurs, ces locuteurs sont de 8 régions différentes des Etats-Unis.
Les données sont échantillonnées avec une fréquence de 16KHz. Notre base est organisée sous formes de deux répertoires : répertoire entraînement et test.

2.2 Organisation de la base de données

La base est un mélanges de fichiers (extension wav) parlés par des locuteurs (34) et locutrices (14). Les locuteurs sont étiquetés de 1 à 48 (par exemple speaker2) et les phrases de chacun sont contenues dans un répertoire portant l'étiquette du locuteur correspondant. Les phrases sont de durée moyenne de 3s chacune. Pour avoir un enregistrement minimum de 14 s à l'entraînement, on a concaténé 7 phrases parmi les 10 de chaque locuteur et les 3 restantes sont à utiliser pour le test.

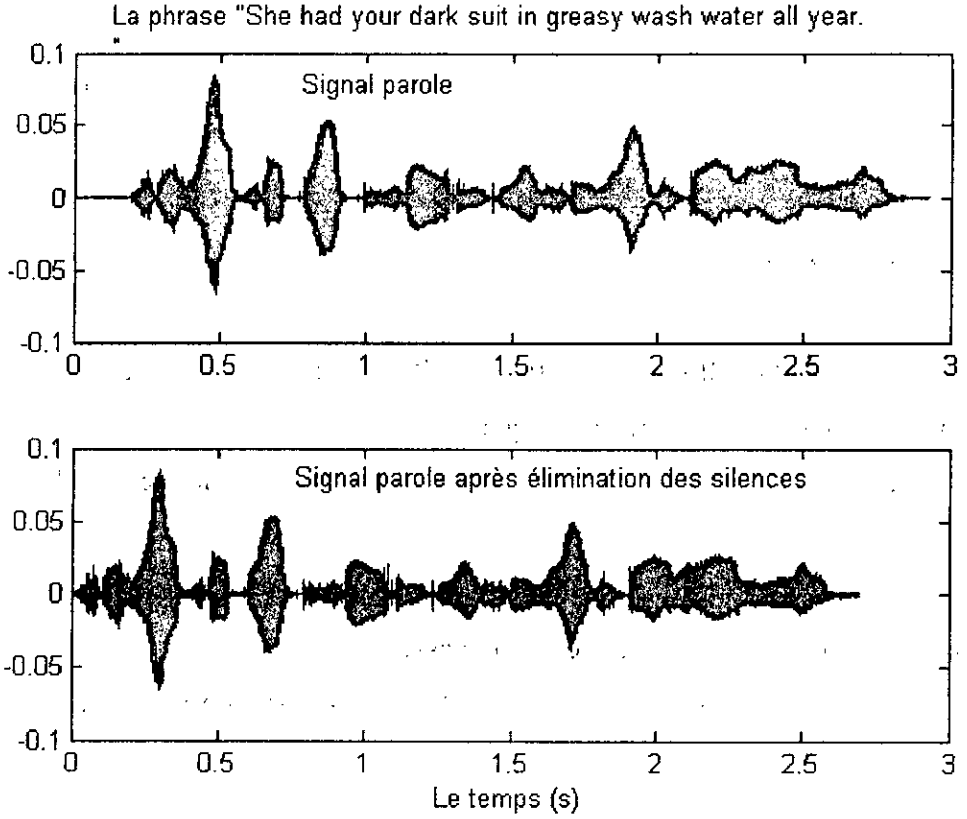


Figure 4.1 : Détection et élimination de silence

2.3 Paramètres

Dans cette simulation qui considère le modèle de perception, on va évaluer l'influence des paramètres suivants : nombre de locuteurs, nombre de classes et dimension du vecteur MFCC sur l'identification.

2.3.1 Nombre de locuteurs

On évalue le système sur 16, 32 et 48 locuteurs. Il est évident que les performances vont diminuer avec l'augmentation du nombre de locuteurs donc on cherche un nombre maximum pour un pourcentage d'identification acceptable.

2.3.2 Nombre de classes

Le nombre de classes est variable de 4 à 50 classes. Ce choix est pris selon la variation du taux d'identification et la convergence de l'algorithme au voisinage de ces classes.

2.3.3 Dimension du vecteur

La quantité d'information contenue dans les harmoniques du signal est modifiée par la dimension du vecteur MFCC. Cette dimension varie de 10 à 40.

2.4 La fréquence d'échantillonnage

L'étude se fait à fréquence d'échantillonnage de 16 kHz. Pour évaluer la robustesse du modèle, on analysera son comportement dans la bande téléphonique.
Notre base de données est échantillonnée à 16 kHz, on l'a donc réechantillonnée à 8 kHz en tenant compte de la bande téléphonique.

Pour l'échantillonnage à 8 kHz, le signal est filtré par un passe bande [300 - 4300] Hz « Bande téléphonique », puis on fait la décimation à 50 %, c'est-à-dire, on prend un échantillon de signal parmi deux.

2.5 Estimation des résultats.

Après le test, l'analyse du résultat. On estime le taux d'identification des locuteurs.
La performance finale d'une identification est calculée d'après le test, en divisant le nombre des segments correctement identifiés sur le nombre total des segments :

$$\%identification\ correcte = \frac{Nombre\ de\ segments\ correctements\ identifiés}{Nombre\ total\ des\ segments}.100$$

2.6 Langage utilisé

Langage de programmation utilisé est MATLAB 6p5. C'est un langage qui possède des bibliothèques dédiées au traitement de signal.

3. Simulation et résultats

3.1 La quantification vectorielle

3.1.1 Fréquence d'échantillonnage 16 kHz

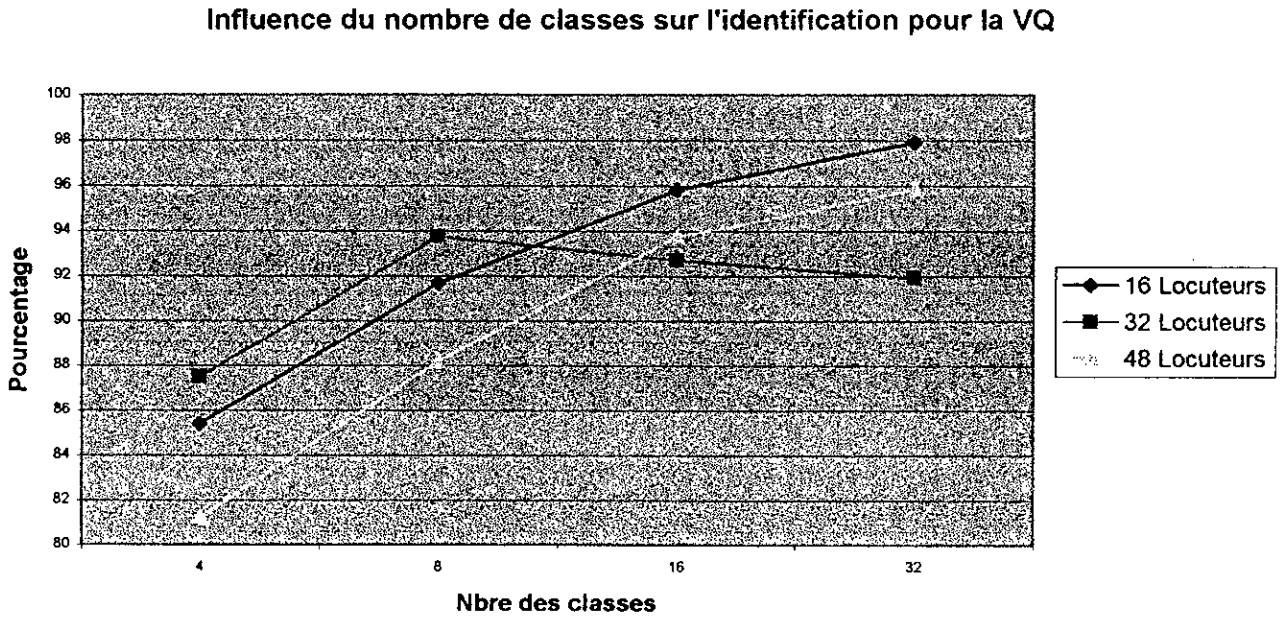
3.1.1.1 Influence du nombre de classes

Le choix important dans la QV est le nombre de centroides c'est-à-dire le nombre de classes. Un compromis sur ce choix s'impose. Un choix d'un petit nombre de classes cause la perte d'information donc dégradation de l'identification. Un choix d'un grand nombre induit de la redondance et cause la saturation de la mémoire.

Nbre de locuteurs	16				32				48			
Nbre de classes	4	8	16	32	4	8	16	32	4	8	16	32
Pourcentage (%)	85,41	91,67	95,83	97,91	87,5	93,75	92,71	91,92	81,25	88,19	93,8	95,8

Tableau 4.1: Influence du nombre de classes

Le graphe :



Graphe 4.3 : Influence du nombre de classes

Commentaire

Le minimum de pourcentage d'identification pour tout les locuteurs est très proche les uns des autres.
D'après les résultats, la courbe est croissante dans l'intervalle de nombre des classes.
Donc, le compromis dans ce cas est évident et dépend de l'application envisagée.

Conclusion

Pour une application de haute précision ou pour une identification d'une base de données large, il faut donc entraîner un nombre important de classes. Inconvénient, le temps pour l'entraînement et le test augmente.

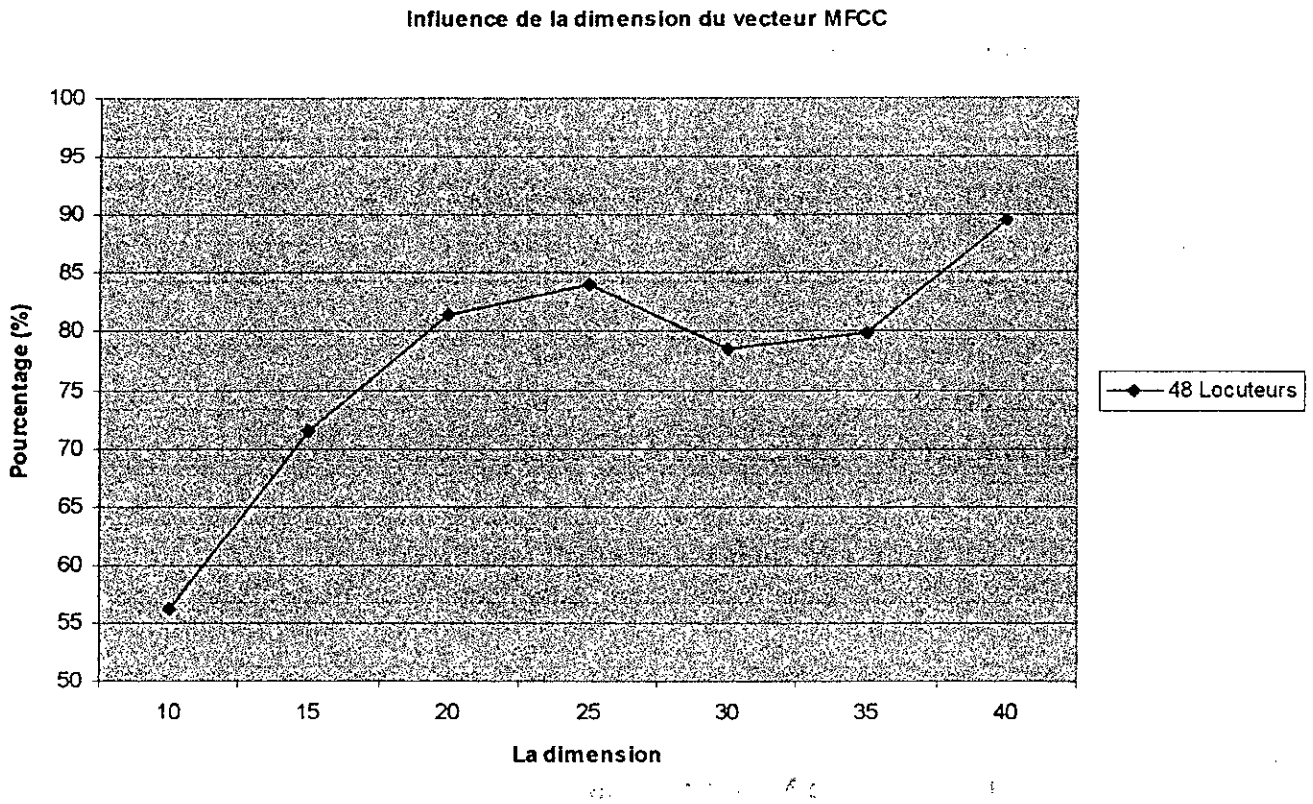
3.1.1.2 Influence de la dimension du vecteur MFCC

On a choisit le point le plus dégradé de l'identification (d'après l'étape précédente). Ce point correspond au :
Nombre de classes = 4
Nombre de locuteurs = 48

Nbre de classes = 4							
Nbre de Locuteurs = 48							
Nbre de MFCC	10	15	20	25	30	35	40
Pourcentage (%)	56,25	71,52	81,52	84,03	78,47	79,86	89,58

Tableau 4.2 : Influence de la dimension du vecteur MFCC

Le graphe :



Graphe 4.4 : Influence de la dimension du vecteur MFCC

Commentaire

La courbe est croissante. Le minimum d'identification est à 56.25 %. On remarque aussi un maximum local dans l'intervalle [20 – 25].

Conclusion

Ce maximum local (84.03 %) est pris en considération car :

-1. On gagne en temps de calcul : l'entraînement se fait pour un nombre de classe variant entre 20 et 25.

-2. Le modèle de classification est presque stable dans cet intervalle et ne perturbe pas le taux d'identification.

L'entraînement dans ce cas là est optimal pour un nombre de classe variant entre [20 – 25].

3.1.2 La fréquence d'échantillonnage 8 kHz

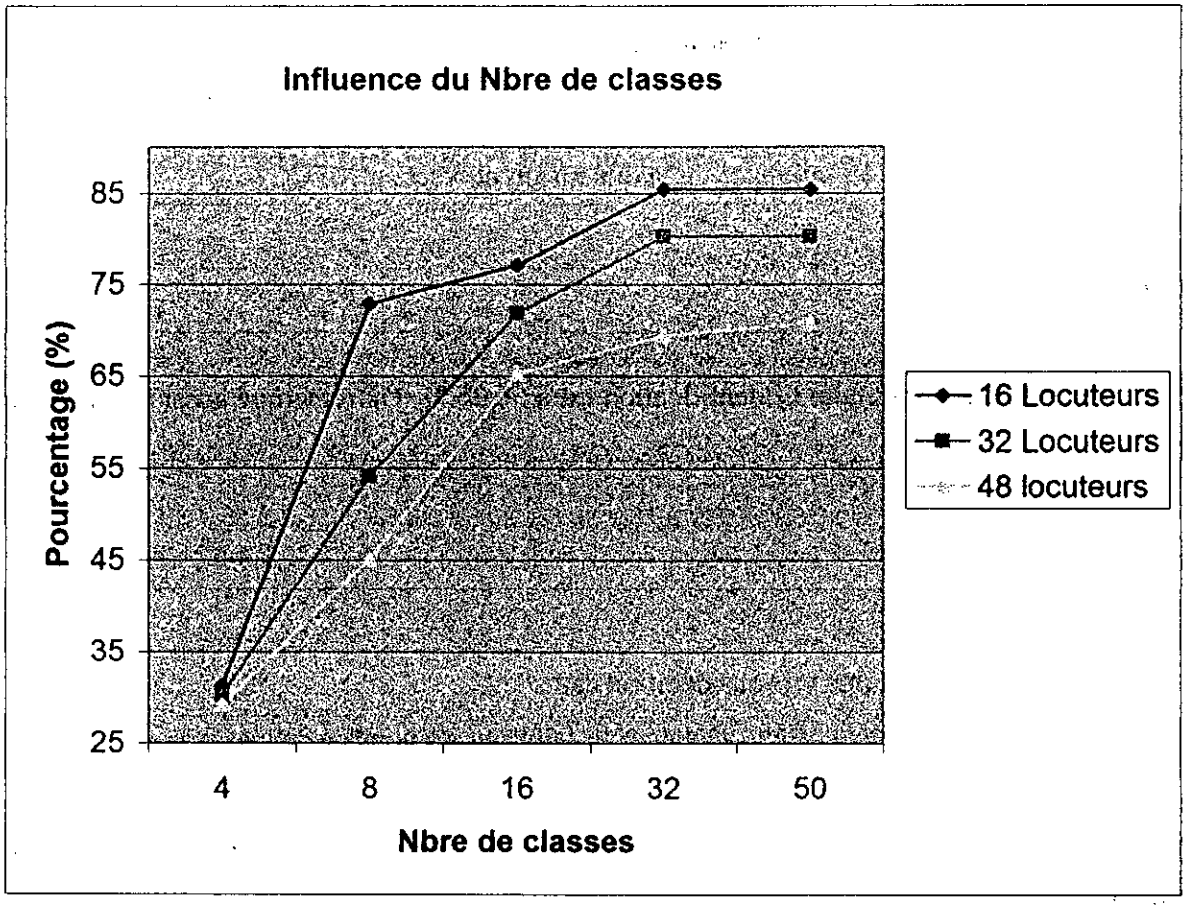
3.1.2.1 Influence du nombre de classes

Nbre de Locuteurs	16					32				
Nbre de Classes	4	8	16	32	50	4	8	16	32	50
Pourcentage (%)	31,25	72,92	77,08	85,42	85,42	30,21	54,17	71,875	80,21	80,21

Nbre de Locuteurs	48				
Nbre de Classes	4	8	16	32	50
Pourcentage (%)	29,17	45,14	65,28	69,44	70,83

Tableau 4.3 : Influence du nombre de classes

Le graphe :



Graphe 4.5 : Influence du nombre de classes

Commentaire

Les courbes des trois locuteurs sont croissantes dans l'intervalle de nombre de classes.
Le minimum d'identification est 29.17 % pour une base de 48 locuteurs, le maximum atteint est 85.42 % pour 16 locuteurs.

Conclusion

Le pourcentage d'identification dans la bande téléphonique est très acceptable pour une base de données de 16 locuteurs, il est acceptable pour 32 locuteurs. Par contre Il est moins acceptable pour 48 locuteurs.

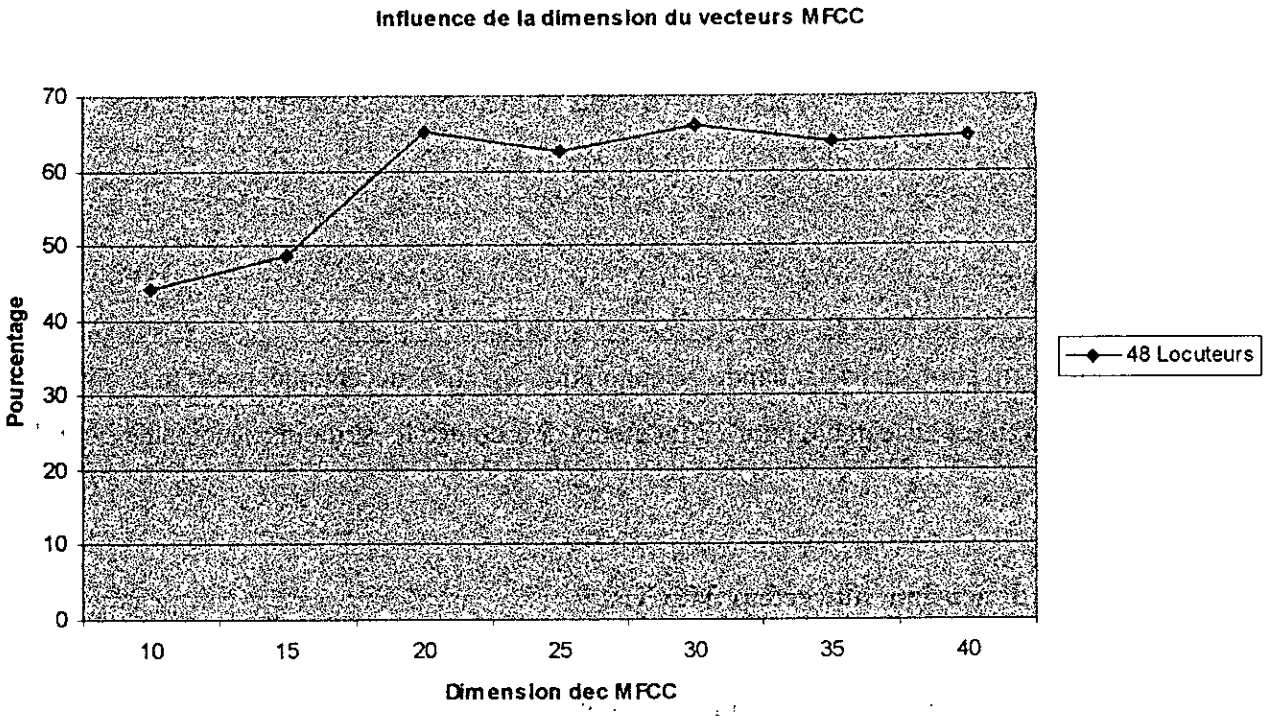
3.1.2.2 Influence de la dimension du vecteur MFCC

On a choisit un point de l'identification. Ce point correspond au :
Nombre de classes = 16
Nombre de locuteurs = 48

Nombre de classes = 16							
Nombre de Locuteurs = 48							
Nombre de MFCC	10	15	20	25	30	35	40
Pourcentage (%)	44,065	48,61	65,28	62,5	65,97	63,89	64,58

Tableau 4.5 : Influence de la dimension du vecteur MFCC

Le graphe :



Graphe 4.6 : Influence de la dimension

Commentaire

C'est une courbe croissante dont le minimum est 44.065 % pour un entraînement d'un vecteur de dimension 10, le maximum est 65.97 % pour un entraînement d'un vecteur de dimension variant entre 20 et 30. On remarque dans ce cas là, l'existence d'un maximum local dans l'intervalle [20 – 25] et un autre dans [25 – 30].

Conclusion

La dimension favorable c'est-à-dire celle qui donne un pourcentage d'identification acceptable appartient à l'intervalle [20 – 25] car, à 16 kHz, on a un maximum local stable dans l'intervalle [20 - 25].

Compromis

Dans cette méthode de classification, on a conclu que le nombre de classes dépend de l'application envisagée et l'entraînement de notre base de données se fait sur vecteur MFCC de dimension variant entre 20 et 25.

3.2 La mixture gaussienne (GMM):

3.2.1 Fréquence d'échantillonnage 16 kHz

3.2.1.1 Influence du nombre de classes

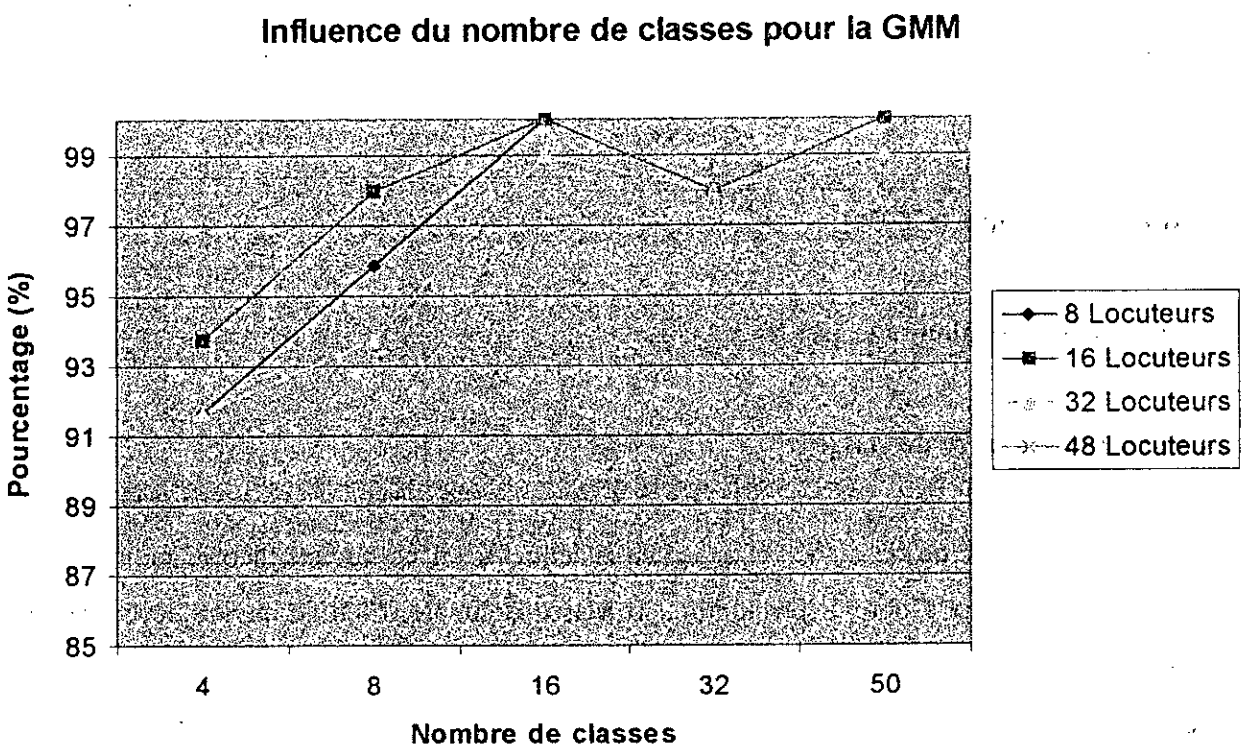
La détermination du nombre de classes est une tache très délicate, il n'y a aucune théorie au préalable qui nous indique la manière d'obtenir un nombre de classe optimale. L'étude envisagée ci-dessous nous indique la démarche pour obtenir ce nombre de classes.

Nombre de Locuteurs	8			16					32				
Nombre de classes	4	8	16	4	8	16	32	50	4	8	16	32	50
Pourcentage (%)	91,67	95,83	100	93,75	97,92	100	97,92	100	91,67	93,75	98,96	97,92	98,96

Nombre de Locuteurs	48				
Nombre de classes	4	8	16	32	50
Pourcentage (%)	86,11	90,97	97,92	97,92	97,92

Tableau 4.6 : Influence du nombre de classes

Le graphe :



Graphe 4.6 : Influence du nombre de classes

Commentaire

Les courbes sont croissantes dans [4 – 16]. Elles possèdent un maximum pour un nombre de classes de 16.

Pour 8 locuteurs le pourcentage est 100 %.

Pour 16 locuteurs le pourcentage est 100 %.

Pour 32 locuteurs le pourcentage est 98.96 %.

Pour 48 locuteurs le pourcentage est 97.92 %.

Conclusion

Nombre de classe = 16 est le nombre optimal pour une identification maximale pour tout les locuteurs. On a besoin donc, pour notre base de données de 48 locuteurs d'un entraînement sur 16 classes.

3.2.1.2 Influence de la dimension du vecteur MFCC

On a choisit un point de l'identification.

Ce point correspond au :

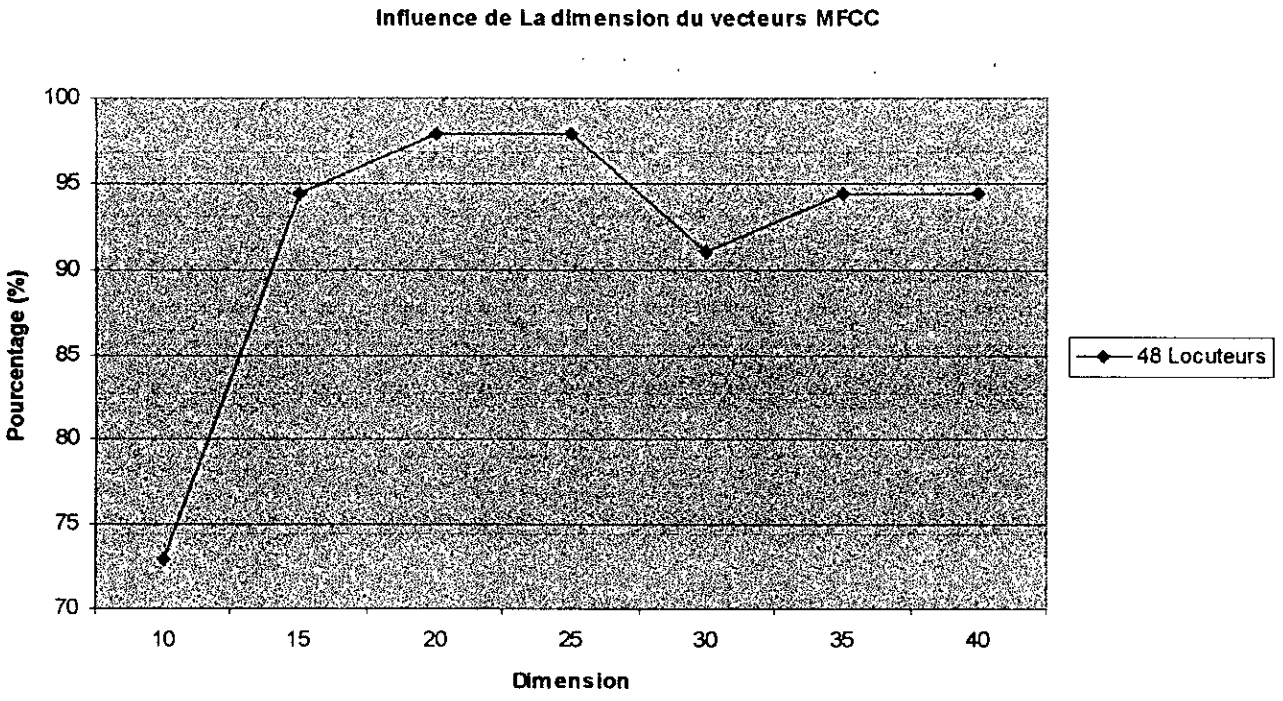
Nombre de classes = 16

Nombre de locuteurs = 48

Nombre de classes = 16							
Nombre de Locuteurs = 48							
Dimension du MFCC	10	15	20	25	30	35	40
Pourcentage (%)	72,92	94,44	97,92	97,92	90,97	94,44	94,44

Tableau 4.7 : Influence de la dimension du vecteur MFCC

Le graphe :



Graphe 4.7 : Influence de la dimension du MFCC

Commentaire

Courbe croissante dont le minimum est 72.92 % pour une dimension 10, le maximum est pour une dimension variant entre [20 – 25]. Ce maximum est un maximum local.

Conclusion

Le pourcentage d'identification le plus favorable est atteint pour une dimension variant toujours d'après le graphe, dans l'intervalle [20 – 25].

3.2.2 Fréquence d'échantillonnage 8 kHz

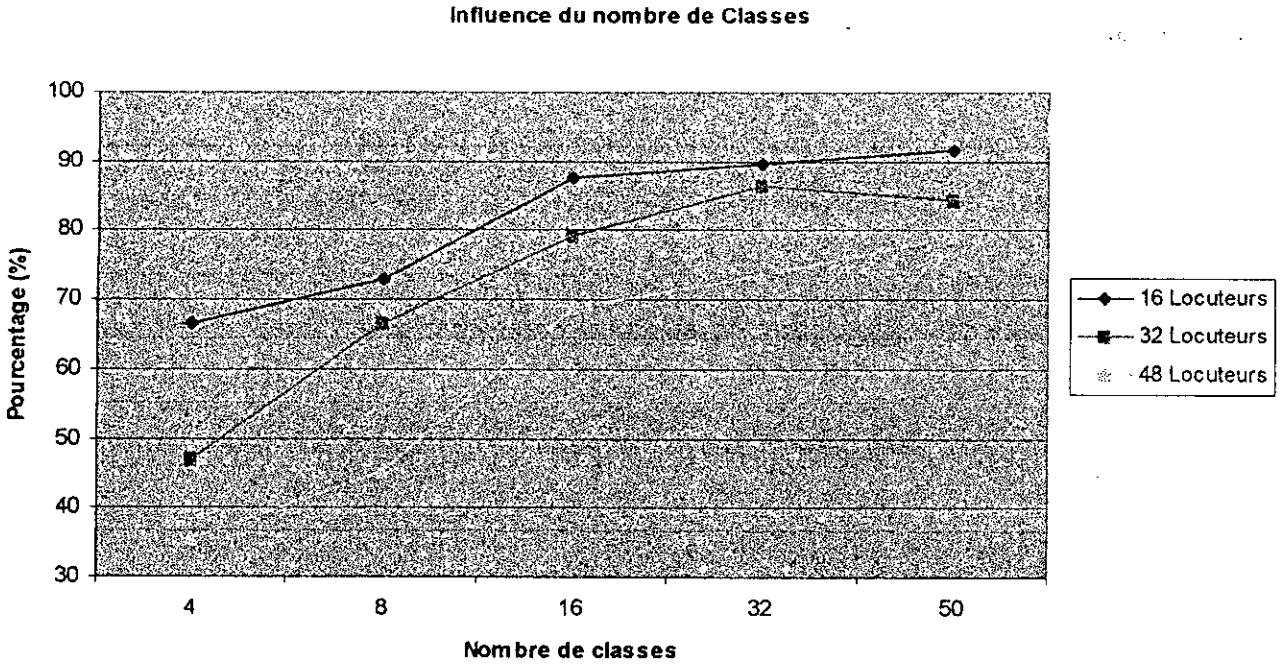
3.2.2.1 Influence du nombre de classes

Nbre de Locuteurs	16					32				
Nbre de classes	4	8	16	32	50	4	8	16	32	50
Pourcentage (%)	66,67	72,92	87,5	89,58	91,67	46,875	66,67	79,17	86,46	84,375

Nbre de Locuteurs	48				
Nbre de classes	4	8	16	32	50
Pourcentage (%)	35,42	45,14	68,75	73,61	79,17

Tableau 4.8 : Influence du nombre de classes

Le graphe :



Graphe 4.8 : Influence du nombre de classes

Commentaire

Les courbes représentées sont croissantes. Le minimum d'identification est de 35.42 % pour un nombre de locuteurs 48 et un nombre de classes 4, le maximum est de 91.67 % pour un nombre de locuteurs 16 et nombre de classes 50.

Conclusion

Le pourcentage d'identification est évalué dans ce cas là selon l'application envisagée. C'est-à-dire, pour une application de haute précision on a besoin d'un entraînement sur un nombre de classes élevé.

3.2.2.2 Influence de la dimension du vecteur MFCC

On a choisit un point de l'identification. Ce point correspond au :

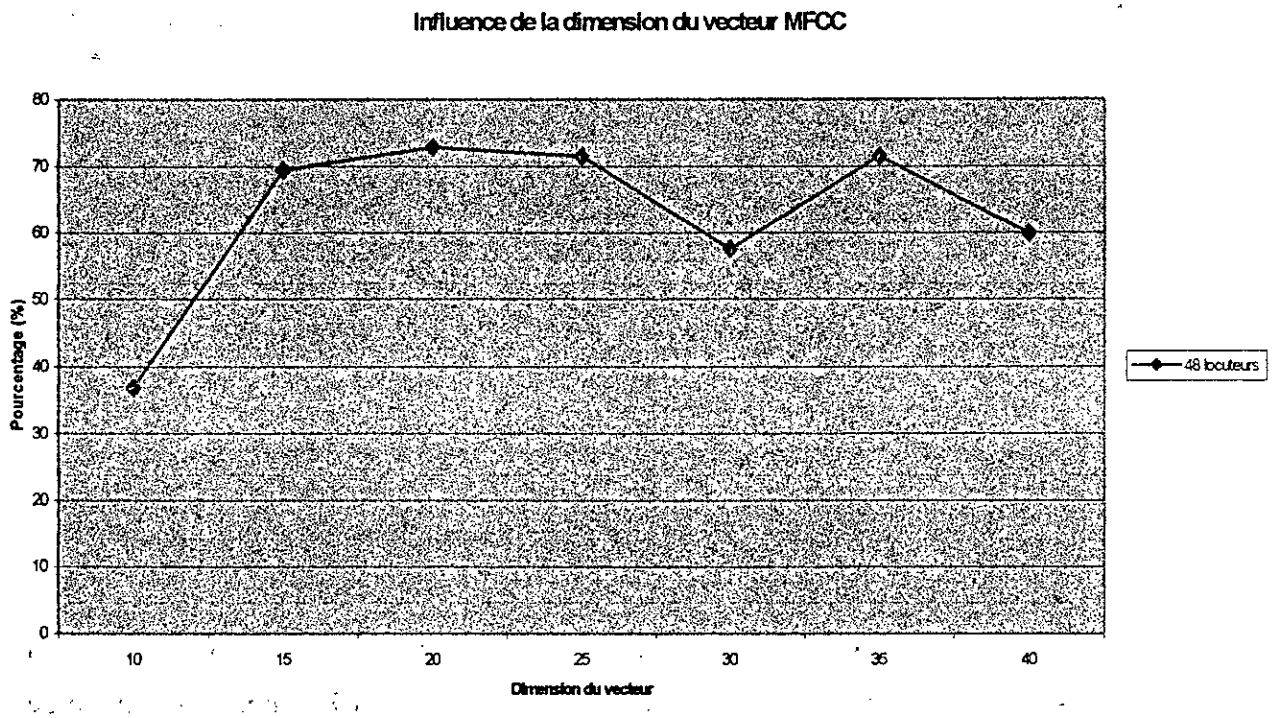
Nombre de classes = 16

Nombre de locuteurs = 48

Nombre de Locuteurs	48						
Dimension des vecteurs MFCC	10	15	20	25	30	35	40
Pourcentage (%)	36,81	69,44	72,92	71,52	57,64	71,53	59,94

Tableau 4.9 : Influence de la dimension des vecteurs MFCC's

Le graphe



Graphe 4.9 : Influence de la dimension du vecteur MFCC

Commentaire

Courbe croissante dont le minimum est 36.81 % pour une dimension 10, le maximum est pour une dimension variant entre [20 – 25]. Ce maximum est un maximum local.

Conclusion

Le pourcentage d'identification le plus favorable est atteint pour une dimension variant toujours d'après le graphe, dans l'intervalle [20 – 25]. Ce maximum est pris en considération d'après :

1. réduction du temps de traitement : l'entraînement se fait pour un nombre de classes variant entre 20 et 25.
2. Le modèle de classification est presque stable dans cet intervalle et ne dégrade pas le pourcentage d'identification.

3.3 Etude comparative

On a testé deux approches différentes : algébrique et statistique.

Un système d'identification basé sur la quantification vectorielle (QV) donne de bonnes performances, aussi la QV à l'avantage d'être simple et rapide à implémenter.

Son inconvénient la dégradation de la performance pour un nombre important de locuteurs ce qui nous oriente vers une approche mixture gaussienne (GMM).

Cette approche assure de bonnes performances même pour un nombre de locuteurs important, pour cela, on remarque que son pourcentage est plus élevé que celui de la quantification vectorielle. On remarque aussi l'existence d'un nombre optimale de classes pour l'entraînement de la GMM à l'opposé de la VQ qui stipule qu'en augmentant en nombre de classes en a une meilleure identification. Ce qui n'est pas le cas.

La dimension du vecteur MFCC est dans les deux approches variables dans l'intervalle [20 – 25] pour une identification optimale.

L'inconvénient de la GMM est la convergence. Il faut s'assurer au préalable de la convergence de l'algorithme au cours de l'entraînement.

4. Conclusions

L'influence de la fréquence d'échantillonnage sur l'identification dégrade son pourcentage. La moyenne de l'identification pour les deux algorithmes reste acceptable.

Les recherches actuelles dans l'identification du locuteur ont fixé des buts, améliorer le taux d'identification dans la bande téléphonique [300 – 4300] Hz et descendre dans les fréquences, au dessous des 4000 Hz. Ces recherches étendront les applications de l'identification dans les domaines de communication.

Notre travail a pour perspective de continuer l'étude de l'identification en se penchant également sur le choix du corpus qui repose sur une connaissance approfondie de l'expert dans le domaine de la phonétique et la linguistique. Rechercher aussi une méthode optimale de segmentation de la séquence d'enregistrement par des inter-silences et évaluer la performance d'identification.

Une autre méthode intéressante est la méthode hybride. Elle utilise deux approches différentes (par exemple : entraînement HMM (réseaux de neurones) et test GMM. Les résultats de performance peuvent être attendus plus intéressants.

Les algorithmes sont écrits en langage MATLAB 6p5 ce qui limite la rapidité d'exécution et notamment l'implémentation.
Dans le cadre de notre travail, le temps réel n'est pas une exigence.
La programmation en Borland C permet d'enlever cet inconvénient et facilite par conséquent, l'implémentation.

Annexes

Annexe A : Estimation des paramètres LPC et des formants

1. Estimation du modèle autorégressif (La prédiction linéaire)

Elle est basée sur l'hypothèse que chaque échantillon du signal original $x(n)$ peut être approché par une combinaison linéaire des p échantillons qui le précédent :

$$x(n)=-a(1)x(n-1)-a(2)x(n-2)-\ldots-a(p)x(n-p)+e(n) \qquad (2.1)$$

$$\text{Donc} \qquad e(n)=\sum_{i=0}^{p}a(i)x(n-i) \quad a(0)=1 \qquad (2.2)$$

$e(n)$: l'erreur de prédiction ou résidu d'ordre p .

- **Variance de l'erreur de prédiction.**

L'estimation des coefficients a_i est basée sur la minimisation de la variance de l'erreur de prédiction :

$$\sigma_e^2=E\left[e(n)^2\right]=E\left[\sum_{i=0}^{p}a(i)x(n-i)\sum_{j=0}^{p}a(j)x(n-j)\right]$$

$$=E\left[\sum_{i,j=0}^{p}a(i)a(j)x(n-i)x(n-j)\right]$$

$$=\sum_{i,j=0}^{p}a(i)a(j)\phi_x(i-j) \qquad (2.3)$$

$\phi_x(k)$ représente la fonction d'autocorrélation du signal x :

$$\phi_x(k)=E[x(n)x(n+k)]=\lim_{N\to\infty}\frac{1}{2N+1}\sum_{i=N+k}^{N-k}x(n)x(n+k); \qquad (2.4)$$

$$\phi_x(0)=\sigma_x^2$$

Remarque.

La moyenne de x est supposée nulle : une composante continus ne porte aucune information utile et on peut l'extraire avec un filtre très simple.

Le vecteur des coefficients de prédiction d'ordre p sera noté :

$$a=[1,a(1),a(2),\ldots,a(p)]^T=[1,a^T]^T \quad (2.5)$$

et la matrice d'autocorrélation de $x(n)$:

$$\phi_x=\begin{bmatrix} \phi_x(0)\phi_x(1)..\phi_x(p-1) & \phi_x(p) \\ \phi_x(1)\phi_x(0).. \quad \ldots & \phi_x(p-1) \\ \vdots & \vdots \\ \vdots & \phi_x(1) \\ \phi_x(p) \quad \ldots \quad \ldots & \phi_x(0) \end{bmatrix} \quad (2.6)$$

Selon (2.3) et (2.5), la variance σ_e^2 de l'erreur de prédiction peut s'écrire :

$$\sigma_e^2=[1,a^T]\phi_x[1,a^T]^T \quad (2.7)$$

C'est une forme quadratique définie positive en les $a(i)$, ce qui assure l'unicité de son minimum.

Remarque.

La matrice d'autocorrélation possède une structure très particulière : les éléments situés le long de chaque diagonale parallèle à la diagonale principale sont égaux; une telle matrice est appelée matrice de Toeplitz. C'est aussi une matrice symétrique.

- **Estimation des coefficients de prédiction.**

On dérive l'expression (2.7) par rapport aux coefficients $a(i)$, $(i=1,2,\ldots p)$ on obtient un système d'équations linéaires en $a(i)$.

Cette résolution rapide est basée sur une méthode récurrente sur l'ordre de la prédiction.

ϕ_x de (2.5) sera notée $\phi^{(p)}$.

$[1,a]^T$ sera notée $[1,a_p]^T$.

$$\phi^{(p)}=\begin{bmatrix} \sigma_x^2 & \phi^{(p)T} \\ \phi^{(p)} & \phi^{(p-1)} \end{bmatrix} \quad (2.8)$$

La forme quadratique (2.7) peut s'écrire

$$\sigma_e^2=\begin{bmatrix}1, a_p^T\end{bmatrix}\begin{bmatrix} \sigma_x^2 & \phi^{(p)T} \\ \phi^{(p)} & \phi^{(p-1)} \end{bmatrix}\begin{bmatrix}1 \\ a_p\end{bmatrix} \quad (2.9)$$

$$=\sigma_x^2+2\phi^{(p)T}a_p+a_p^T\phi^{(p-1)}a_p$$

On a donc

$$\frac{\partial\sigma_e^2}{\partial a_p}=2\phi^{(p)}+2\phi^{(p-1)}a_p=0 \quad (2.10)$$

$$\phi^{(p-1)}a_p=-\phi^{(p)} \quad (2.11)$$

Soit sous la forme développée

$$\phi_x(k)=\sum_{i=1}^{p}a_p(i)\phi_x(k-i) \quad \text{pour } k=1,2,\ldots,p \quad (2.12)$$

$$\text{avec } \phi_x(0)=\sigma_x^2$$

La valeur minimisée de la variance de l'erreur de prédiction σ_e^2 qui sera notée

$$\sigma_p=\sigma_{e,\min}^2$$

$$\sigma_{e,\min}^2=\sigma_x^2+\phi_x^{(p)T}a_p=\sum_{i=0}^{p}a_p(i)\phi_x(i)=\alpha_p \quad (2.13)$$

D'autre part, si l'on réduit (2.11) et (2.13), on obtient :

$$\begin{bmatrix} \sigma_x^2 & \phi^{(p)T} \\ \phi^{(p)} & \phi^{(p-1)} \end{bmatrix}\begin{bmatrix}1 \\ a_p\end{bmatrix}=\begin{bmatrix}\alpha_p \\ 0 \\ \vdots \\ 0\end{bmatrix} \quad (2.14)$$

Remarque.

Ce système est évidemment redondant puisqu'il comporte $p+1$ équations en p inconnues : il ne sera résolu sous cette forme qui par contre convient parfaitement pour l'élaboration des algorithmes de résolution.

-Estimation du gain du modèle.

Après avoir évalué les coefficients *a(i)*, on veut choisir une valeur adéquate pour le gain du modèle :

$\sigma_{e,\min}=\sqrt{\alpha_p}$ donné par (2.13).

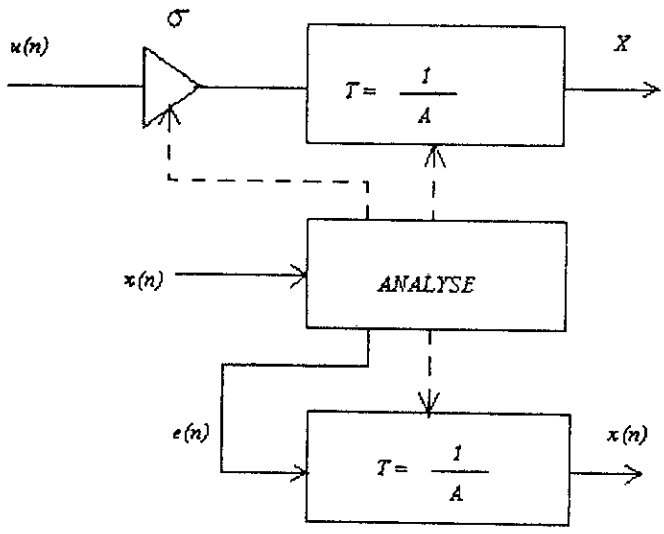


Figure 2.1 Modélisation AR d'ordre p

2. Filtre inverse.

On désigne *X(z) et E(z)* les transformées respectives de *x(n) et e(n)*
La relation (2.1) conduit à :

$$X(z)=\frac{1}{1+a_p(1)z^{-1}+a_p(2)z^{-2}+...+a_p(p)z^{-p}}E(z)=A(z)E(z)$$

A(z) est la transmittance d'un filtre FIR qui, excité par *x(n)* reproduit le signal *e(n)*: ce filtre est appelé filtre inverse.
Le filtre inverse permet de reconstruire le résidu *e(n)* ou une approximation de l'excitation du modèle comme illustré à la figure

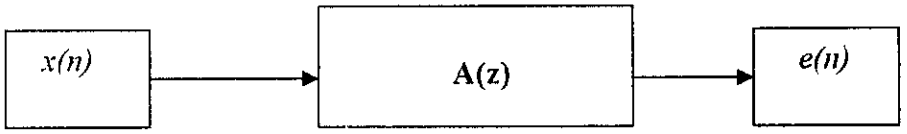


Figure 2.2 Filtre inverse

3. Propriétés de la fonction d'autocorrélation.

La fonction d'autocorrélation du signal engendré par le modèle AR d'ordre p et de gain $\sqrt{\alpha_p}$ excité par un bruit blanc de variance unitaire coincide avec celle du signal original pour $k=0,1,...,p$ au moins.

On peut en effet écrire pour le signal x_{AR} engendré par le modèle AR :

$$x_{AR}(n)=-\sum_{i=1}^{p} a_p(i)x_{AR}(n-i)+\sigma u(n) \qquad (2.15)$$

Avec $\mu_u=0,\sigma_u^2=1$; on a donc :

$$E[x_{AR}(n)x_{AR}(n+k)]=-\sum_{i=1}^{p} a_p(i)E[x_{AR}(n-i)x_{AR}(n+k)]+\sigma E[u(n)x_{AR}(n+k)] \quad (2.16)$$

Le second terme du second membre est nul pour $k=0,1,...,p$ car $u(n)$ est un bruit blanc, on a donc :

$$\phi_{xAR}(k)=-\sum_{i=1}^{p} a_p(i)\phi_{xAR}(k+i)=-\sum_{i=1}^{p} a_p(i)\phi_{xAR}(k-i) \quad (2.17)$$

La dernière égalité résulte de la symétrie de la fonction d'autocorrélation .

Selon (2.12), cette meme relation existe pour $\phi_x(k)$; de plus $\phi_{xAR}(0)=\phi_x(0)$ car le choix du gain du modèle $\sigma=\sigma_{e,\min} \Rightarrow \sigma_{xAR}^2=\sigma_x^2$. (fig 2.1)

On a donc bien l'identité

$$\phi_{xAR}(k)=\phi_x(k) \quad \text{pour} \quad k=0,1,...,p \qquad (2.18)$$

Remarque importante

La propriété (2.17) de la modélisation AR est très importante, en effet, si l'on choisit l'ordre p de la prédiction suffisamment élevé, la fonction d'autocorrélation du signal $x(n)$ et celle de son modèle coincident d'après (2.17) pour un grand nombre de valeurs du délai k . Ceci assure une bonne identification du spectre du modèle avec celui du signal.

4. Algorithme de résolution-Algorithme de Levinson-Durbin.

Rappelons que la fonction d'autocorrelation est supposée connue et que pour un signal stationnaire on a :

$$\phi_{xx}(i,j)=\phi_{xx}(i-j)=\phi_{xx}(k)$$

Initialisation.

$$a_m(0)=1, \quad (m=1,2,...,p) \quad \alpha_0=\phi_{xx}(0,0)=\sigma_x^2$$

Récursion

-Pour $m=1,2,...,p$

$$k_m=-\frac{1}{\alpha_m}\sum_{i=0}^{m-1} a_{m-1}(i)\phi_{xx}(m-i)$$

-Pour $i=1,2,...,m-1$

$$a_m(i)=a_{m-1}(i)+k_m a_{m-1}(m-i)$$

$$a_m(m)=k_m$$

$$\alpha_m=\alpha_{m-1}\left(1-k_m^2\right)$$

Commentaire.

La méthode décrite offre un grand nombre d'avantages :la qualité de calcul n'est pas trop grande et surtout on peut garantir la stabilité du modèle AR et celle de l'algorithme de résolution.

5. Relation avec les coéfficients de prédiction linéaire .

$\tilde{C}(n)$ peut être estimé à partir des coefficients de prédiction $a_p(i)$.

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

on dérive par rapport à z^{-1} on aura.

$$\frac{A_p'}{A_p}=\sum_{n=1}^{\infty}nz^{-(n+1)}c(n)$$

Or $A_p(z)=\sum_{i=0}^{p}a_p(i)z^{-i}$ et $A_p'(z)=\sum_{i=1}^{p}ia_p(i)z^{-i+1}$

$$-\sum_{i=1}^{p}ia_p(i)z^{-i+1}=\left[\sum_{j=0}^{p}a_p(j)z^{-j}\right]\left[\sum_{n=1}^{\infty}nc(n)z^{-n+1}\right]$$

Soit : $-ia_p(i)=\sum_{n=1}^{i-1}nc(n)a_p(i-n)+ic(i)$

on obtient donc la récurrence suivante qui permet le calcul aisé du cepstre :

$$\boxed{c(i)=-a_p(i)-\sum_{n=1}^{i-1}\left(1-\frac{n}{i}\right)a_p(n)c(i-n)} \qquad i>0$$

avec $c(0)=\ln\sigma^2$

$c(0)$ contient l'information sur l'énergie du signal

6. Estimation de la trajectoire des formants.

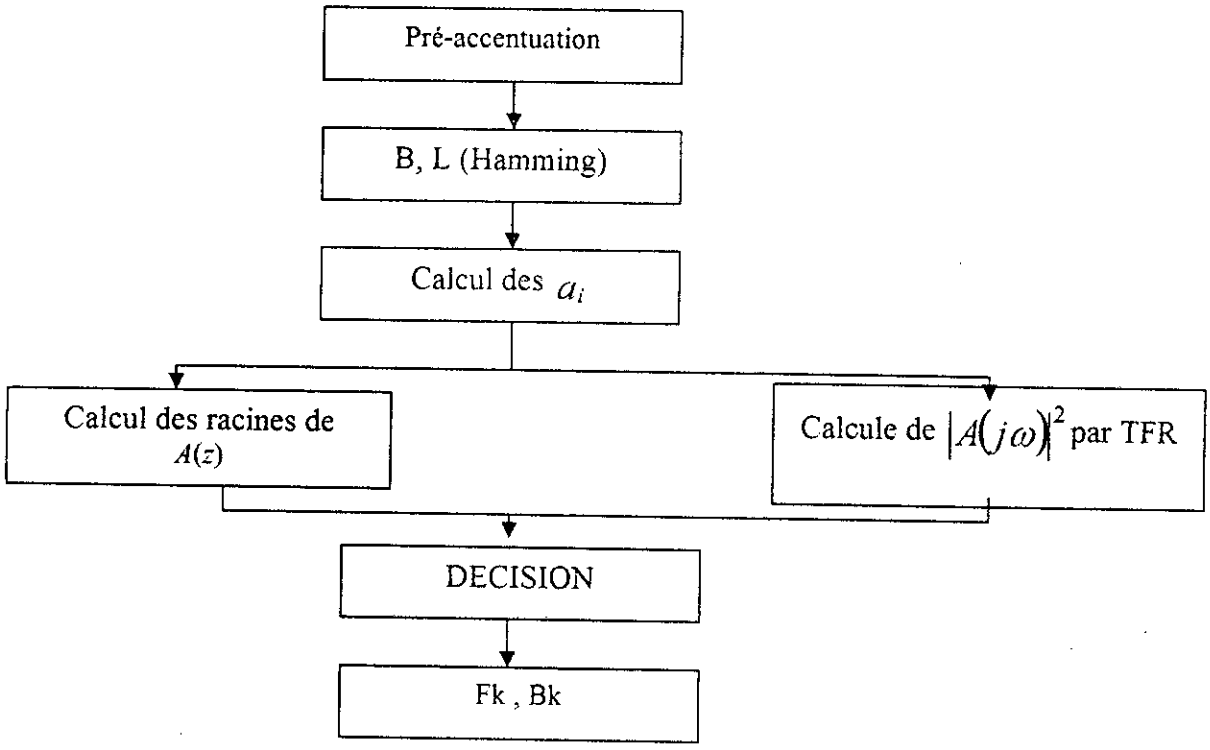
Les fréquences propres de ces formants sont F_k, les amortissements correspondants sont définis par les bandes passantes relatives à –3dB et notées B_k.

Les trois premiers formants sont indispensables pour caractériser le spectre vocal.

On estime ces paramètres par LPC. On recherche les maximas du spectre du modèle, chaque maximum correspond à une paire de pôles complexes de la transmittance $\frac{\sigma}{A_p(z)}$.

Le degré p du polynôme $A_p(z)$ doit être adapté au nombre de formant.

L'organigramme de l'estimation de la trajectoire des formants :



Les maximums du spectre peuvent etre déterminés par la factorisation du polynome $A_p(z)$.

Il a une racine $z_k=\mu_k \exp(j\theta_k)$ proche du cercle unité correspond un certain formant, une première estimation des paramètres sera :

$$F_k=\frac{\theta_k}{2\pi f_e}, B_k=(\Delta_{f,3dB})=\frac{1}{\pi}(1-\mu_k)f_e$$

7. Méthodes de détection de la fréquence fondamentale.

1. Méthode basée sur la fonction d'autocorrélation.

On calcule la fonction d'autocorrélation sur une tranche de N échantillons qui doit recouvrir plusieurs périodes du fondamental :

$$r(k)=\sum_{n=1}^{N-k-1} x(n)x(n+k);k=0,1,...,K$$

Pour l'échantillon contenant la fréquence fondamentale, la fonction d'autocorrélation est maximale. Cette méthode est utilisée dans le domaine temporel.

Inconvénient.

Il apparaît clairement que le coût du calcul est très important ; la fonction d'autocorrélation contient une information surabondante sur le signal.

2. Méthode basée sur la fonction d'autocorrélation en utilisant le noyau .

La solution au problème précédent consiste à ne retenir de chaque échantillon *x(n)* que la partie qui excède un certain seuil; la valeur de ce dernier est définie par rapport à la plus grande valeur en module de *x(n)* .

3. Méthode du cepstre.

L'estimation du pitch peut être faite sur la cepstre réel. le cepstre *c(n)* est observé à travers la fenêtre temporelle. Les maximas sont recherchés dans un intervalle de 2 à 15 ms, un maximum est accepter lorsqu'il excède la courbe du seuil ce qui permet en principe d'éviter le phénomène de doublement du pitch.

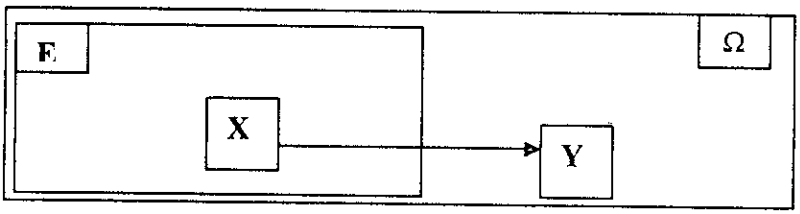
Annexe B : Modèle de Markov caché

1. Structure générale du modèle.

Un modèle MMC se caractérise par un système à états (ou une source) comportant 2 processus.

Le premier est un processus de changement d'états que nous appellerons processus caché X_t et qui n'est pas observable.

Le second est un processus d'émission observable Y_t, que nous appellerons processus externe.



Constituant d'un MMC

La chaîne interne est supposée, pour chaque instant, être dans un état où la fonction correspondante génère une composante de l'observation. La chaîne interne change d'état en suivant une loi de transition. L'observateur ne peut voir que les sorties des fonctions aléatoires associées aux états et ne peut pas observer les états de la chaîne sous-jacente, d'où le terme de Modèles de Markov Cachés (ou Hidden Markov Model).

Les réalisations du premier processus sont des chaînes cachées $X=x_0x_1\cdots x_T x_{T+1}$ ou x_0 est l'état initial de la source avant le début des émissions, noté 0 ou I.

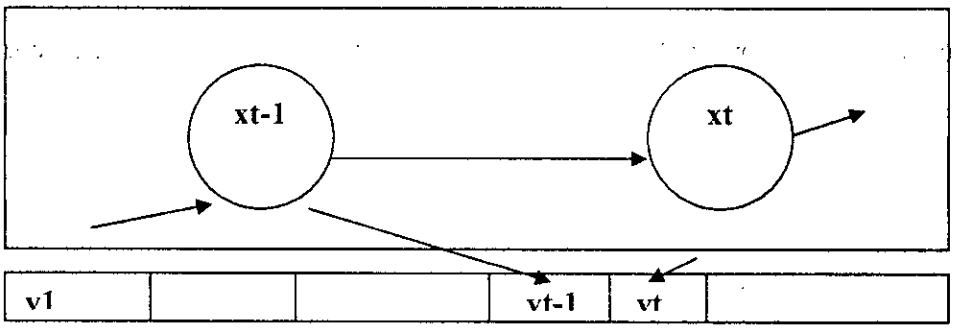
$x_t(1\le t\le T)$ est élément d'un ensemble de N états que nous conviendrons de noter par des entiers et x_{T+1} est l'état final, noté N+1 ou F de sorte que $E=\{0,1,...,N,N+1\}$ est l'ensemble de tous les états.

Le processus (Xt) 0<=t<=T est une chaîne de Markov d'ordre 1, il doit vérifier :

$$P(Xt+1 = qj / Xt = qi, \ldots, X0 = q0) = P(Xt+1 = qj / Xt = qi)$$

$$= aij \text{ pour tout } t >= 0. \ldots (3)$$

Les réalisations du second processus sont des chaînes externes ou observable $Y=y_1 y_2 \cdots y_T$ ou chaque y_t est élément d'un espace d'observation Ω. Celui-ci peut être discret, par exemple s'il s'agit d'un modèle linguistique ou les éléments émis sont des mots, ou continu, par exemple si l'on représente l'émission de la parole par une suite de spectres à valeurs dans un continuum. L'élément y_t est émis quand la chaîne interne atteint x_t.



Transition dans un modèle MMC

Ce processus (Yt) 0<=t<=T, vérifie :

P(Yt = yt / Xt = qi,...,X1=q1, Yt-1=yt-1,...,Y1=y1) = P(Yt=yt / Xt = qi)

= bi(yt) **(2)**

Les observations sont supposées indépendantes les unes des autres conditionnellement à la suite d'états. Chaque réalisation de Yt ne dépend que de l'état courant caché. Les observations yt peuvent être de nature :

- discrète => bi est une distribution de probabilité discrète : une loi discrète est généralement représentée par les fréquences d'apparitions des observations discrètes.
- continue => bi est une fonction de densité de probabilité définie sur R^d : les densités traditionnelles utilisées sont des densités gaussiennes, entièrement définies par le vecteur moyenne et la matrice de covariance, ou des densités de type multi-gaussiennes (sommes pondérées de densités gaussiennes).

Il s'ensuit qu'un modèle de Markov caché est caractérisé par :

- son ensemble fini d'états Q = (q1, ..., qN)
- son ensemble de probabilités de transitions entre les états A=(aij)1<=i<=N 1<=j<=N
- son ensemble de lois (ou densités) de probabilités associées à un état B = (bi (...)) 1<= i <=N
- son ensemble de probabilités initiales $\Pi=(\pi_i)$ 1<= i <=N , π_i désigne la probabilité d'entrer dans le modèle par l'état initial qi. Cette probabilité est généralement égale à (1 / nombre d'états initiaux). Un modèle de Markov caché est décrit par le jeu de paramètres Θ = (Π, A, B).

2. La fonction de vraisemblance (ou likelihood).

La vraisemblance d'une suite d'observations par rapport à un tel modèle, est calculée comme ci-dessous [1] :

Soient :

Y= y1,.....,yT la suite d'observations

Qi = qi1,....., qiT la suite d'états de longueur T le long du chemin i

$$p(Y_1=y_1,\dots,Y_T=y_T/\Theta) = \sum Q_i p(X_1=q_{i1},\dots,X_T=q_{iT},Y_1=y_1,\dots,Y_T=y_T /\Theta)$$

$$= \sum Q_i p(Y_1=y_1,\dots,Y_T=y_T /X_1=q_{i1},\dots,X_{iT}=q_{iT} /\Theta) p(X_1=q_{i1},\dots,X_T=q_{iT}/\Theta)$$

En réutilisant les formules (1) et (2) et après réarrangement, on obtient :

$$\boxed{p(Y_1=y_1,\dots,Y_T=y_T/\Theta) = \sum Q_i \left[\pi_{i1} b_{i1}(y_1) \left[\prod_{n=2}^{T} a_{in-1in} b_{in}(y_n) \right] \right]}$$

Remarque.

Trés souvent on utilise un modèle simplifié de MMC ou l'émission à l'instant t est supposée ne dépendre que de l'état x_t au temps t, et non du couple (x_t, x_{t-1}). L'intérêt est pratique : on a ainsi des matrices **b** de dimension plus réduite .

3. Modèle d'observation.

Le modèle d'observation est défini par les probabilités conditionnelles suivantes :

$$p(Y_{1:T} / X_{1:T}=i_{1:T})=\prod_{t=1}^{T} p(Y_t / X_t=i_t)=\prod_{t=1}^{T} f_{it}(X_t)$$

Paramètres des N densités conditionnelles $f_i(X)$ (cas gaussien : $\mu_i et \sum_i$)

Annexe C : Les tableaux

Résultat

La quantification vectorielle

Fréquence d'échantillonnage 16 kHz

1, 9, 10 : Phrases à tester

16 Locuteurs			
Nbre de classes = 4			
Locuteur	1	9	10
1	0	0	1
2	1	1	1
3	1	1	1
4	1	0	1
5	0	0	1
6	1	1	1
7	1	1	1
8	1	1	1
9	0	1	1
10	1	1	1
11	1	1	1
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	0	1
Pourcentage = 85,41 %			

16 Locuteurs			
Nbre de classes = 8			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	0	0	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	0	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	0	1
Pourcentage = 91,67 %			

16 Locuteurs			
Nbre de classes = 16			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	0	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	0	1
Pourcentage = 95,83 %			

16 Locuteurs			
Nbre de classes = 32			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	0	1
Pourcentage = 97,91%			

32 Locuteurs			
Nbre de classes = 4			
Locuteur	1	9	10
1	0	0	1
2	1	1	1
3	1	1	1
4	1	0	1
5	0	0	0
6	1	1	1
7	1	1	1
8	1	1	1
9	0	1	1
10	1	1	1
11	1	1	1
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	0	1
17	0	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	0	1	1
24	1	1	0
25	1	1	1
26	1	1	1
27	1	1	0
28	1	1	1
29	1	1	1
30	1	1	1
31	1	1	1
32	1	1	1
Pourcentage = 87,5 %			

32 Locuteurs			
Nbre de classes = 8			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	0	0	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	0	1
13	1	1	1
14	1	1	0
15	1	1	1
16	1	0	1
17	1	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	1	1	1
24	1	1	1
25	1	1	1
26	1	1	1
27	1	1	0
28	1	1	1
29	1	1	1
30	1	1	1
31	1	1	1
32	1	1	1
Pourcentage = 93,75 %			

32 Locuteurs			
Nbre de classes = 16			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	0
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	0	1
13	1	1	1
14	1	1	0
15	1	1	0
16	1	0	1
17	1	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	1	1	1
24	0	1	1
25	1	1	1
26	1	1	1
27	1	1	0
28	1	1	1
29	1	1	1
30	1	1	1
31	1	1	1
32	1	1	1
Pourcentage = 92,71 %			

32 Locuteurs			
Nbre de classes = 32			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	0	1
17	1	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	0	1	1
24	1	1	1
25	1	1	1
26	1	1	1
27	1	1	1
28	1	1	1
29	1	1	1
30	1	1	1
31	1	1	1
32	1	1	1
Pourcentage = 91,92 %			

48 Locuteurs			
Nbre de classes = 4			
Locuteur	1	9	10
1	0	0	1
2	1	1	1
3	1	1	1
4	1	0	0
5	0	0	0
6	1	1	1
7	1	1	1
8	1	1	1
9	0	1	1
10	1	1	0
11	1	1	1
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	0	1
17	0	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	0	1	1
24	1	1	0
25	0	1	0
26	1	1	1
27	1	1	0
28	1	1	1
29	1	1	0
30	1	1	1
31	1	1	1
32	1	1	1
33	1	1	1
34	1	1	0
35	1	1	1
36	1	1	0
37	0	1	1
38	1	0	0
39	1	1	1
40	1	1	1
41	1	1	0
42	1	1	1
43	1	0	1
44	1	1	1
45	1	1	0
46	1	1	1
47	1	0	0
48	1	1	1
Pourcentage = 81,25 %			

48 Locuteurs			
Nbre de classes= 8			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	0	0	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	0	1
13	1	1	1
14	1	1	0
15	1	1	1
16	1	0	1
17	1	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	1	1	1
24	1	1	1
25	1	0	1
26	1	1	1
27	1	1	0
28	1	1	1
29	1	1	0
30	1	1	1
31	1	1	1
32	1	1	1
33	1	1	1
34	1	1	0
35	1	1	1
36	1	1	1
37	1	0	1
38	1	0	0
39	1	1	1
40	1	1	1
41	1	1	1
42	1	1	0
43	1	0	0
44	1	1	1
45	1	1	1
46	1	1	1
47	0	1	1
48	1	0	1
Pourcentage = 88,19 %			

48 Locuteurs			
Nbre de classes = 16			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	0
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	0	1
13	1	1	1
14	1	1	0
15	1	1	0
16	1	0	1
17	1	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	0	1	1
24	1	1	1
25	1	1	1
26	1	1	1
27	1	1	0
28	1	1	1
29	1	1	1
30	1	1	1
31	1	1	1
32	1	1	1
33	1	1	1
34	1	1	0
35	1	1	1
36	1	1	1
37	1	1	1
38	1	1	1
39	1	1	1
40	1	1	1
41	1	1	1
42	1	1	1
43	1	0	1
44	1	1	1
45	1	1	1
46	1	1	1
47	1	1	1
48	1	1	1
Pourcentage = 93,75 %			

48 Locuteurs			
Nbre de classes = 32			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	1	1	1
7	1	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	1	1	1
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	0	1
17	1	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	0	1	1
24	1	1	1
25	1	1	1
26	1	1	1
27	1	1	1
28	1	1	1
29	1	1	1
30	1	1	1
31	1	1	1
32	1	1	1
33	1	1	1
34	1	1	1
35	1	1	1
36	1	1	1
37	1	1	1
38	1	1	1
39	1	1	1
40	1	1	1
41	1	1	1
42	1	0	0
43	1	1	0
44	1	1	1
45	1	1	1
46	1	1	1
47	1	1	1
48	1	0	1
Pourcentage = 95,83 %			

1.2 Fréquence d'échantillonnage 8 kHz

16 Locuteurs			
Nbre de classe = 4			
Locuteur	1	9	10
1	1	1	1
2	1	1	0
3	0	0	1
4	1	0	0
5	1	1	1
6	0	0	0
7	0	0	0
8	1	0	0
9	0	0	0
10	1	0	0
11	0	0	0
12	0	0	0
13	0	1	1
14	1	0	0
15	0	0	0
16	0	0	0
Pourcentage = 31,25 %			

16 Locuteurs			
Nbre de classe = 8			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	0	0
5	1	1	1
6	1	1	0
7	0	1	1
8	1	1	1
9	0	1	1
10	1	1	1
11	0	0	0
12	0	0	0
13	0	1	1
14	1	1	1
15	1	1	0
16	1	1	1
Pourcentage = 72,92 %			

16 Locuteurs			
Nbre de classe = 16			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	0	0
5	1	1	1
6	1	1	1
7	0	1	1
8	1	1	1
9	0	1	1
10	0	0	0
11	0	0	0
12	0	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
Pourcentage = 77,08 %			

16 Locuteurs			
Nbre de classe = 32			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	0	1	1
4	1	0	0
5	1	1	1
6	1	1	1
7	0	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	0	1	0
12	0	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
Pourcentage = 85,42 %			

16 Locuteurs			
Nbre de classe = 50			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	0	1	1
4	1	0	0
5	1	1	1
6	1	1	1
7	0	1	1
8	1	1	1
9	1	1	1
10	1	1	1
11	0	1	0
12	0	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
Pourcentage = 85,42 %			

32 Locuteurs			
Nbre de classes = 4			
Locuteur	1	9	10
1	1	1	1
2	1	1	0
3	0	0	1
4	1	0	0
5	1	1	1
6	0	0	0
7	0	0	0
8	1	0	0
9	0	0	0
10	1	0	0
11	0	0	0
12	0	0	0
13	0	1	1
14	0	0	0
15	0	0	0
16	0	0	0
17	0	1	1
18	0	1	0
19	0	0	1
20	1	1	1
21	0	0	0
22	0	1	0
23	0	0	0
24	0	0	0
25	0	0	0
26	0	1	0
27	0	0	0
28	0	1	1
29	0	0	0

32 Locuteurs			
Nbre de classes = 8			
Locuteur	1	9	10
1	0	1	1
2	1	1	1
3	1	1	1
4	1	0	0
5	1	1	1
6	1	1	0
7	0	0	0
8	1	1	1
9	0	1	1
10	0	0	0
11	0	0	0
12	0	0	0
13	0	1	1
14	1	0	0
15	1	1	0
16	1	1	1
17	0	0	0
18	1	1	1
19	1	1	1
20	1	1	1
21	1	0	1
22	1	1	1
23	1	0	0
24	0	0	0
25	0	1	1
26	0	1	1
27	1	0	0
28	0	0	0
29	0	0	0

30	0	0	0
31	1	1	1
32	1	0	0
Pourcentage = 30,21 %			

30	0	0	0
31	1	1	1
32	1	0	1
Pourcentage = 54,17 %			

32 Locuteurs			
Nbre de classes = 16			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	1	0	0
5	1	1	1
6	1	1	1
7	0	1	1
8	1	1	1
9	0	1	1
10	0	0	0
11	0	0	0
12	0	1	1
13	1	1	1
14	1	0	0
15	1	1	1
16	1	1	1
17	0	1	1
18	0	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	0	1	0
24	1	1	1
25	0	1	0
26	0	1	1
27	1	0	0
28	0	1	1
29	1	0	0
30	0	0	1
31	1	1	1
32	1	1	1
Pourcentage =71,875 %			

32 Locuteurs			
Nbre de classes = 32			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	0	1	1
4	1	0	0
5	1	1	1
6	1	1	1
7	0	1	1
8	1	1	1
9	0	1	1
10	1	1	1
11	0	0	0
12	0	0	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
17	1	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	1	1	0
24	1	1	1
25	0	1	0
26	1	1	1
27	1	0	0
28	0	1	1
29	0	0	0
30	1	1	1
31	1	1	1
32	1	1	1
Pourcentage = 80,21 %			

32 Locuteurs			
Nbre de classe = 50			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	0	1	1
4	1	0	0
5	1	1	1
6	1	1	1
7	0	1	1
8	1	1	1
9	0	1	1
10	1	1	1
11	0	0	0
12	0	0	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
17	1	1	1
18	1	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	1	1	1
23	1	1	0
24	1	1	1
25	0	1	0
26	1	1	1
27	1	0	0
28	0	1	1
29	0	0	0
30	1	1	1
31	1	1	1
32	1	1	1
Pourcentage = 80,21 %			

48 Locuteurs			
Nbre de classe = 8			
Locuteur	1	9	10
1	0	1	1
2	1	1	1
3	1	1	1
4	0	0	0
5	0	0	1
6	1	1	0
7	0	0	0
8	0	1	1
9	0	1	1
10	0	0	0
11	0	0	0
12	0	0	0
13	0	1	1
14	1	0	0
15	1	1	0
16	1	1	1
17	0	0	0
18	1	1	1
19	1	1	1
20	1	1	1
21	1	0	1
22	0	1	0
23	1	0	0
24	0	0	0
25	0	1	0
26	0	1	1
27	1	0	0
28	0	0	0
29	0	0	0
30	0	0	0
31	1	1	1
32	1	0	1
33	1	1	0
34	1	0	0
35	1	1	0
36	1	0	0
37	1	1	1
38	1	0	0
39	1	1	1
40	1	1	1
41	1	1	1
42	0	0	0
43	1	0	0
44	0	0	0
45	0	0	0
46	0	0	0
47	0	0	0
48	0	0	0
Pourcentage = 45,14 %			

48 Locuteurs			
Nbre de classe = 16			
Locuteur	1	9	10
1	1	1	1
2	1	1	1
3	1	1	1
4	0	0	0
5	1	1	1
6	1	1	1
7	0	1	1
8	1	1	1
9	0	1	1
10	0	0	0
11	0	0	0
12	0	1	1
13	1	1	1
14	1	0	0
15	1	1	1
16	1	1	1
17	0	1	1
18	0	1	1
19	1	1	1
20	1	1	1
21	1	1	1
22	0	1	0
23	0	0	0
24	1	1	1
25	0	1	0
26	0	1	1
27	0	0	0
28	0	1	1
29	1	0	0
30	0	0	1
31	1	1	1
32	1	1	1
33	1	1	1
34	1	0	0
35	1	1	0
36	1	0	1
37	1	1	1
38	1	1	1
39	1	1	1
40	1	1	1
41	1	1	1
42	0	0	0
43	1	0	0
44	0	0	0
45	0	0	0
46	1	1	1
47	1	0	0
48	0	1	1
Pourcentage = 65,28 %			

BIBLIOGRAPHIE

[1] Reconnaissance automatique de la parole. Auteurs : J-P.Haton, J-M. Pierrel, G. Perennou, J. Caëlen, J-L. Gauvin. Ed: dunod informatique. Année 1991.

[2] Fondements théoriques de la radiotechnique statistique T2. Auteur : B.Lévine. Ed :Mir Année : 1973.

[3] Traitement automatique de la parole. Auteurs : Boite René, Hervé Bourlard, Thierry Dutoit, Joel Hancq et Henri Leich. Ed : presses polytechniques et universitaires romandes. Année 2000.

[4] Thèse de docteur-ingénieur : Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique. Auteur : Yves Grenier -Département Système et Communications-ENST. Année : 1977.

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

[6] Thèse de docteur-ingénieur : Etude pour un vocodeur à classification par prédiction linéaire en vue de la transmission de la parole à très faible débit. Auteur : Antonio DE LA O-Département Système et Communications-ENST. Année :1982.

[7] Thèse d'ingéniorat à l'ENP : Reconnaissance de la parole par la méthode des chaines de Markov cachées. Auteur : T.Berbar. Année :1991.

[8] Thèse de magister à l'ENP: Detection de la fatigue vocale à l'aide de paramètres prosodiques et formantiques. Auteur : N. Abina. Année :1995.

[8] International Phonetic Association (IPA) homepage. http://www.arts.gla.ac.uk/IPA/ipa.html.

[9] ATAL, B. S., Efficient coding of LPC parameters by temporal decomposition. Proc. IEEE ICASSP 83, pp. 81-84, 1983.

[10] GAROFOLO, J.S., L.F.LAMEL, W.M. FISHER, J.G. FISCUS, D.S. PALLETT N.L. DAHLGREN, DARPA-TIMIT acoustic-phonetic speech corpus. NISTIR 4930, U.S. Department of Commerce, National Institute of Standards and Technology, Computer Systems Laboratory, 1993.

[11] GERSHO, A., Advances in speech and audio compression. Proc. IEEE, 82(6), pp. 900-918, 1994.

[12] ISMAIL, M. K. PONTING, Between recognition and synthesis - 300 bits/second speech

coding. Proc. EUROSPEECH 97, pp. 441-444, Rhodes, Greece, 1997.

[13] RABINER, L. R. L. W. SCHAEFFER, Digital processing of speech signals. Prentice Hall, 1978.

[14] RIBEIRO, C.M. I.M. TRANCOSO, Application of speaker modification techniques to phonetic vocoding. Proc. ICSLP 96, pp. 306-309, Philadelphia, 1996.

[15] RIBEIRO, C.M. I.M. TRANCOSO, Phonetic vocoding with speaker adaptation. Proc. EUROSPEECH 97, pp. 1291-1294, Rhodes, Greece, 1997.

[16] Gurmeet Singh, Ashsh Panda, Saurav Bhattacharyya, Thambipillai Srikanthan "vector quantization techniques for GMM based speaker verification" *IEEE 2003.*

[17] D.A Reynolds and R.C Rose, « Robust text-independent Speaker Identification Using Gaussian Mixture Speaker Models », *IEEE Trans. Speech and Audio Processing, Vol. 3, no. 1, 1995,pp. 72-82.*

[18] Alex S. Park, « ASR Dependent Techniques for Speaker Recognition MASSACHUSETTS INSTITUTE OF TECHNOLOGY, May 2002.

[19] Douglas A. Reynolds " An Overview of Automatic Speaker Recognition Technology" *IEEE 2002.*

[20] JOSEPH P. CAMBELL, JR " Speaker Recognition: A Tutorial" *proceeding of the IEEE, VOL 85, NO 9, september1997.*

[21] Sadaoki Furuui " RECENT advances in speaker recognition" *Tokyo Institue of Technology 2-12-1, O-okayama-ku, Tokyo 152, Japan (Pattern recognition Letters 18 (1997)859-872).*

[22] Douglas A. Reynolds, Thomas F. Quateieri and Robert B. Dunn."Speaker Verification Using Adapted Gaussian Mixture modes" *M.I.T Lincoln Laboratory, 244 wood St., Lexington, Massachusetts 02420.*

[23] Vincent Wan "speaker Verification Using Support vector Machines", *Departement of Computer Science. University of Sheffield United Kingdom, Jun 2003.*

[24] Patrick StrÖm, Tobias Ljungkvist, Wojtek Dabrowski, Tomas Andersson, P°al Brevik, Tony Lindblom "SPEAKER RECOGNITION FOR USER IDENTIFICATION AND VERIFICATION" , *THE ROYAL INSTITUTE OF TECHNOLOGY STOCKHOLM, SWEDEN MAY2001.*