République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche
Ecole Nationale Polytechnique

المدرسة الوطنية المتعددة التقنيات
Ecole Nationale Polytechnique

Département d'Electronique

Projet de fin d'étude

En vue de l'obtention du Diplôme d'Ingénieur d'Etat en Electronique

Thème

Différentes méthodes de traitement de signal appliquées à l'identification des locuteurs indépendants du texte.

Etudié par :
- BOUFFOUGROUNE Zakaria

Proposé et dirigé par :
- Mr: *B. BOUSSEKSOU*

Juin 2014
ENP 10 avenue Hassan Badi El-Harrach Alger

Remerciements

Je tiens à remercier en premier lieu « ALLAH » le tout puissant, qui m'a donné la force, le courage et la volonté pour mener à bien ce modeste travail.

J'exprime ma profonde gratitude, mon grand respect et ma sincère reconnaissance à mon promoteur monsieur ***B. BOUSSEKSOU*** *pour avoir assumé la lourde responsabilité de m'encadrer, de m'avoir orienté et conseiller tout au long de ce travail ainsi pour la confiance qu'il m'a accordée.*

Je remercie profondément ***Mlle Moussaoui*** de nous avoir fait l'honneur de présider le jury.

Mes remerciements vont aussi à **Mr Mameri** pour avoir accepté de juger notre travail.

Je remercie vivement tous mes enseignants de l'Ecole Nationale polytechnique.

Dédicaces

Je dédie ce modeste travail :

A mes très chers parents et ma grande famille.

A mes sœurs.

A tous mes amis et spécialement Nounou, Lamaidi, Salah, Hamdan, Osmane, Mohamed Anis, allawa, sans oublier Gamiz.

A toute la famille et spécialement à mes oncles Morad, Kamilou et hacéne. Mes tantes Warda, Soria et hassiba. Et sans oublier ma chère cousine Sarah et ami hmed.

A tous ceux qui m'aiment et que j'aime.

A tous mes collègues de la promotion 2014.

A vous.

ملخص

يندرج هذا العمل ضمن التشخيص الأوتوماتيكي للمتكلم، هذا الميدان الغني بالتطبيقات البالغة الأهمية بدءا من تامين المعابر و التطبيقات القضائية إلى تنظيم الملفات الصوتية. حتى نترك المجال مفتوح، اهتممنا بالتشخيص الأوتوماتيكي للمتكلم المستقل عن النص المنطوق. اهتممنا بكيفية تمثيل المتكلمين و استخراج الشارات الصوتية للمتكلمين حتى نتمكن من تطبيقها في نظام تشخيص المتكلم.

كلمات مفتاحية : تشخيص المتكلم، الفضاء الصوتي، وسائط مال سابستر، التكميم الشعاعي، النموذج متعدد الغوصيات.

Résumé

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

Mots clés : identification du locuteur, espace acoustique, MFCC, quantification vectorielle QV, modèles de mélanges des gaussiennes GMM.

Abstract

This work relates to the automatic speaker recognition which has many potential applications ranging from access security to audio indexing. In this thesis, the text-independent speaker identification is studied with a specific focus on speaker modeling and representation. We are especially interested to extract, from speech signals, the relative information of the speaker identity and estimate a sufficiently robust speaker's model.

Keywords: speaker identification, acoustic space, Mel frequency cepstral coefficients MFCC, vector quantization VQ, Gaussian mixture models GMM.

Table des matières

Chapitre 1 : Généralités

Chapitre 2 : Paramétrisation du signal vocal

Chapitre 3 : Modélisation des locuteurs

Table des matières

Chapitre 4 : Evaluations expérimentales

Liste des Figures

Liste des Tableaux

Acronymes

AR : **A**uto **R**égressif.

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

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

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

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

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

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

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

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

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

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

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

QV : **Q**uantification **V**ectorielle.

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

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

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

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

Chapitre 1

Généralités

Introduction

Chaque être humain peut, dès son plus jeune âge, reconnaître les voix des personnes qui lui sont familières. Bien que le processus de reconnaissance de la parole soit fort développé chez l'homme, il ne lui est cependant pas immédiat de caractériser les indices qui permettent de distinguer un locuteur d'un autre. La figure 1.1 illustre ce problème en montrant que le signal parole émis par un être humain transmet son identité en plus du message, ces deux entités étant intimement liées. D'évidence, l'interprétation du message transmis par la parole est fortement dépendante des interlocuteurs impliqués dans le processus de production et de reconnaissance de la parole. Le décodage des émotions ainsi que les variations physiologiques des intervenants peuvent en effet changer non seulement les caractéristiques physiques du signal parole, mais également son sens.

La parole est certainement le moyen de communication directe entre humains qui est le plus sophistiqué. Les subtiles variations du langage sont capables de susciter chez l'auditeur non seulement une palette forte variée d'émotions et de sentiments, mais aussi une attention complète de son cerveau. Les ordinateurs et les logiciels qui se construisent actuellement, bien que capables de traiter énormément d'informations en un temps très court, n'ont pas encore la capacité de générer ou de comprendre les finesses de la parole humaine. Cependant, de nombreuses applications en reconnaissance de la parole sont déjà industrialisées, allant de la dictée vocale à la commande d'opérations diverses dans les navettes spatiales. De plus en plus, les entreprises de télécommunications et de services (banques, assurances), désireuses d'améliorer leur service à la clientèle, tentent d'introduire des applications basées sur les technologies de la parole. La palette de ces technologies est fort riche, partant de systèmes de reconnaissance de la parole entraînés pour un seul locuteur à des systèmes capables de reconnaître des centaines de milliers de mots. Dans un autre registre, un grand nombre de services demandent une reconnaissance de l'identité du locuteur (accès aux boîtes vocales, à des services par abonnements, consultation de comptes en banques, etc. . .). Finalement, pour que le dialogue homme-machine soit complet, le domaine

de la synthèse de la parole essaie de produire de la voix humaine (ou y ressemblant fort) automatiquement.

Cette thèse est plutôt orientée sur la reconnaissance de l'identité d'un locuteur par sa voix. Cependant, comme toutes les technologies vocales font appel à différents aspects du même phénomène (la voix humaine), elles sont, par beaucoup d'aspects, indissociables, ce qui nous induira à traiter également des aspects de reconnaissance de la parole et d'analyse/synthèse de celle-ci.

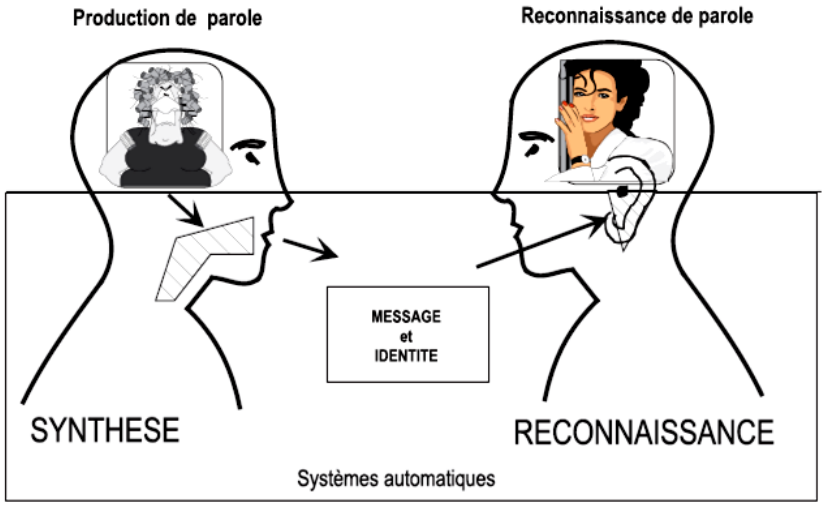


FIG. 1.1 – Production et reconnaissance de la parole.

Par mesure de simplification, nous supposerons que le message émis par un locuteur peut être considéré comme sans ambiguïtés pour un système de reconnaissance automatique de la p*arole*.

1.1 Le signal de la parole

La figure 1.2 [1] nous montre l'appareil phonatoire humain et les éléments qui le définissent. La commande de ces différents éléments physiologiques s'effectue à partir du cerveau lui-même soumis à des influences psychologiques pouvant modifier fortement le signal de parole lui-même (peur, colère, joie, etc. . .) .

Lorsqu'on observe le signal de parole durant le temps (voir la figure 1.3) on peut constater de larges différences d'amplitude et de durée lors de la prononciation d'un même mot, d'un locuteur à l'autre, mais aussi d'une prononciation à l'autre émanant du même locuteur.

La transformation en temps/fréquences (figures 1.4 et 1.5) nous permet cependant de constater l'existence de lignes d'énergie à certaines fréquences. Si l'on utilise une analyse adaptée (spectrogramme large bande de la figure 1.5), on voit même l'apparition des zones fréquentielles de forme similaire. Ce sont ces propriétés du signal que les phonéticiens et la **reconnaissance automatique de la parole** tentent d'exploiter. Comme ce signal fréquentiel transporte simultanément des informations sur l'identité du locuteur, l'analyse de ces mêmes motifs spectraux doit permettre d'effectuer une **reconnaissance du locuteur**.

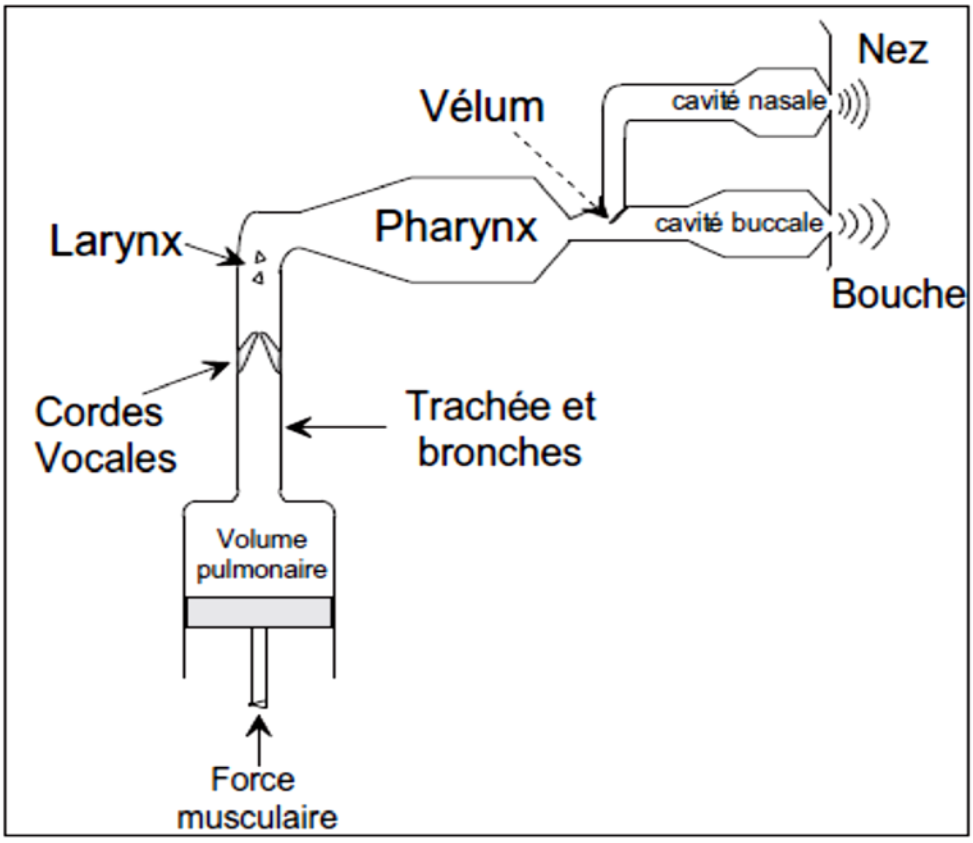


FIG. 1.2 – Les différentes parties constituant le conduit vocal.

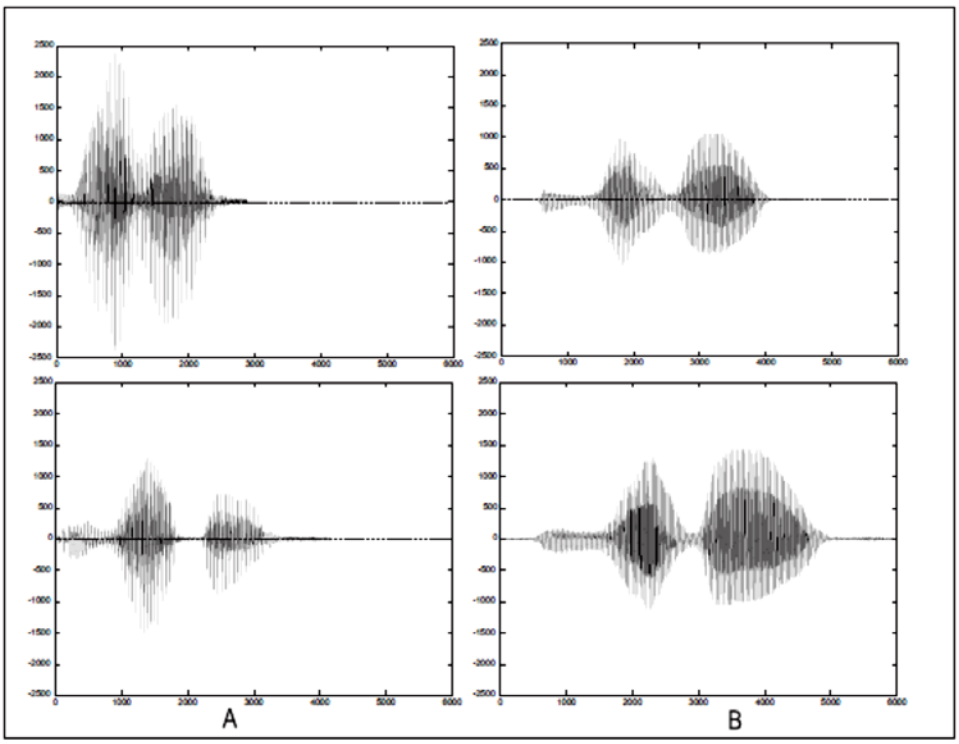


FIG. 1.3 – Exemple de signaux parole prononciations du mot zéro: à gauche deux prononciations de la personne A à droite deux prononciations de la personne B.

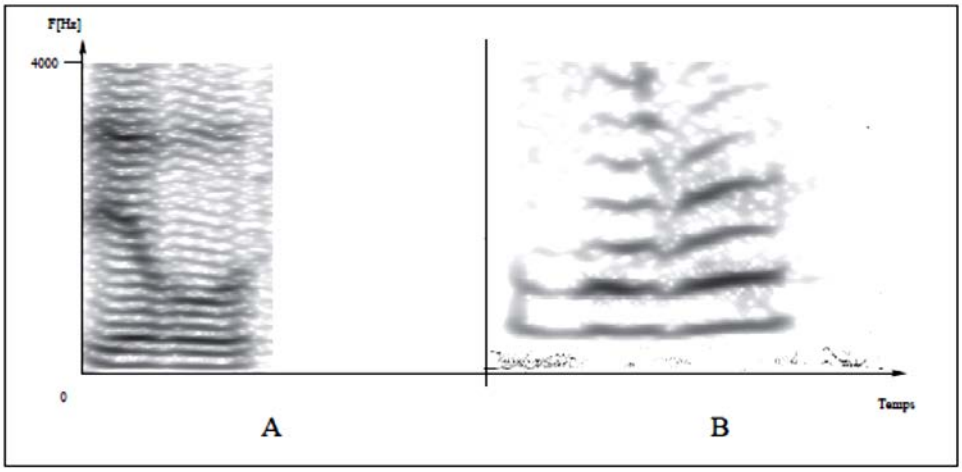


FIG. 1.4 – Spectrogrammes en bande étroite pour le mot “zéro” des locuteurs A et B (extrait de la première prononciation du mot de la figure 1.3).

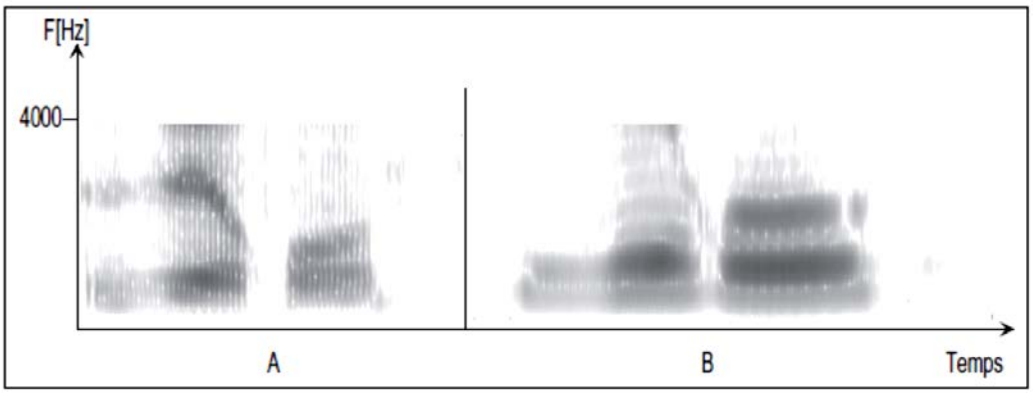


FIG. 1.5 – Spectrogrammes en bande large pour le mot “zéro” des locuteurs A et B (extrait de la première répétition du mot de la figure 1.3).[2]

1.1.1 Motifs spectro-temporels

L’analyse plus précise d’un spectrogramme permet de déterminer un espacement régulier entre les différentes lignes spectrales (voir la figure 1.6), représentant en fait les **harmoniques** d’une **fréquence fondamentale** (F0). Cette fréquence fondamentale est directement liée à la source du signal de parole que sont les cordes vocales. La figure 1.8 (selon [1]) nous donne une idée de l’évolution de la vitesse du volume d’air à la sortie des cordes vocales. Nous le verrons par la suite, cette fonction est souvent modélisée par un train d’impulsion à la fréquence fondamentale (F0).

L’inspection d’un spectrogramme généré à partir d’une analyse temps-fréquence en bande large permet de faire ressortir une accumulation d’énergie des harmoniques dans certaines zones. La figure 1.7 montre le déplacement du centre de ces zones, appelées **formants**. Les parties où la fréquence fondamentale existe sont appelées parties **voisées** et correspondent aux parties du signal où les cordes vocales sont en activité. On distingue aussi des parties du spectrogramme où il semble qu’aucun motif fréquentiel ne se dessine, elles correspondent aux régions où les cordes vocales ne vibrent pas.

Une des caractéristiques majeures du signal de parole est la variation de la valeur de la fréquence fondamentale selon l’état psychologique ou physiologique du locuteur et le sens que celui-ci veut donner à ce qu’il prononce. Cette variabilité pose un problème d’identification de motifs à des fréquences données. La figure 1.9 montre l’évolution de F0 et

des formants tout au long d'une phrase (ici l'extrait de phrase "reconnaissance du locuteur?"). Remarquons les variations de la fréquence vocale importantes dans la dernière partie du mot "locuteur" dues à la forme interrogative.

Cette évolution de la fréquence fondamentale (et des formants) contribue à la **prosodie**, que l'on pourrait aussi expliquer comme la *musique* de la parole.

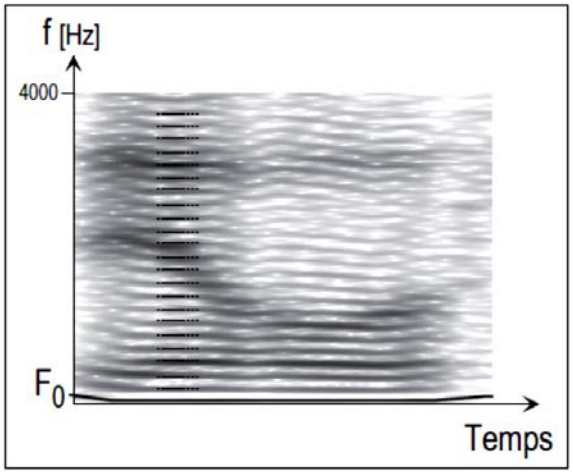


FIG. 1.6 – Fréquence fondamentale et harmoniques du mot "zéro"

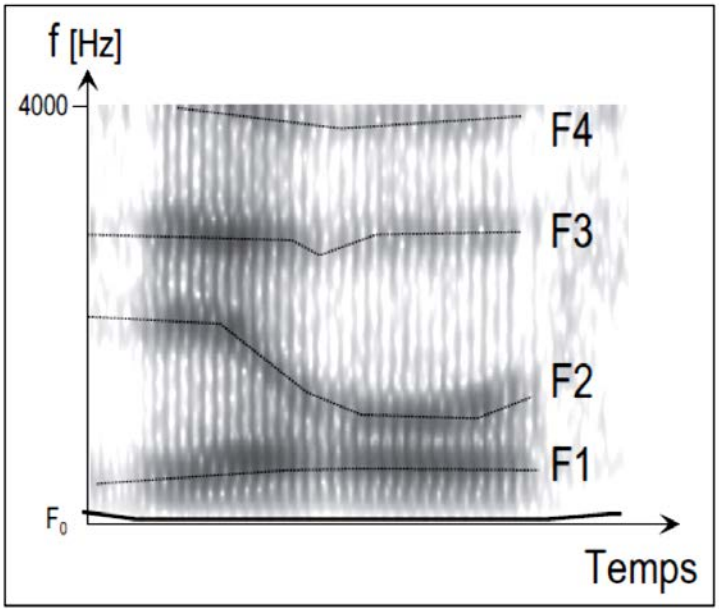


FIG. 1.7 – Fréquence fondamentalenet les 4 premiers formants.

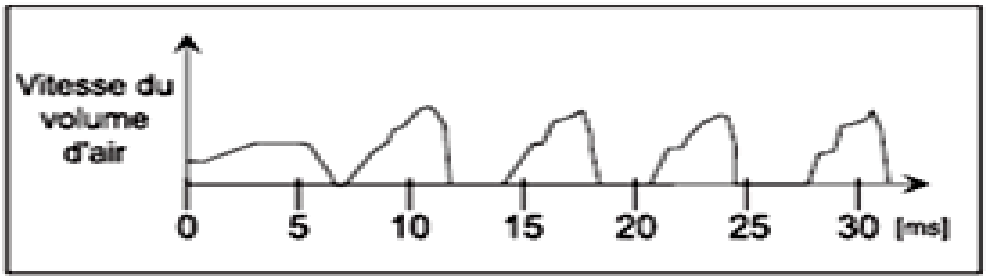


FIG. 1.8–Vitesse du volume d'air a la sortie des cordes vocales

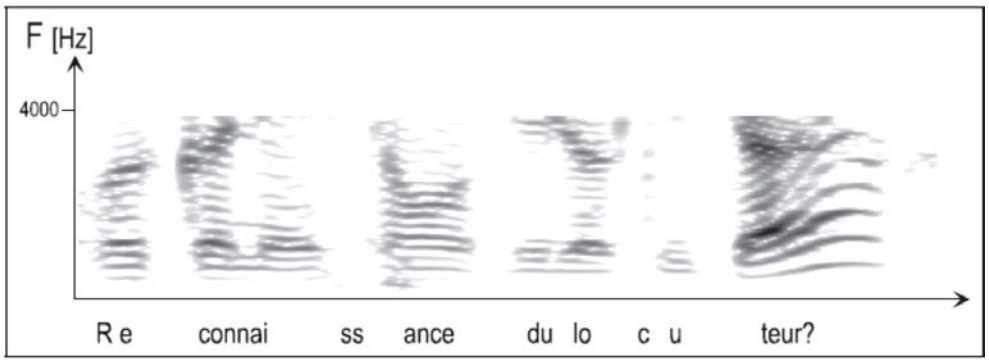


FIG. 1.9 – Variation du spectre dû à la forme interrogative

1.1.2 Unités temporelles

Si l'on analyse la figure 1.9, on peut constater que certains motifs formantiques semblent se répéter. Ces motifs constituent les éléments d'une langue. Selon la durée à laquelle on les associe, on leur donne des noms différents. Les plus petites unités que les phonéticiens définissent sont des unités abstraites dont le corrélat acoustique est le **phonème**. Le nombre exact de phonèmes dépend des écoles, mais on en dénombre environ 35 dans la langue française, que l'on peut diviser en plusieurs types (ici aussi les classifications varient selon les phonéticiens) comme par exemple les voyelles (a, e, i, o, u), les fricatives (s,f), les plosives (p,t,k) ou les liquides (l). On peut définir des unités temporelles plus longues qui correspondent, par exemple, à des segments compris entre les maxima de stabilité spectrale de 2 phones consécutifs (diphones) ou des unités plus longues encore, tels les tri-phones, les poly-phones ou les syllabes.

1.2 Reconnaissance automatique du locuteur

Comme l'a montré la section précédente, la parole est un signal particulier, de par sa variabilité et sa richesse. C'est probablement pour cela que depuis plus de 30 ans, de nombreux chercheurs se sont penchés sur sa reconnaissance automatique sans vraiment parvenir à résoudre le problème complètement. Nous porterons notre contribution à l'éclaircissement de certains points en nous intéressant à reconnaître la voix de différents locuteurs.

Dans le cadre d'une application de reconnaissance du locuteur, nous avons à disposition une base de données dans laquelle sont stockées les références des voix des locuteurs qui vont accéder à cette application.
Ces locuteurs seront appelés les **clients** de l'application. La reconnaissance d'un client par sa voix se déroule en 2 phases: tout d'abord, nous devons identifier la voix de celui-ci parmi les voix stockées dans la base de données de l'application. Si cette opération est effectuée en analysant le signal de parole et en regardant à quel client la séquence appartient avec **la plus grande vraisemblance**, on parlera d'**identification du locuteur**. Lorsque l'identification du locuteur s'effectue par un autre moyen (code personnel, etc. . .) il reste à vérifier que le segment de parole testé appartient bien au locuteur, on parlera alors de **vérification du locuteur**. Selon qu'on tienne compte du texte prononcé par le locuteur ou que l'on ne s'intéresse qu'à des paramètres de celui-ci sans tenir compte de ce qu'il prononce, on parlera d'applications **dépendantes du texte** ou **indépendantes du texte** c'est ce mode qui nous nous intéresse. Si le locuteur ne connaît pas à l'avance le texte qu'il doit prononcer et que l'application le lui impose, on parle alors de "**textprompted**" . Quel que soit le type choisi, le processus de reconnaissance automatique du locuteur peut être décomposé en quatre parties principales ordonnées chronologiquement [2]:

1. Le **prétraitement** du signal qui permet de compenser les déformations dues à la transmission du signal de parole, tels que le micro ou le canal téléphonique (section 1.2.1).
2. L'**analyse** du signal qui extrait les éléments caractéristiques du signal de parole.
3. La **modélisation** et mémorisation des paramètres caractéristiques du locuteur.

4. Le module de **décision** qui permet de tester si un échantillon de parole appartient bien au locuteur dont on vérifie l'identité.

1.2.1 Prétraitement du signal de parole

De manière à atténuer les déformations du signal dues à l'environnement (échos, bruits de fond) et à tous les éléments intermédiaires nécessaires à le capter (micros), à le transmettre lignes téléphoniques) ou à l'enregistrer (convertisseurs analogique/numérique, déformations dues aux têtes d'enregistrement magnétique), un certain nombre de stratégies, de méthodes et d'algorithmes sont déployés. Pour la plupart, ce sont ceux utilisés dans le domaine du traitement du signal, avec cependant quelques particularités dues au signal de parole lui-même, citons-en quelques-unes ici:

– La plus grande partie de l'énergie du signal de parole se trouve entre 0 et 4000 [Hertz].

– Le signal de parole est très redondant.

– On peut considérer que le signal varie de manière lente et qu'il est stationnaire sur une période d'environ 5 à 10 ms.

Ces considérations sur le signal de parole sont relativement manquent de finesse puisqu'on peut discerner en tous cas 5 formants, le 5ème formant étant pour les hommes au-dessus de 4000 [Hz] en général et que pour les femmes les 4ème et 5ème formants se situent au-dessus de 4000 [Hz] . De plus, certaines plosives peuvent avoir une durée plus courte que 10 ms. Cependant ces caractéristiques sont celles qui ont permis de définir la largeur de bande téléphonique (300-3400 [Hz]) et qui sont utilisées encore de nos jours en codage GSM par exemple, puisque nous utilisons de la parole téléphonique, échantillonnée à 8 [kHz], Notons encore que certains traitements, visant à compenser les déformations de la bande téléphonique, s'appliquent sur les coefficients cepstraux directement (chapitre 2).

1.2.2 Analyse de la parole

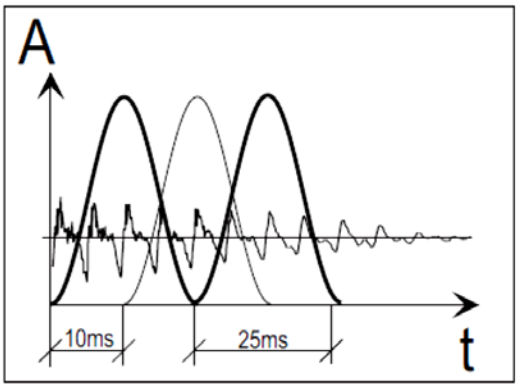


FIG. 1.10 – Analyse du signal de parole par fenêtrage court terme.

Il est possible d'identifier des motifs correspondant à des unités spectro-temporelles reconnaissables d'un individu à l'autre (les formants). Cependant, il est nécessaire, nous l'avons vu, que l'analyse du spectre se fasse sur des fenêtres temporelles de courte durée. La largeur des fenêtres d'analyse choisies sont de 20 à 40 ms réparties toutes les 10 ms (figure 1.10). De manière à compenser la distorsion crée par l'analyse fréquentielle sur des durées finies, on peut utiliser une fenêtre de type Hamming, bien connue en traitement de signal. Une phase de préaccentuation du signal est aussi utilisée pour compenser la pente du spectre (le premier formant contient plus d'énergie que les suivants). On utilise pour cela généralement un filtre RIF de premier ordre avec un coefficient ($0.9<\alpha<1.0$), on va voir les détails sur l'analyse du signal parole dans le deuxième chapitre [1].

1.2.3 Modélisation de locuteurs

Que ce soit pour reconnaître le message prononcé par un locuteur ou son identité, il nous est nécessaire de modéliser les entités que nous voulions reconnaître automatiquement. Notre connaissance du cerveau humain ne nous aide pas beaucoup ici, car si l'analyse du signal effectuée par l'oreille humaine semble plus ou moins connue, il en va tout autrement de ce que fait le cerveau avec les informations reçues par la cochlée, de leur stockage et de leur interprétation.

Les systèmes de reconnaissance automatique de la parole et du locuteur actuels utilisent pour la plupart des algorithmes de comparaison de motifs ("patterns matching" en anglais). Les motifs utilisés sont basés sur des parties de spectrogrammes qui ont été évoqués dans la section 1.1 et dont on utilise les représentations cepstralesduchapitre2.

Dans le cadre de la **reconnaissance de locuteurs**, nous modéliserons les différentes prononciations qu'*un locuteur* peut avoir effectuées pour le *même motif*. En étudiant la parole de locuteurs sur plusieurs prononciations des mêmes motifs, nous pouvons distinguer des variabilités caractéristiques du signal de parole nous permettant de séparer les locuteurs les uns des autres (variabilités inter-locuteurs) et d'autres, intrinsèques au locuteur (variabilités intra-locuteur).

1.2.3.1 Les variabilités caractéristiques du signal de parole

Variabilités intra-locuteur La voix humaine, à la différence des empreintes digitales, varie avec le temps ou les conditions physiologiques et psychologiques du locuteur. Cependant, ces variations intra-locuteur ne sont pas identiques pour tous les humains.
En effet, hormis les variations lentes de la voix dues au vieillissement, certains phénomènes extérieurs tels que la fumée ou l'état de santé d'une personne ont une influence variable sur sa voix [3].

Variabilités inter-locuteurs Elles proviennent des différences physiologiques (différences dimensionnelles du conduit vocal, fréquence d'oscillation des cordes vocales) et de différences de style de prononciation (accent, niveau social, etc. . .).
Certaines de ces différences, qui influencent le spectre associé à chaque locuteur, vont nous permettre de les séparer.

1.2.3.2 Le locuteur et le monde

Reconnaître un locuteur revient à essayer de le distinguer des autres.
Cependant, quelle que soit la modélisation choisie, il est nécessaire de définir ce qui n'est pas le locuteur, ou, en d'autres termes, de trouver une mesure qui permette d'estimer les dimensions de l'hypercube dans lequel varient les paramètres du locuteur.

Plusieurs manières de faire existent comme les modèles de voisinage ou les modèles de cohortes ou de monde [4].

Dans les applications pratiques, on distingue aussi les **clients**, qui sont des locuteurs dont on a enregistré des références et qui sont autorisés à pénétrer dans le système et les **imposteurs** qui sont des locuteurs qui tentent de se faire passer pour un client donné.

(voir la figure 1.11 du Schéma modulaire d'un système d'IAL).

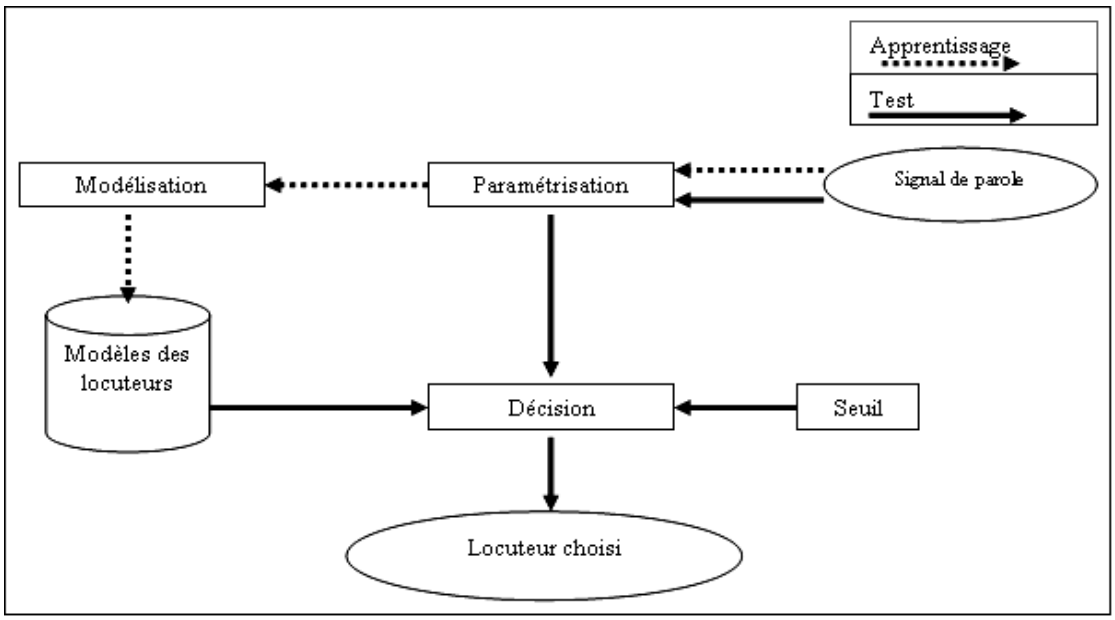


FIG. 1.11 Schéma modulaire d'un système d'IAL

1.2.4 MESURES ET DECISIONS EN RECONNAISSANCE DU LOCUTEUR

1.2.4.1 TEST D'HYPOTHESES (CALCUL DES SCORES)

En effet, dans tous les systèmes de reconnaissance du locuteur il faut, à un moment ou à un autre, prendre la décision d'accepter ou de rejeter un segment de parole (noté O_t ici la suite des vecteurs issus de l'étage de paramétrisation) comme appartenant au client dont on cherche à vérifier l'identité. Ce genre de décisions peut se comprendre comme un test d'hypothèse statistique H_0 (le segment considéré appartient au locuteur) contre H_1 (ce segment n'appartient pas au locuteur). Ce qui revient à tester la probabilité conditionnelle d'un événement X sachant les hypothèses $H0$ et Hi (équations 1.20). Les quantités $P(H0|X)$ et $P(H_1\backslash X)$ sont appelées probabilités *a posteriori* de l'hypothèse $H0$, respectivement H_1 sachant l'événement X.

$$H_0 : \ Locuteur, \ H_1 = \overline{H_0}$$

$$P(H_0) \underset{reject}{\overset{accept}{\gtrless}} P(H_1), \quad P(H_0|X) \underset{reject}{\overset{accept}{\gtrless}} P(H_1|X) \tag{1.1}$$

Comme pratiquement nous ne savons pas modéliser ce qui n'est pas le locuteur (virtuellement tous les autres locuteurs passés, présents et futurs de cette planète vivant ou ayant vécu en même temps que le locuteur considéré), nous traduisons l'hypothèse H_1 en : *"Ce segment appartient à un grand nombre de locuteurs qui ne sont ni des clients ni des imposteurs de l'application considérée ". H_1* sera modélisée par un *modèle de monde* qui respectera cette approximation de l'hypothèse de départ. Appliquée au problème de la vérification, l'équation 1.20 devient donc le test de la probabilité *a posteriori* du modèle du client Me sachant la séquence d'observation 0_t, contre la probabilité *a posteriori* du modèle de monde *Mw* sachant la même séquence d'observation 0_t (équation1.2) :

$$P(M_C|O_t) \underset{reject}{\overset{accept}{\gtrless}} P(M_W|O_t) \tag{1.2}$$

En utilisant la première règle de Bayes (équation 1.3) et connaissant la probabilité *a priori* qu'une séquence de parole appartienne au locuteur $P(M_C)$ ou au modèle de monde $P(M_W)$, on peut en déduire les probabilités *a posteriori* $P(O_t \setminus M_C)$et $P(O_t \setminus M_W)$ que la séquence O_t soit issue du modèle M_C ou du modèle M_W respectivement :

$$P(B/A) = \frac{P(A/B)P(B)}{P(A)} \qquad \textbf{(Bayes)} \tag{1.3}$$

Fonction d'erreur Supposons maintenant que pour un système donné nous cherchions à minimiser le coût total de ses erreurs. On peut pour cela définir une **fonction de coût** qui est la somme des erreurs faites en acceptant faussement une séquence qui n'est pas du locuteur (**fausse acceptation: FA**) ou en rejetant faussement une séquence qui appartient au locuteur (**faux rejet: FR**) [5].

1.2.4.2 **Taux d'erreur**

De manière à représenter les performances d'un système de reconnaissance de locuteur, plusieurs mesures sont possibles. En effet, selon que l'on veuille connaître ses performances intrinsèques ou en exploitation, nous n'utiliserons pas les mêmes mesures. On calcule les performances intrinsèques d'un système à partir des scores obtenus en utilisant des données de test de clients et d'imposteurs. On estime ensuite *a posteriori* les taux d'erreur lorsque l'on fait varier le seuil de décision sur toute la plage de réglage du système.

Le résultat d'une telle mesure est une courbe **COR** (Caractéristique Opérationnelle du Récepteur) qui se situe dans le plan des Faux Rejets/ Fausses Acceptations. Un des points de la courbe COR est très populaire car il correspond à un taux d'égale erreur de fausses acceptations et de faux rejets.

Pour connaître les performances d'un système en exploitation, on calcule d'abord un **point de fonctionnement** en utilisant des données de réglage. Ce point est estimé à partir du facteur de risque qui définit un seuil de décision. On évalue ensuite les performances du système avec le seuil (-) déterminé *a priori* ce qui nous permet de calculer un taux de faux rejets **FR** et un taux de fausses acceptations **FA** en exploitation. On calcule également souvent la moyenne de ces deux erreurs, le taux de 1/2 erreur (**HTER**, Half Total Error Rate enanglais):

$$\text{HTER} = \frac{FA + FR}{2} \tag{1.4}$$

De manière à évaluer la significativité des résultats, ceux-ci seront présentés soit avec un taux en pourcent et le nombre de tests effectués, soit avec un taux d'erreur E en pourcent et une variance sur l'erreur var calculée de la manière suivante:

$$var = \pm 1.96 \cdot \sqrt{\frac{E \cdot (1 - E)}{N}} \tag{1.5}$$

Avec une limite de confiance de 0.95, cette formule n'est valable que pour un nombre d'échantillons N grand [2].

Conclusion

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

Il est alors indispensable de caractériser l'individu par une empreinte afin de le différencier des autres sans aucune ambiguïté, cette empreinte est une clé codant l'identité d'une personne sans redondance ni variabilité. La plupart des indices biométriques, comme les empreintes digitales ou génétiques, répondent à ces critères.
Il en est différemment pour la voix dont la disposition à varier est inscrit dans sa nature même. Si nous voulons vraiment parler d'empreinte vocale, il faut tenir compte du fait que la variabilité interlocuteur est plus importante que la variabilité intra locuteur.

La voix devient donc un indice biométrique intéressant à exploiter parce que, elle est disponible via le réseau téléphonique, contrairement aux autres indices.

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

Chapitre 2

Paramétrisation du signal

2.1 INTRODUCTION

Cette étape consiste à extraire du signal de parole les caractéristiques du locuteur qui vont permettre de le reconnaître. Généralement, on calcule un jeu de coefficients acoustiques à des intervalles de temps réguliers, sur des blocs de signal de longueur fixe. Ce jeu de coefficients constitue un vecteur acoustique (voir FIG 2.1). Les techniques de paramétrisation acoustique sont nombreuses. Néanmoins, on peut les regrouper en trois grandes familles :

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

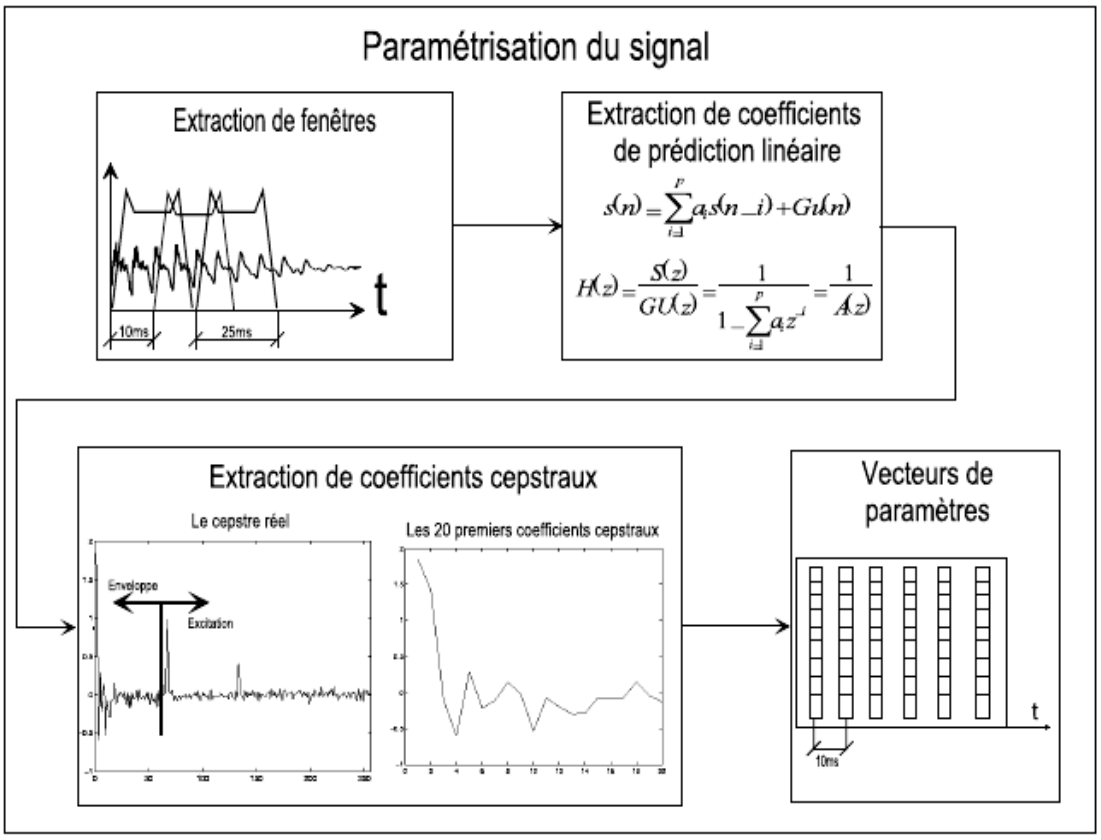


FIG.2 .1– Etage de paramétrisation de la parole [6]

2.2 Analyse et paramétrisation du signal vocal

2.2.1 Pré-traitement du signal

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

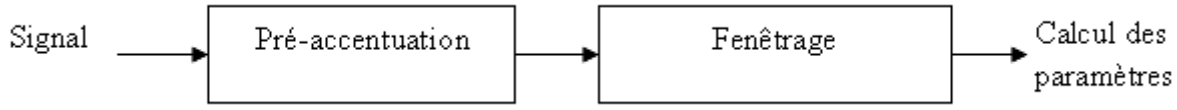


FIG. 2.2 Pré-traitement et extraction des paramètres

2.2.1.1 La pré-accentuation

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

$$H(z) = 1 - \alpha z^{-1} \tag{2.1}$$

Avec : $0.9 \leq \alpha \leq 1$.

2.2.1.2 Le fenêtrage

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

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

La fenêtre de Hamming est donnée par l'expression :

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad , \ 0 \le n \le N-1 \tag{2.2}$$

N : Le nombre d'échantillons dans une fenêtre.

2.2.2 Les paramètres acoustiques

2.2.2.1 L'énergie du signal

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

$$E = \sum_{n=0}^{N-1} s^2(n) \tag{2.3}$$

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

$$E = \ln\left(\sum_{n=0}^{N-1} s^2(n)\right) \tag{2.4}$$

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

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

2.2.2.2 Les coefficients de prédiction linéaire LPC

Ce modèle est basé sur l'appareil de production de la parole humaine. Ils font l'hypothèse que l'appareil phonatoire de la figure 1.2 peut se modéliser par une série de tubes sans pertes (voir la figure 2.3) et que la source du signal est soit un train d'impulsion de période 1/ F0

(approximation de la fonction de vibration des cordes vocales de la figure 1.8) pour les parties voisées du signal, soit une source de bruit Gaussien pour les parties non-voisées, soit les deux simultanément.

Les tubes sans pertes sont équivalents à des filtres tous pôles appliqués aux sources du signal. De manière à estimer les coefficients de ces filtres, on suppose que le signal de parole $(s(n)\ ,\ n \in \{1, \cdots N\})$ se prédit, à chaque instant, comme une combinaison des échantillons aux instants précédents.

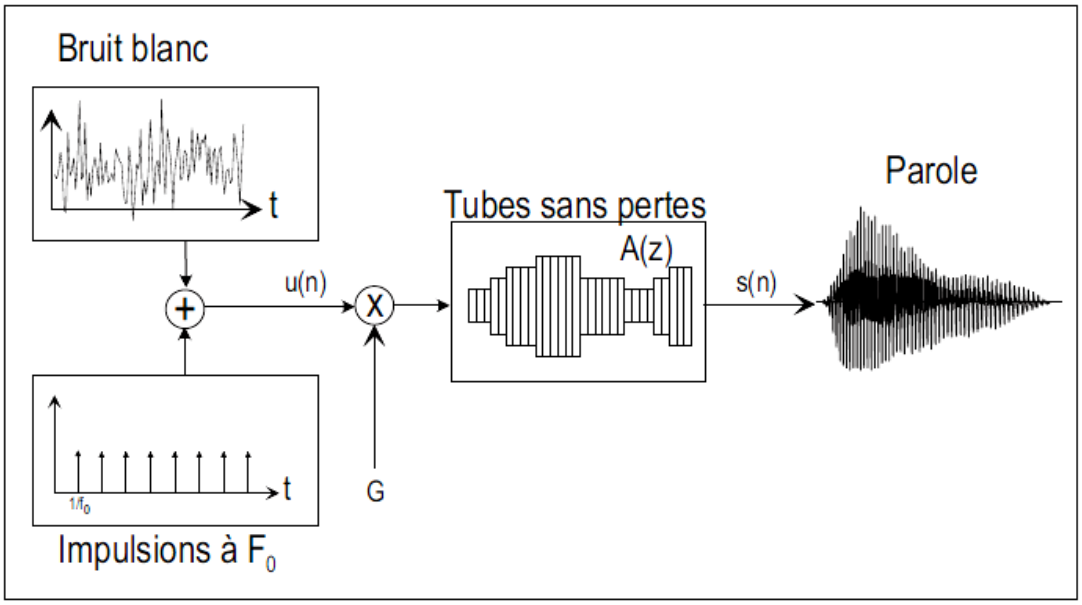


FIG. 2.3 – Modèle de production de la parole faisant l'hypothèse que le conduit vocal est une série de tubes sans perte, auquel on fournit une source de bruits et/ou une source impulsionnelle.

La fonction de transfert du système donné par la figure 2.3 $H(z)$ peut se mettre sous la forme :

$$H(z) = \frac{\sigma}{A(z)} \tag{2.5}$$

Avec :

$$A(z) = 1 + \sum_{i=1}^{P} a_i \ z^{-i} \tag{2.6}$$

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

covariance, soit sur le calcul de la matrice d'autocorrélation, pour plus des détails sur le modèle de la production de la parole voir annexe A [6].

2.2.2.3 Les coefficients cepstraux de prédiction linéaire LPCC

Les coefficients cepstraux peuvent être calculés à partir de la sortie d'un banc de filtres ou à partir des coefficients de prédiction linéaire, ainsi les coefficients LPCC (*LinearPredictionCepstral Coefficients*) sont dérivés directement des coefficients LPC [7].

Les coefficients cepstraux c_k sont obtenus par :

$$c_k = -a_k - \sum_{i=1}^{k-1} (1-\frac{i}{k}) a_i \; c_{k-i} \qquad , k > 0 \tag{2.7}$$

2.2.2.4 Les coefficients MFCC (Mel Frequency Cepstral Coefficients)

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

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

$$s(t) = e(t) * h(t) \tag{2.8}$$

Où $s(t)$ est le signal de parole, $e(t)$ est la source d'excitation glottique et $h(t)$ est la réponse impulsionnelle du conduit vocal.

L'application à l'équation (2.8) du logarithme du module de la transformée de Fourier donne :

$$log\left|S(f)\right| = log\left|E(f)\right| + log\left|H(f)\right| \tag{2.9}$$

Par une transformée de Fourier inverse, on obtient :

$$s'(cef) = e'(cef) + h'(cef) \tag{2.10}$$

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

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

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

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

- Echelle Mel :

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{2.11}$$

- Echelle Bark :

$$Bark(f) = 6\, Arc\sinh\left(\frac{f}{1000}\right) \tag{2.12}$$

f : représente la fréquence [Hz].

La procédure de calcul des coefficients MFCC est illustrée sur la figure 2.4 :

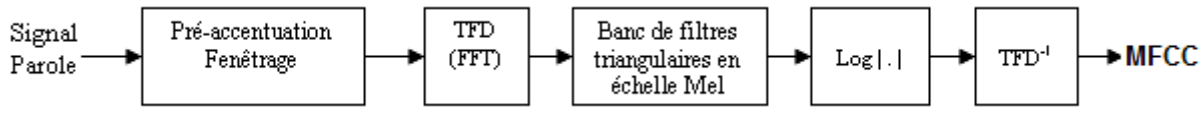


FIG. 2.4 Calcul des coefficients MFCC

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

$$S(k) = \sum_{n=0}^{N-1} s(n)\ exp\left(\frac{-j\,2\,\pi\,n\,k}{N}\right), \quad 0 \le k \le N-1 \tag{2.13}$$

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

$$B_m = B_b + m\ \frac{B_h - B_b}{M + 1}\ , \quad 0 \le m \le M+1 \tag{2.14}$$

M : Le nombre de filtres.

B_h : La fréquence la plus haute du signal.

B_b : La fréquence la plus basse du signal.

Dans le domaine fréquentiel, et d'après (2.11), les points f_m discrets correspondants sont calculés d'après :

$$f_m = B^{-1}\left(B_b + m\ \frac{B_h - B_b}{M + 1}\right) \tag{2.15}$$

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

$$B^{-1}(x) = 700\left(10^{\frac{x}{2595}} - 1\right) \tag{2.16}$$

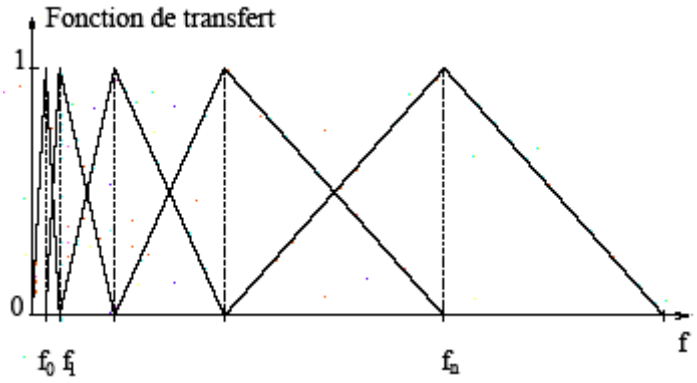


FIG. 2.5 Banc de filtres sur l'échelle linéaire

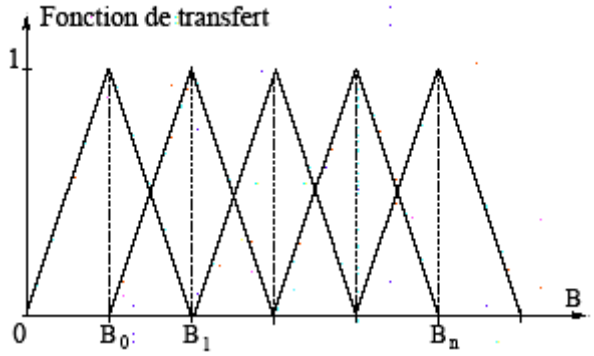


FIG. 2.6 Banc de filtres sur l'échelle Mel

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

$$c_k = \sum_{i=1}^{M} log\ E_i\ cos\left[\frac{\pi k}{M}\left(i-\frac{1}{2}\right)\right] \quad , 1 \le k \le d \tag{2.17}$$

et qui permet d'obtenir des coefficients peu corrélés.

Le coefficient c_0 qui est la somme des énergies n'est pas utilisé ; il est éventuellement remplacé par le logarithme de l'énergie totale E calculée dans le domaine temporel et normalisée [8][9].

2.2.2.5 Les coefficients LFCC (Linear Frequency Cepstral Coefficients)

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

2.2.3 Distances et mesures de dissemblance dans l'espace acoustique

2.2.3.1 Définitions et propriétés

Dans toute approche de reconnaissance, le choix de la distance associée à l'espace des paramètres est important. Il est possible d'utiliser toutes les distances classiques, en particulier les distances de Minkovski, parmi lesquelles la distance euclidienne, et la distance de Mahalanobis qui normalise les coefficients par leur matrice de covariance.

Des distances spécifiques aux espaces de représentation de la parole existent aussi, comme les distances cepstrales pondérées et la mesure d'Itakura pour les coefficients LPC. [11]

Dans un espace métrique, la distance entre deux vecteurs X et Y doit satisfaire les conditions suivantes :

1. $d(X,Y) \geq 0$
2. $d(X,Y) = d(Y,X)$ (2.18)
3. $d(X,Y) \leq d(X,U) + d(U,Y)$

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

La condition 2 peut être assurée en posant

$$d_S(X,Y) = \frac{1}{2}\left[d(X,Y) + d(Y,X)\right] \quad (2.19)$$

La condition 3 est rarement utile en traitement de la parole [12].
Pour les détails sur les distances (voir annexe B).

2.2.3.2 Distances adaptées à une représentation

1. Mesure d'Itakura

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

$$d_I(A,B)=log\left[\frac{A^t R_b A}{B^t R_b B}\right] \tag{2.20}$$

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

2. Distance cepstrale

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

$$V(\theta)=ln\, f(\theta)-ln\, f'(\theta) \tag{2.21}$$

et la distance spectrale logarithmique est la norme :

$$d_p=\left[\frac{1}{2\pi}\int_{-\pi}^{\pi}|V(\theta)|^p d\theta\right]^{1/p} \tag{2.22}$$

La norme d_2 est la plus utilisée. Toutefois, la distance spectrale logarithmique donnée par (2.39) est coûteuse en temps de calcul, alors on lui substitue la distance cepstrale.

Les coefficients du cepstre réel sont donnés par :

$$\ln f(\theta) = \sum_{n} c(n) \exp(-jn\theta) \tag{2.23}$$

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

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

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

$$d_2^2 = \sum_{l} \left(c(l) - c'(l) \right)^2 = \left(c(0) - c'(0) \right)^2 + 2 \sum_{l=1}^{\infty} \left(c(l) - c'(l) \right)^2 \tag{2.24}$$

La distance cepstrale est basée sur un nombre fini L de termes [12]:

$$d_{CEP} = \left(c(0) - c'(0) \right)^2 + 2 \sum_{l=1}^{L} \left(c(l) - c'(l) \right)^2 \tag{2.25}$$

2.3 Conclusion

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

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

Chapitre 3

Modélisation des locuteurs

Introduction

De manière à mémoriser des caractéristiques qui dépendent du locuteur, nous utilisons des algorithmes capables de capturer les points communs entre différentes représentations de motifs spectraux issus du même locuteur (constituant ainsi un **modèle du locuteur**), tout en ayant la possibilité de s'adapter aux variations d'échelles fréquentielles et temporelles liées au signal de parole.

Ces motifs peuvent être soit des segments de parole déterminés (mots, phonèmes) si nous travaillons en mode dépendant du texte, soit des segments de parole dont on ne connaît pas le contenu phonétique si l'application fonctionne en mode indépendant du texte. Ces algorithmes doivent être couplés avec une mesure qui permettra de donner une valeur de distorsion (ou de similitude) entre le modèle du locuteur et un motif inconnu dont on cherche à déterminer la provenance.

Le chapitre 2 a montré quels sont les paramètres discriminant les locuteurs. Dans ce chapitre, nous allons nous intéresser à modéliser ces paramètres de manière à modéliser les caractéristiques du locuteur considéré en mode indépendant du texte en utilisant la modélisation par mélanges de gaussiennes qui fournit de bonnes performances et qui constitue l'état de l'art en la matière.

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

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

Dans la dernière partie, nous allons introduire la technique de modélisation : QV,
en suite on fait une étude comparative entre les deux techniques (GMM, QV).

.

3.1 Modélisation des locuteurs

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

- L'approche vectorielle : le locuteur est représenté par un ensemble de vecteurs issus directement de la phase de paramétrisation. Ses principales techniques sont la reconnaissance à base de l'alignement temporel dynamique (DTW) et par quantification vectorielle.
- L'approche statistique : consiste à représenter chaque locuteur par une densité de probabilité dans l'espace des paramètres acoustiques. Elle couvre les techniques de modélisation par les modèles de Markov cachés (HMM), par les mélanges de gaussiennes (GMM) et par des mesures statistiques du second ordre.
- L'approche connexionniste : consiste principalement à modéliser les locuteurs par des réseaux de neurones.
- L'approche relative : il s'agit de modéliser un locuteur non pas de façon absolue mais relativement par rapport à d'autres locuteurs de référence, dont les modèles sont bien déterminés.

3.1.1 L'approche vectorielle

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

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

Malgré les bonnes performances obtenues par cette technique, elle reste très sensible à la qualité de l'alignement et notamment le choix du point de départ des deux formes à comparer (voir la figure 3.1 qui montre le principe de la DTW) [13].

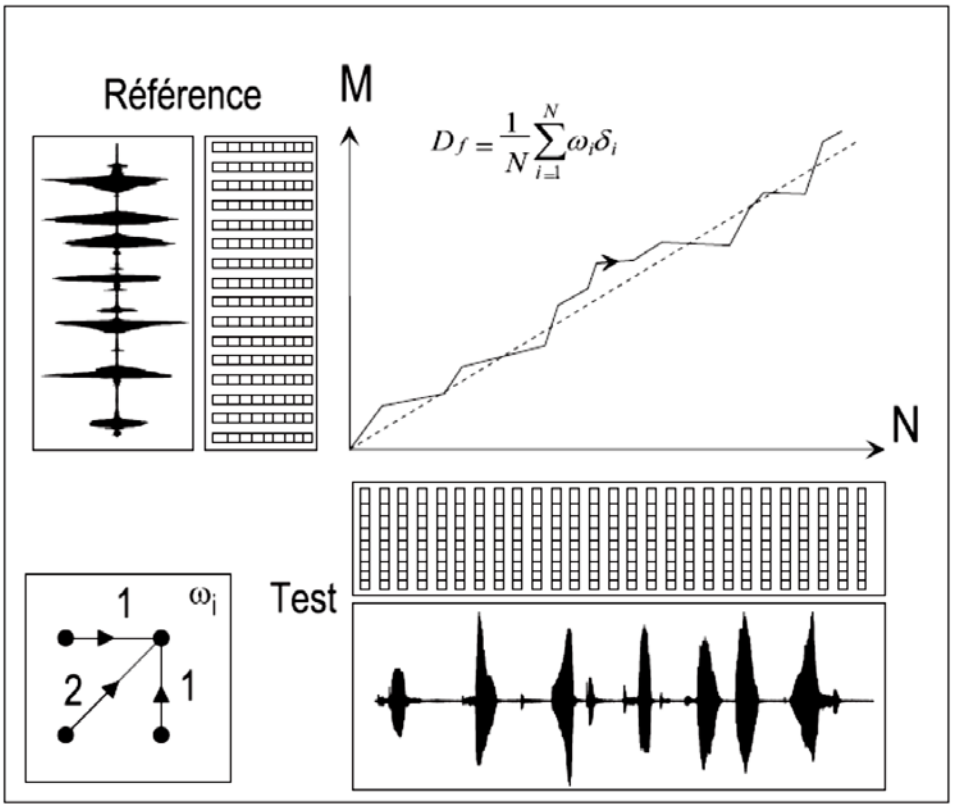


FIG. 3.1-Calcul de la distance dynamique pour un des coefficients extrait du signal parole par DTW.

Lors de la **phase d'entraînement**, les références des locuteurs sont simplement stockées.

Lors de la **phase d'exploitation**, la distance finale D_f entre une référence et la séquence de test est calculée comme la somme des distances partielles (δ_i) entre vecteurs le long du chemin optimal, pondérées par des contraintes locales (voir la figure 3.1 pour un exemple de contrainte locale). Le chemin optimal est le chemin dont la distance globale est minimale. Afin d'augmenter l'efficacité de la méthode, les chemins possibles sont limités et toute une gamme de contraintes locales peut être imposée. Lorsque l'on possède plusieurs exemplaires de référence R_j, le score de sortie final est la moyenne de toutes les distances calculées sur toutes les références.

Avantages de la DTW L'algorithme de DTW est rapide, bien adapté à la parole parce que capable de tenir compte des variations temporelles du signal. Il ne nécessite pas beaucoup de données pour fonctionner correctement.

Inconvénients de la DTW La DTW est très sensible à la segmentation du signal. En effet, si le point de départ du calcul dynamique n'est pas bon, l'algorithme peut rapidement diverger du chemin optimal. Il existe cependant des possibilités de corriger partiellement cette erreur [14]. De plus, ses capacités de modélisation des variabilités intrinsèques du locuteur sont relativement limitées puisqu'il n'est capable d'estimer que des points sur le meilleur chemin et non des distributions.

3.1.1.2 La Quantification Vectorielle

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

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

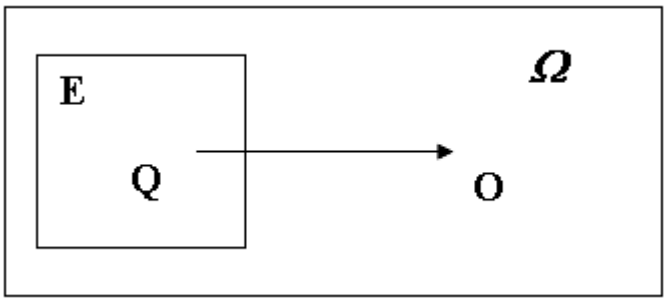
3.1.2 L'approche statistique

3.1.2.1 Les Modèles de Markov Cachés HMM : Hidden Markov Models

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

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

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



[4]

FIG. 3.2 Constituants d'un HMM

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

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

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

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

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

- Un ensemble de probabilités initiales de se trouver dans chaque état :

$$\pi = \{\pi_i \ / \ \pi_i = P(q_1 = S_i) \ i = 1, .., N \ \}.$$

Un modèle de Markov caché est donc spécifié par un triplet $\lambda = \{A, B, \pi\}$ où *A* est la matrice des probabilités de transition, *B* la matrice des probabilités d'observation et π les probabilités initiales [15].

Problèmes des modèles HMM

Trois problèmes se posent avec les modèles de Markov cachés :

1. L'évaluation

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

Il existe deux méthodes pour résoudre ce problème. La méthode dite directe et qui consiste à calculer cette probabilité en énumérant toutes les séquences d'états possibles de même longueur que la séquence d'observation. Cette technique demande beaucoup de temps de calcul. Un moyen plus rapide pour calculer cette probabilité est l'utilisation des algorithmes de programmation dynamique [16].

2. Estimation des états cachés

Le deuxième problème posé avec les HMM est le décodage qui consiste à chercher la séquence $Q = q_1 \, q_2 \, ... q_T$ d'état qui maximise la probabilité $P(O,Q / \lambda)$, étant donné une séquence d'observations $O = o_1 \, o_2 \, ... o_T$ et un modèle $\lambda = \{A, B, \pi\}$. Pour cela, l'algorithme de Viterbi est le plus utilisé. Il permet de chercher la séquence d'états cachés la plus probable en ne gardant que les états S_i qui maximisent la probabilité à chaque instant t [17].

3. Apprentissage

C'est le problème principal d'un modèle HMM. En effet, la qualité d'un système utilisant une modélisation HMM dépend principalement de la qualité de ses modèles. C'est pourquoi l'étape d'apprentissage qui consiste à estimer les paramètres des modèles HMM est très importante.

Il existe plusieurs méthodes pour résoudre ce problème, les plus utilisées sont :

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

La phase de reconnaissance

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

Le principal avantage de l'approche HMM est sa grande capacité d'apprendre les propriétés statistiques. En reconnaissance de locuteur le choix le plus fréquent consiste à utiliser un modèle dont la distribution conditionnelle dans chaque état est un mélange de gaussiennes. L'utilisation de ces modèles est plus importante dans le mode dépendant du texte parce qu'en mode indépendant du texte l'information supplémentaire apportée par les transitions entre états n'améliore pas les performances de la reconnaissance du locuteur [18].

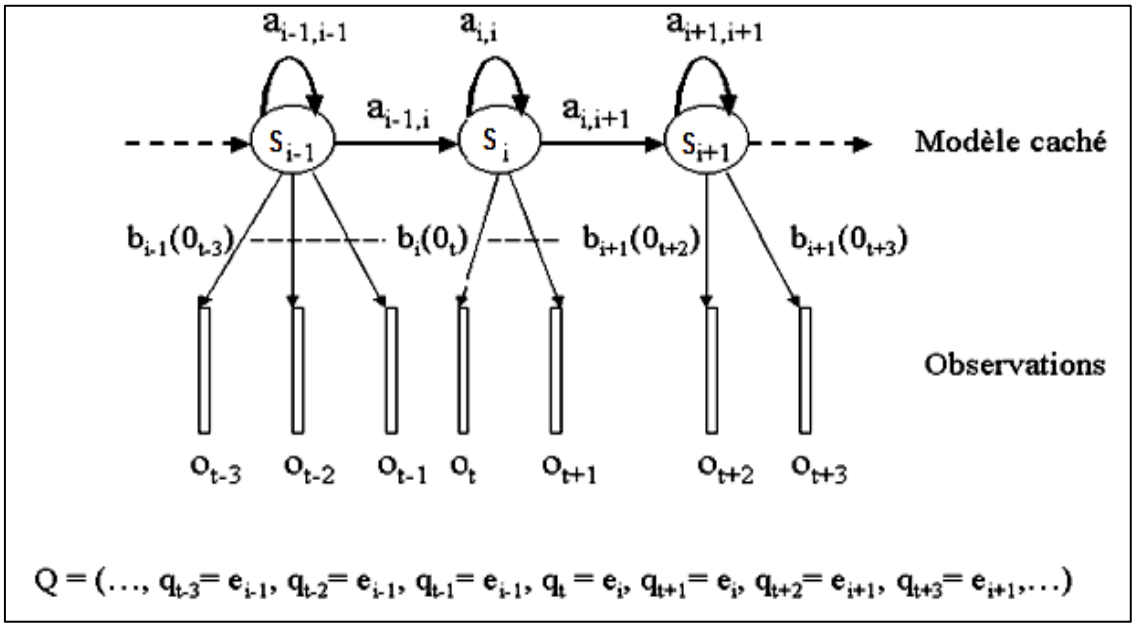


FIG. 3.3 Exemple d'une machine Markovienne [17]

3.1.2.2 Les mélanges de gaussiennes

La reconnaissance du locuteur par mélanges de gaussiennes (ou GMM pour *Gaussian Mixture Models*) consiste à modéliser un locuteur par une somme pondérée de composantes gaussiennes. Chaque composante gaussienne est supposée modéliser un ensemble de classes acoustiques.

Les GMM sont considérés comme un cas particulier des HMM et une extension de la quantification vectorielle [19].

3.1.2.3 Mesures statistiques du second ordre

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

3.1.3 L'approche connexionniste

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

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

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

On peut aussi utiliser les réseaux de neurones en les couplant à d'autres techniques, comme par exemple les modèles de Markov cachés. On parle alors de méthodes hybrides [20].

3.1.4 L'approche relative

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

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

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

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

3.2.1 Les mélanges de gaussiennes

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

L'utilisation d'un modèle GMM se justifie essentiellement en faisant appel à l'interprétation des classes du mélange. En effet, les vecteurs de paramètres vont se répartir différemment selon les caractéristiques du son de parole considéré (son voisé / non voisé, ou plus finement en fonction du phonème). Chaque composante va modéliser des ensembles sous-jacents de classes acoustiques, chaque classe représentant des évènements acoustiques (voyelles, nasales, …). Ainsi, l'allure spectrale de la ième composante pourra être représentée par sa moyenne et sa matrice de covariance. Ces classes caractérisent l'espace acoustique propre à chaque locuteur.

L'autre raison poussant à utiliser les GMM est qu'à l'aide d'une combinaison linéaire de composantes gaussiennes, on peut représenter une large gamme de distributions.

3.2.2 Modèle du mélange

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

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

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

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

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

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

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

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

Ce modèle peut prendre plusieurs formes, notamment en ce qui concerne les matrices de covariance. On peut utiliser une matrice de covariance pour chaque gaussienne, ou bien une matrice de covariance globale, commune à toutes les gaussiennes [21].

3.2.3 Apprentissage du modèle

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

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

Avantages des GMM [19] Avec un mélange contenant beaucoup de Gaussiennes, la modélisation GMM donne d'excellents résultats en reconnaissance du locuteur indépendante du texte (cet algorithme est à l'état de l'art dans ce domaine). Il est possible de fusionner des GMM de manière à tenir compte de l'environnement.

Inconvénients des GMM Bien qu'ils soient capables de capturer les informations à plus long terme d'un locuteur, ils ne contiennent pas d'aspects dynamiques. Pour une bonne modélisation (i.e. beaucoup de Gaussiennes) nécessitent beaucoup de données.

3.2.3.1 Quantification Vectorielle

La quantification vectorielle est une généralisation de la quantification scalaire. Elle consiste à substituer à un vecteur x, dont les composantes sont à valeurs réelles continues ($x \in R^k$), un vecteur voisin appartenant à l'ensemble fini $\left\{ y_i \in R^k \ , \ i = 1, 2, ..., M \right\}$. Les vecteurs y_i sont dits vecteurs quantifiés, et constituent un dictionnaire (code-book) de points dans R^k.

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

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

3.2.3.2 Algorithme LBG

L'algorithme LBG (Linde Buzo Gray) est développé pour la quantification vectorielle. Son objectif est de minimiser la distorsion totale donnée par :

$$D = \sum_{i=1}^{M} \sum_{t=1}^{T} \left\| x_t - \mu_i \right\| \tag{3.4}$$

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

Pour l'apprentissage des modèles GMM avec l'algorithme LBG, on suivra les étapes suivantes [22] :

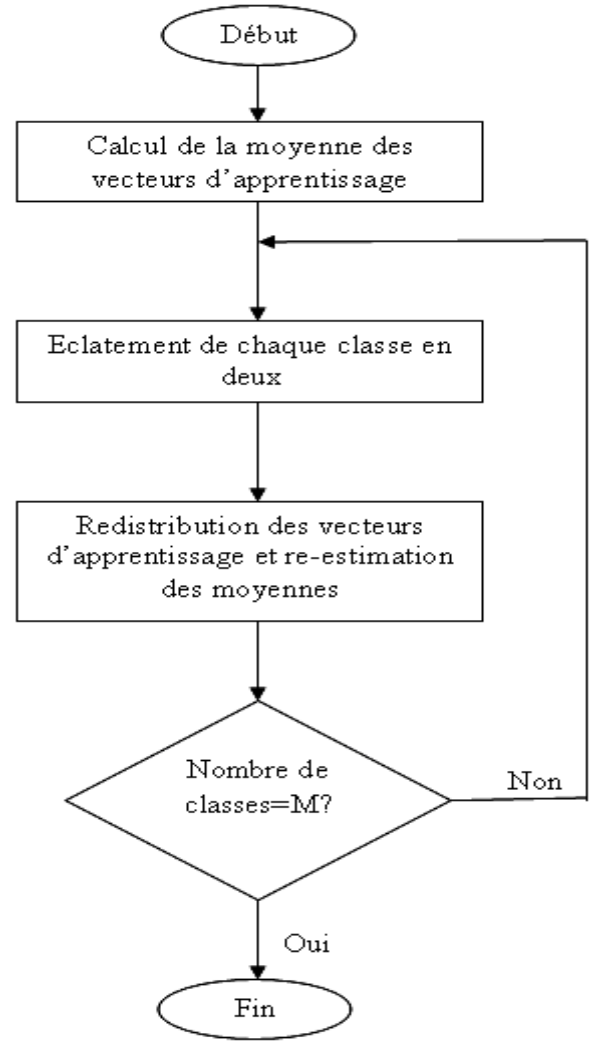


FIG. 3.4 L'algorithme LBG

1. Initialisation

Les vecteurs d'apprentissage $\{x_1, x_2, \ldots\ldots, x_T\}$ sont supposés appartenir à une seule classe dont le centroïde est la moyenne de cette classe, et est donné par :

$$\mu_1 = \frac{1}{T}\sum_{t=1}^{T} x_t \ . \tag{3.5}$$

2. Eclatement du dictionnaire

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

3. Optimisation du dictionnaire

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

4. Teste d'arrêt

Tant que $M < 2^p$ le dictionnaire est à nouveau éclaté et optimisé on réitérant les étapes 2 et 3. La figure 3.4 illustre l'organigramme de l'algorithme LBG.

3.2.3.3 Apprentissage par Maximum de vraisemblance

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

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

$$p(X/\lambda) = \prod_{n=1}^{N} p(x_n/\lambda) = \prod_{n=1}^{N}\sum_{m=1}^{M} p(x_n/\pi_m, \mu_m, \Sigma_m) \tag{3.6}$$

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

- Une étape d'initialisation qui permet l'obtention des valeurs approximatives des paramètres du modèle par l'algorithme LBG.

- Une étape d'optimisation des valeurs de ces paramètres par un algorithme de type EM (Expectation-Maximisation).

Algorithme EM (Expectation-Maximisation)

L'algorithme EM fait intervenir des variables latentes que l'on ne peut observer directement. Dans notre cas, chaque vecteur x_j est défini non seulement par les D paramètres acoustiques mais aussi par le sous-ensemble S_i (défini par un centroïde) auquel il se rattache. Dans le cas de l'algorithme LBG, on a vu que chaque vecteur x se rattache réellement à un sous-ensemble. Dans le cas de l'algorithme EM ce ne sera plus le cas. Celui-ci va maximiser la vraisemblance de façon itérative, mais le vecteur x sera maintenant rattaché aux M sous-ensembles S_i avec une probabilité particulière, sans que l'on puisse déterminer à quel sous-ensemble S_i il appartient. C'est ce paramètre que l'on qualifie de donnée cachée ou latente.

La maximisation de la fonction de vraisemblance fait intervenir la fonction auxiliaire $Q(\theta,\theta^{(t)})$ qui est définie comme étant l'espérance mathématique du logarithme de la vraisemblance jointe (incluant les variables observée et les variables cachées) sur l'ensemble complet des variables d'apprentissage.

$$Q(\theta,\theta^{(t)}) = \sum_{m=1}^{M} \sum_{n=1}^{N} p(m / x_n, \theta^{(t)}) \log p(x_n, m / \theta) \tag{3.7}$$

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

$$\begin{aligned} Q(\theta,\theta^{(t)}) = & \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\log \pi_m - \frac{D}{2} \log(2\pi) - \frac{1}{2} \log|\Sigma_m| \right] \\ & - \sum_{m=1}^{M} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \left[\frac{1}{2} (x_n - \mu_m)' \Sigma_m^{-1} (x_{n,} - \mu_m) \right] \end{aligned} \tag{3.8}$$

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

$$\gamma_{n,m}^{(t)} = \frac{\pi_m^{(t)}\ p(x_n / \mu_m^{(t)}, \Sigma_m^{(t)})}{\sum_{k=1}^{M} \pi_k^{(t)}\ p(x_n / \mu_k^{(t)}, \Sigma_k^{(t)})} \tag{3.9}$$

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

Les formules de calcul des paramètres $(\pi_m^{(t+1}, \mu_m^{(t+1)}, \Sigma_m^{(t+1)})$ sont données par [23] :

$$\pi_m^{(t+1)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,m}^{(t)} \tag{3.10}$$

$$\mu_m^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}\ x_n}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{3.11}$$

$$\Sigma_m^{(t+1)} = \frac{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}\ (x_n - \mu_m^{(t)})(x_n - \mu_m^{(t)})'}{\sum_{n=1}^{N} \gamma_{n,m}^{(t)}} \tag{3.12}$$

3.2.4 Décision d'un système d'identification

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

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

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

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

$$\hat{s} = arg \max_{1 \le s \le \ell} \frac{p(X/\lambda_s)}{p(X)} p(\lambda_s) \tag{3.14}$$

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

$$\hat{s} = arg \max_{1 \le s \le \ell} p(X / \lambda_s) \tag{3.15}$$

En utilisant le logarithme et l'indépendance entre les observations, le système d'identification se base sur l'équation[24] :

$$\hat{s} = arg \max_{1 \le s \le \ell} \sum_{n=1}^{N} log\, p(x_n / \lambda_s) \tag{3.16}$$

3.2.5 Mesure des performances d'un système d'identification

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

$$I_c = \frac{\text{Nombre de segments de test correctement identifiés}}{\text{Nombre total de segments de tests}} \times 100 \tag{3.17}$$

$$I_i = \frac{\text{Nombre de fausses identifications}}{\text{Nombre total de segments de tests}} \times 100 \tag{3.18}$$

Avec :

$$I_c + I_i = 100\,\% \tag{3.19}$$

3.3 Quantification vectorielle (QV)

Cette technique permet une compression considérable des données. Il s'agit de représenter l'espace acoustique par un nombre fini de vecteurs acoustiques formant un dictionnaire, et ce en faisant un partitionnement de cet espace en régions ou classes, qui seront représentées par leurs vecteurs centroïdes. Ainsi, chaque locuteur va être représenté par son dictionnaire de quantification [25]. Les performances et la rapidité de cette technique dépendent fortement de la taille du dictionnaire. En effet, plus la taille du dictionnaire est grande, meilleures sont les performances, mais le processus de test devient lent.

3.3.1 Définition

La quantification vectorielle consiste à représenter tout vecteur x de dimension k par un autre vecteur y_i de même dimension appartenant à un ensemble fini D de Lvecteurs. Les y_i sont appelés les codes vecteurs. D est appelé le dictionnaire ou catalogue des formes. La quantification vectorielle permet d'avoir une constellation qui minimise la distorsion moyenne pour un dictionnaire de taille k donnée. La quantification vectorielle peut fournir un décodage rapide en utilisant une table simple d'identification. La figure 3.5 illustre ce principe.

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

$Q: R^k \rightarrow D$

$x \rightarrow Q(x) = y_i$

$D = \{y_i \in R^k / i=1, 2,\ldots, L\}$

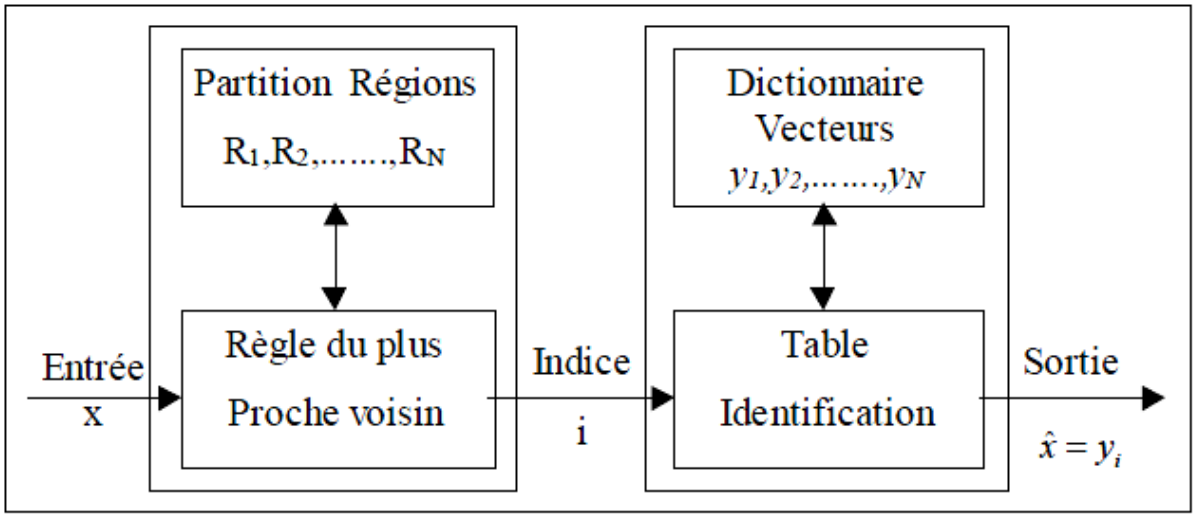


Fig. 3.5 Modèle du quantificateur vectoriel.

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

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

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

$$D_Q = E\left[d\left(x, Q(x)\right)\right] = \int d\left(x, Q(x)\right) p(x)\, dx \qquad (3.20)$$

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

$$D_Q = \frac{1}{N}\sum_{j=1}^{N} d(x_j, Q(x_j)) \qquad (3.21)$$

On appelle centroïde de la classe C_i,le vecteur c_i tel que sa distance moyenne à tous les éléments de la classe soit minimale (en géométrie euclidienne, le centroïde correspond au centre de gravité).

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

3.3.2 Quantificateur Vectoriel Optimal

Un quantificateur se décompose en deux applications : un codeur et un décodeur.

Le quantificateur optimal est alors celui réunissant le codeur optimal et le décodeur optimal.

- Le codeur optimal : étant donné un dictionnaire $\{y_1,y_2,\ldots,y_L\}$, la meilleure partition est celle qui vérifie :
- $R_i = \{ x \in R^k / d(y_i, x) \leq d\left(y_j, x\right) \forall\, j \in \{1,...,L\}\}$ pour $i=1,2,\ldots,L$.
  C'est la règle dite du plus proche voisin.
- Le décodeur optimal : étant donné une partition $\{R_1,R_2,\ldots,R_L\}$, les meilleurs représentants sont obtenus par la condition du centroïde (centre de gravité c_i de la partie de la densité de probabilité placée dans la région R_i).
- Une troisième condition est nécessaire : il faut que la probabilité qu'un vecteur à coder se trouve à la même distance de deux représentants soit nulle, sinon ce vecteur source est affecté à l'un des deux représentants, et dans ce cas, la partition de l'espace n'est

plus optimale. Si les vectrices sources sont à amplitude continue, cette troisième condition est toujours vérifiée.

Ces trois conditions conduisent à la conception d'un algorithme qui réalise, à partir d'une séquence d'apprentissage représentative de la statistique de la source à coder, la construction d'un dictionnaire optimal.

Pour l'apprentissage de ce modèle on utilise l'algorithme LBG de la façon suivante :

1. Initialiser le dictionnaire.
2. Construction des *L* régions de Voronoï associées aux *L* représentants du dictionnaire.
3. Calcul des *L* nouveaux centroïdes (représentants) correspondants à partir de la relation : $\mathrm{c_i} = \frac{1}{N_i}\sum_{\{k;x_k \in R_i\}} x_k$ (3.22)
4. Répéter les étapes 2 et 3 tant que la croissance de la distorsion moyenne reste importante.

L'algorithme présente le problème lié à l'initialisation du dictionnaire : il n'est pas sûr que l'algorithme converge vers le minimum global [26]. Pour éviter ce problème, l'algorithme LBG possède la procédure d'initialisation efficace dite par dichotomie vectorielle (Split BinaryMethod) que voici (figure 3.6) :

1. Calculer le centroïde de l'ensemble de toute la séquence d'apprentissage.
2. A partir de chaque centroïdec_i, on fabrique, par un petit déplacement dans deux directions opposées, deux nouveaux vecteurs :
   $c_i^+ = c_i(1+\varepsilon)$et $c_i^- = c_i(1-\varepsilon)$ avec : $0.01 \leq \varepsilon \leq 0.05$
   Cela a pour conséquence de doubler le nombre de vecteurs représentants.
3. Déterminer les régions de Voronoï associées à l'ensemble des vecteurs obtenus précédemment, ainsi que leurs centroïdes respectifs.
4. Répéter les étapes 2 et 3 jusqu'à l'obtention d'un ensemble de L centroïdes.
   La taille du dictionnaire L est une puissance de 2.

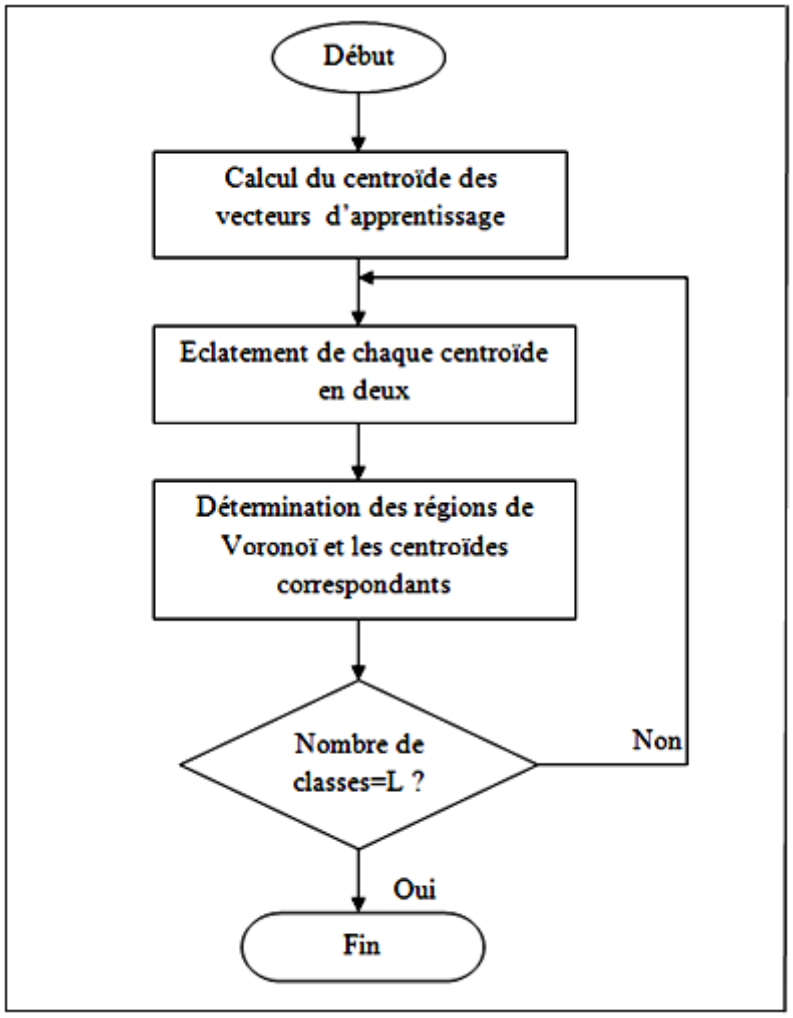


Fig. 3.6 Initialisation de l'algorithme LBG.

4 Conclusion

Nous avons décrit les différents modèles de classification. Nous nous sommes basés essentiellement sur l'approche statistique (GMM) et l'approche vectorielle (QV). Cette description a pour objet l'application de ces différentes approches sur une base de données, et évaluer leur performance en fonction des paramètres choisis.

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

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

L'introduction de l'algorithme LBG réduit considérablement les temps de calcul, et remédie ainsi aux problèmes posés par l'algorithme EM.

Chapitre 4

Evaluations expérimentales

Introduction

Ce chapitre est décomposé en deux grandes parties, la première partie décrit le contexte expérimental de toutes les expériences effectuées le long de cette étude, et dans la deuxième partie on va exposer et commenter les différents résultats obtenus et donner quelques conclusions.

4.1 Contexte expérimental

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

4.1.1 Base de données utilisée

Nous avons pris un échantillon de 38 locuteurs (33 hommes et 5 femmes) extrait de la base de données de l'Ecole Nationale Polytechnique. C'est une base acoustique dédiée seulement à la reconnaissance du locuteur. Elle est constituée de 45 locuteurs (37 hommes et 8 femmes) algériens. Les données sont échantillonnées avec 16KHz, sur 16 bits. L'enregistrement a été fait dans le laboratoire du signal et communication de l'Ecole Nationale Polytechnique sous les conditions :

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

Pour chaque locuteur, nous disposons de 10 phrases chacune de 3 secondes en moyenne. Nous avons concaténé 7 phrases pour l'apprentissage et les 3 autres phrases sont utilisées pour le test.

4.1.2 Analyse acoustique et paramétrisation du signal vocal

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

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

Dans mes expériences, une analyse est appliquée toutes les 15 ms sur des fenêtres d'analyse de 32 ms (par glissement et recouvrement des fenêtres d'analyse). A chaque trame, on associe un vecteur de représentation acoustique, composé des 16 premiers coefficients MFCC.

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

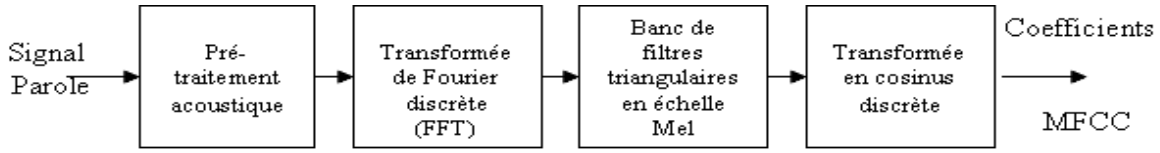


FIG. 4.1 Extraction des coefficients MFCC

La phase de pré-traitement acoustique contient deux étapes :

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

2. L'étape de fenêtrage qui consiste à multiplier le signal vocal par une fenêtre de pondération glissante. Dans notre travail, on a utilisé une fenêtre de Hamming de durée de 32 ms avec déplacement de 15 ms.

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

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

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

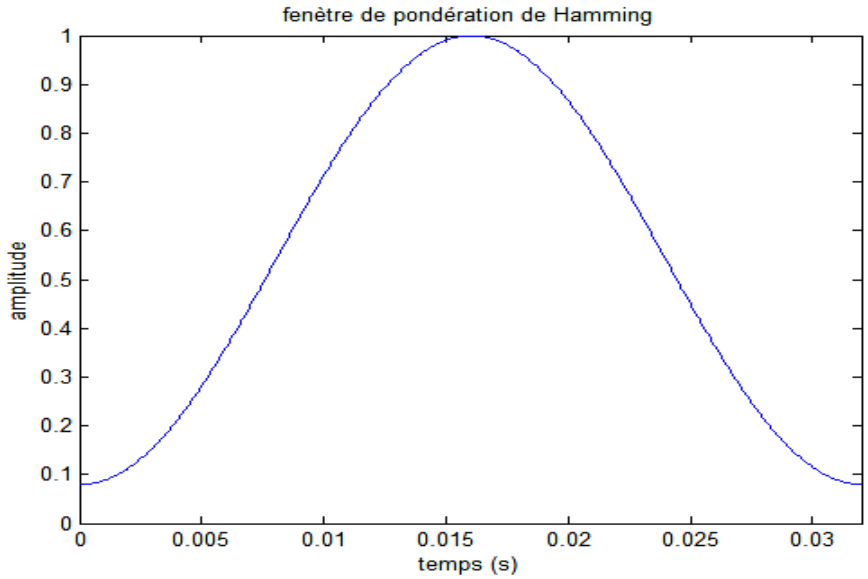


FIG. 4.2 Fenêtre de pondération de Hamming

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

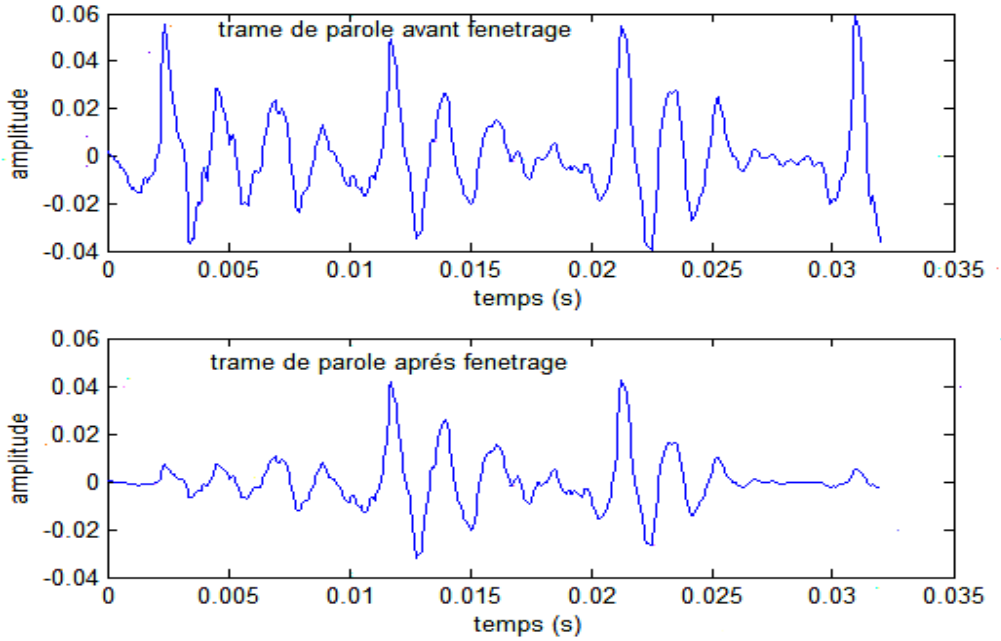


FIG. 4.3 Fenêtrage d'une trame de parole

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

a. Transformée de Fourier discrète

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

b. Banc de filtres triangulaire en échelle Mel

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

c. Transformée en cosinus discrète

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

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

4.1.3 Détection et élimination de silence

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

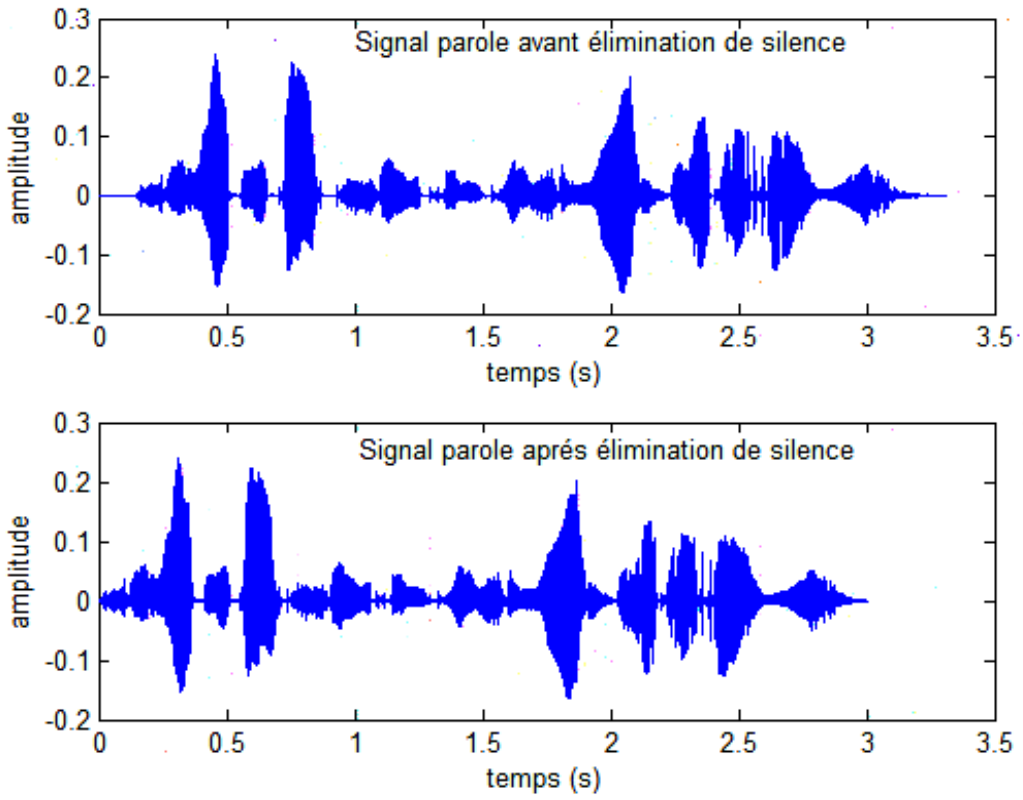


FIG. 4.4 Elimination de silence

4.1.4 Filtrage dans la bande téléphonique et ré-échantillonnage

Nous commençons à travailler avec une fréquence d'échantillonnage de 16 KHz. Ensuite, nous introduisons une dégradation sur les données à utiliser afin d'approcher la qualité du réseau téléphonique commuté RTC (filtrage dans la bande téléphonique [300Hz-3400KHz], sous-échantillonnage à 8KHz et l'ajout d'un bruit blanc gaussien avec un rapport signal sur bruit SNR variant de 10dB à 100dB).

4.1.5 Apprentissage des modèles

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

4.1.6 Protocole d'évaluation

Nous allons évaluer les performances des deux approches GMM et OGMM sur un ensemble de 38 locuteurs (ensemble fermé). Il s'agit d'identifier un locuteur parmi les 38 locuteurs et de calculer le taux d'identification correcte défini par :

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

Le test est effectué sur l'ensemble de tous les locuteurs, chaque locuteur a trois segments (phrases) de test, soit un total de 114 tests.

4.1.7 Langage utilisé

On a utilisé **MATLAB** version **7.5** qui possède des boites à outils spécialisées. L'ensemble des fonctions de ces boites à outils facilitent beaucoup la simulation.

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

4.2 Evaluations expérimentales

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

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

a. Qualité des données d'apprentissage et de test

On commence avec une fréquence d'échantillonnage de 16 KHz. Puis Nous introduisons une dégradation sur les données à utiliser afin d'approcher la qualité du réseau téléphonique commuté RTC. La fréquence d'échantillonnage devient égale à 8KHz. Nous faisons varier le rapport signal sur bruit SNR de 10dB à 100dB

b. La dimension du vecteur de paramètres MFCC

Pour voir l'apport de la dimension du vecteur acoustique sur le taux d'identification, on va varier le nombre de coefficients MFCC de 4 à 40.

c. L'ordre du modèle

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

d. L'algorithme d'apprentissage

On va évaluer et comparer les résultats obtenus avec les deux algorithmes d'apprentissage : EM et LBG.

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

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

4.2.1 Les mélanges de gaussiennes standards (GMM)

4.2.1.1 La fréquence d'échantillonnage : 16 KHz

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

L	1	2	4	8	16	32	64
Ic%	94.11	95.48	96.59	96	100	100	100

TAB. 4.1 GMM - 16 KHz : Influence de l'ordre du modèle

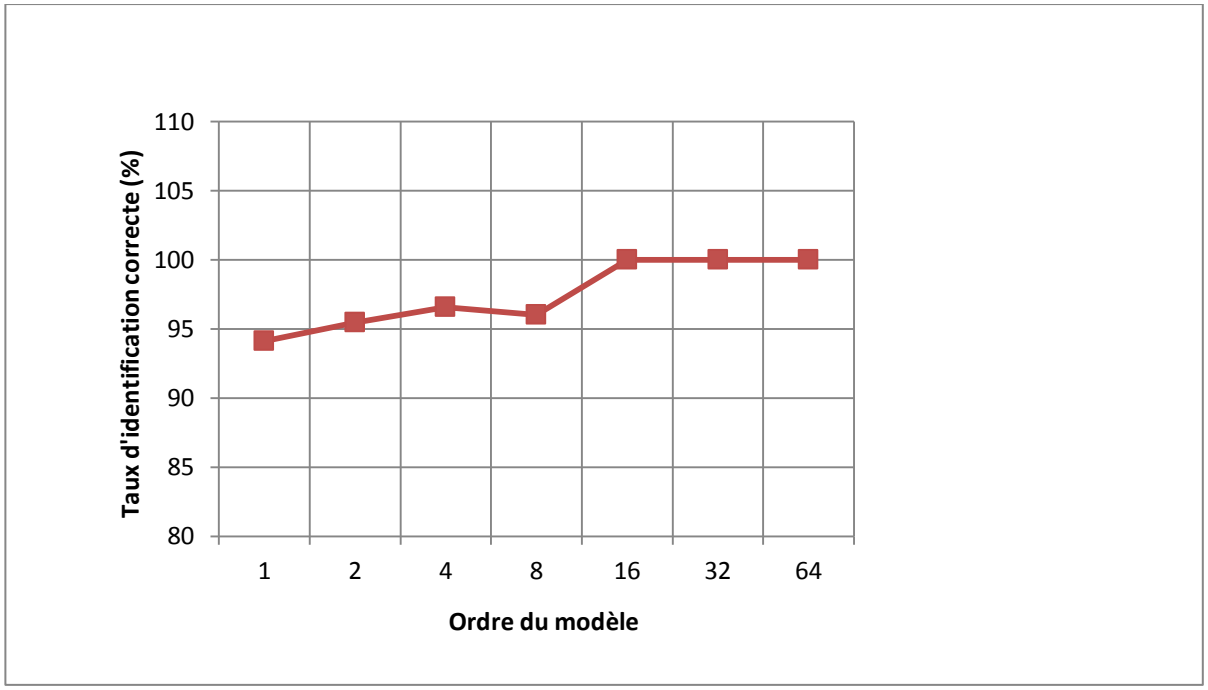


FIG. 4.5 GMM - 16 KHz : Influence de l'ordre du modèle

c**ommentaires et conclusions**

- Comme l'illustre la figure 4.5, l'ordre du modèle apporte une amélioration significative au taux d'identification correcte. Néanmoins, au delà de 16 gaussiennes où le taux d'identification atteint les 100 %, on remarque qu'un régime permanent s'établit.
- Pour ce qui concerne le nombre de locuteurs, le nombre de segments correctement identifiés diminue avec l'augmentation du nombre de locuteurs que le système doit identifier.

- L'augmentation de l'ordre du modèle permet d'affiner la séparation des classes acoustiques, ce qui se traduit par un accroissement du taux d'identification.
- On constate qu'avec une fréquence d'échantillonnage de 16 KHz, 16 composantes gaussiennes sont amplement suffisantes pour modéliser un locuteur.

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

P	4	8	12	16	20	24	28	32	36	40
Ic (%)	90.22	95.13	97.63	97.53	97.31	98	99.85	100	98.18	99.14

TAB. 4.2 GMM - 16 KHz : Influence du la dimension du vecteur acoustique

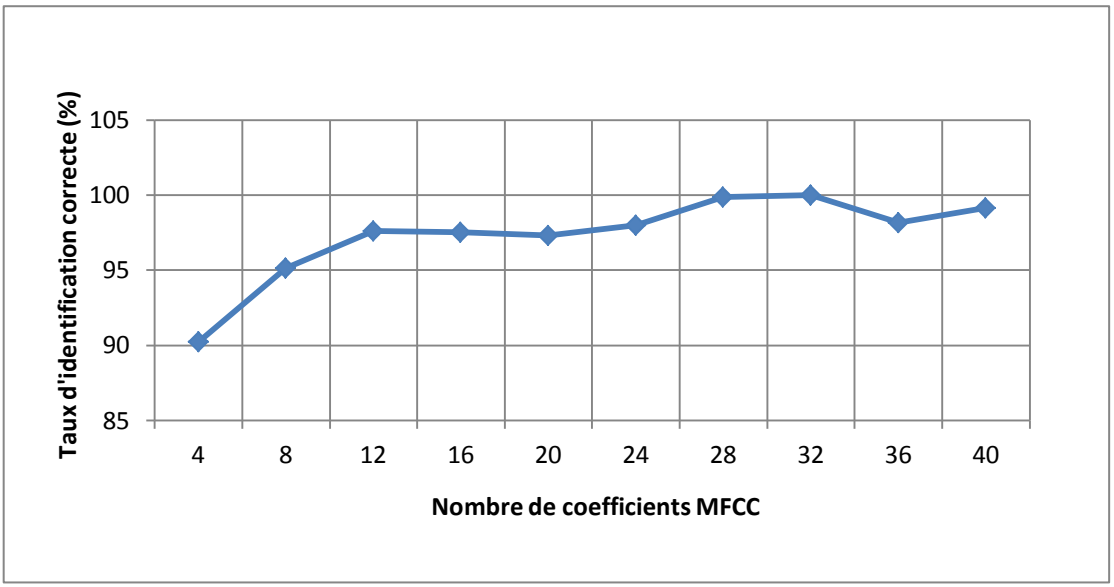


FIG. 4.6 GMM - 16 KHz : Influence de la dimension du vecteur acoustique

Commentaires et conclusions

- La figure 4.6 illustre l'augmentation du taux d'identification correcte avec le nombre de coefficients MFCC utilisé, avec un maximum entre 25 et 30 coefficients. La pente de la courbe est beaucoup plus importante entre 5 et 25 coefficients. Au delà de 25 coefficients MFCC, il n y a pas des grandes changements, donc pour optimiser le système on utilise 28 coefficient.
- On constate que la quasi-totalité de l'énergie du signal parole utilisé est contenue dans les 24 premiers MFCC, et les coefficients MFCC d'ordre supérieurs n'apportent pratiquement pas un plus d'information sur l'identité du locuteur.
- Pour une fréquence d'échantillonnage de 16 *KHz* , il faut utiliser entre 25 et 30 coefficients MFCC pour avoir de bons taux d'identification.

4.2.1.2 La fréquence d'échantillonnage : 8 KHz

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

a. Algorithme EM :

L	1	2	4	8	16	32	64
Ic%	73.45	79.38	82.68	91.49	91.94	99.81	97.43

TAB. 4.3 GMM - 8 KHz - EM : Influence de l'ordre du modèle

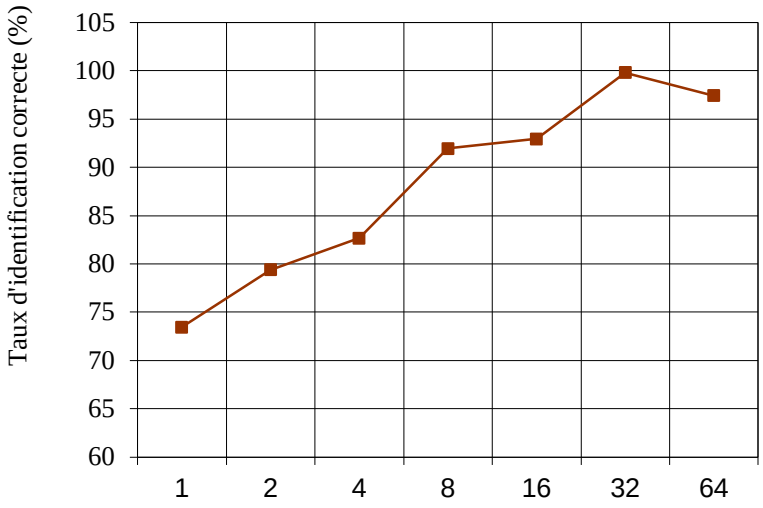


FIG. 4.7 GMM - 8 KHz - EM : Influence de l'ordre du modèle

Commentaires

Sur les courbes de la figure 4.7, on remarque que l'ordre du modèle améliore le taux d'identification correcte jusqu'à 32 gaussiennes, au delà de lesquelles le pourcentage d'identification diminue.

b. L'algorithme LBG

L	1	2	4	8	16	32	64
Ic%	65.94	74.88	86.25	89.81	95	97.23	96.13

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

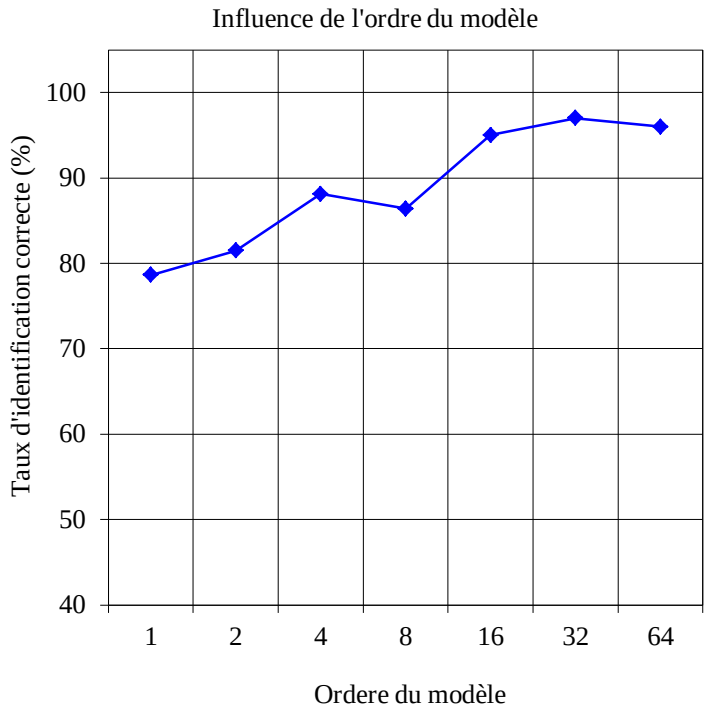


FIG. 4.8 GMM - 8 KHz - LBG : Influence de l'ordre du modèle

Commentaires

Sur les courbes de la figure 4.8, on remarque que l'ordre du modèle améliore le taux d'identification correcte, le maximum est atteint pour 32 gaussiennes.

Conclusions

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

Comparaisons

En comparant les taux d'identification obtenus par les deux algorithmes d'apprentissage utilisés, on a trouvé que l'algorithme de quantification vectorielle : LBG donne 97.41 % de la précision de l'algorithme EM. Cependant, cet algorithme nous a permis de diminuer de façon considérable les temps de calculs (phase d'apprentissage) et la complexité relativement par rapport à l'algorithme EM, ce qui le rend le choix idéal pour les applications grand public qui ne nécessitent pas un niveau de sécurité élevé.

Dans tout ce qui suit, l'apprentissage des modèles va se faire avec l'algorithme EM qui, d'après les résultats obtenus précédemment, constitue le cas le plus défavorable.

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

P	4	8	12	16	20	24	28	32	36	40
Ic (%)	74.14	81.95	89.87	96.61	94.38	95.94	97.81	93.17	97.80	96.9

TAB. 4.5 GMM - 8 KHz : Influence du la dimension du vecteur acoustique

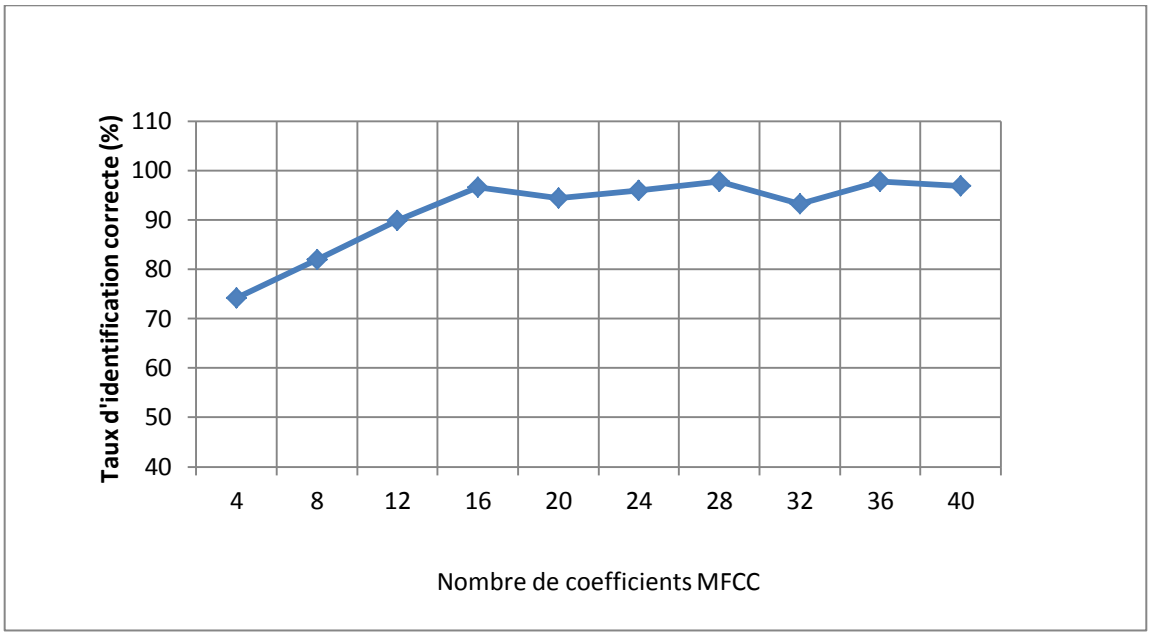


FIG. 4.9 GMM - 8 KHz : Influence de la dimension du vecteur acoustique

Commentaires et conclusions

- Sur la courbe de la figure 4.9, on peut remarquer que l'accroissement du nombre de coefficients apporte une amélioration sur le taux d'identification correcte. A partir du maximum atteint pour 36 coefficients MFCC, on remarque une certaine diminution du taux d'identification.
- On constate que la quasi-totalité de l'énergie du signal parole utilisé est contenue dans les 36 premiers MFCC, et les coefficients MFCC d'ordre supérieurs n'apportent pratiquement pas d'information sur l'identité du locuteur.
- Pour une fréquence d'échantillonnage de 8 KHz et un rapport signal sur bruit de 50 dB, on constate qu'il faut utiliser 36 coefficients MFCC pour avoir de bons taux d'identification.

4.2.1.2.3 Etude de l'influence du rapport signal sur bruit

Rapport SNR (dB)	10	20	30	40	50	60	70	80	90	100
Taux d'identification correcte (%)	4.56	15.94	63.06	88.44	92.94	93.75	96.69	95.88	96.50	97.19

TAB. 4.6 GMM - 8 KHz : Influence du rapport signal sur bruit

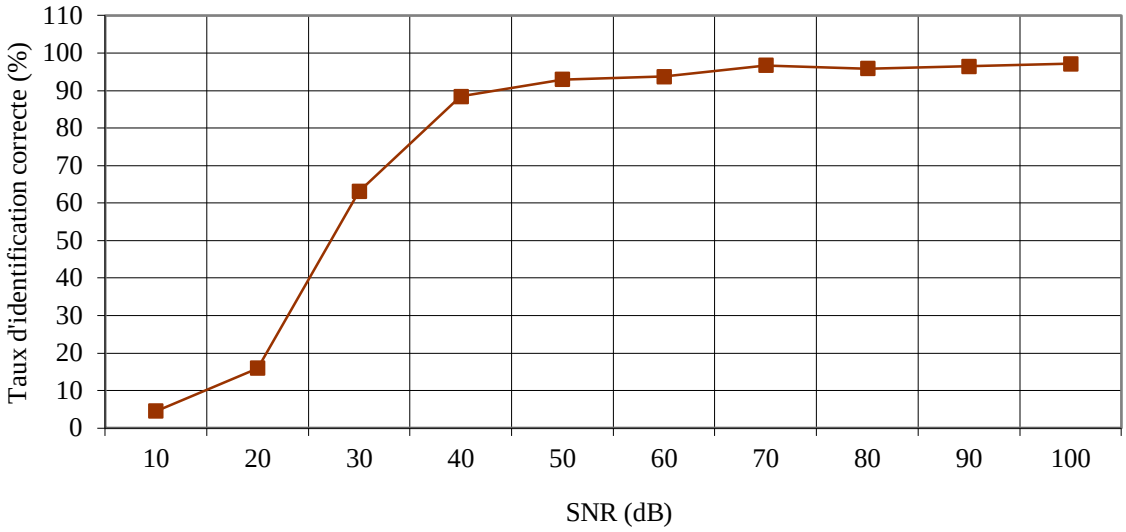


FIG. 4.10 GMM - 8 KHz : Influence du rapport signal sur bruit

Commentaires et conclusions

- La figure 4.10 montre que le taux d'identification augmente en améliorant la qualité des données, c'est-à-dire en augmentant le rapport signal sur bruit. La pente de la courbe est plus importante entre 10 et 45 dB, et au delà de 45 dB le régime permanent s'installe.

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

- Nécessite des quantités importantes de données d'apprentissage.
- Nécessite un nombre important de composantes gaussiennes, ce qui implique des temps de calcul très importants.

4.2.2 Quantification vectorielle QV

4.2.2.1 Influence du l'ordre du modèle

La fréquence d'échantillonnage est 16KHz. Nous faisons varier la taille L des dictionnaires en maintenant le nombre de coefficients à 12. Les résultats obtenus sont dans le tableau ci-dessous :

L	1	2	4	8	16	32	64
I_c%	74.5	74.5	79.8	87.7	89.5	92.1	92.1

Tab. 4.7 Influence du l'ordre du modèle QV

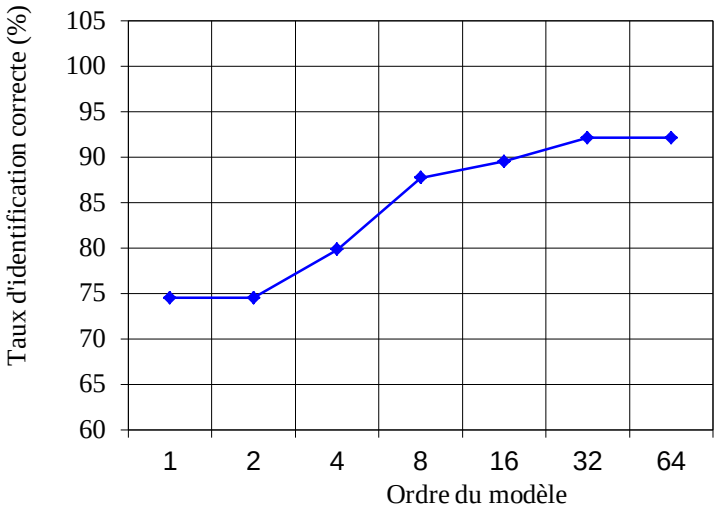


Fig.4.11 Influence du l'ordre du modèle QV.

D'après ces résultats, nous constatons que la quantification vectorielle est très performante , notant que les conditions d'enregistrement de la base de données ne sont pas assez bonnes (bruit ambiant, microphone utilisé, adaptation homme-machine, etc). Le taux d'identification

augmente avec l'ordre du modèle (figure 4.11). Il atteint sa valeur maximale à partir de L=32 et devient constant.

4.2.2.2Influence de la dimension des vecteurs acoustiques

La fréquence d'échantillonnage est 16KHz. Nous faisons varier la dimension P des vecteurs acoustiques en maintenant la taille des dictionnaires à 32. Les résultats obtenus sont dans le tableau ci-dessous :

P	4	8	12	16	20	24	28	32	40
I_c(MFCC)	84.2	89.5	92.1	92.1	92.1	92.1	92.1	92.1	93.0

Tab. 4.8 Influence de la dimension des vecteurs acoustiques QV.

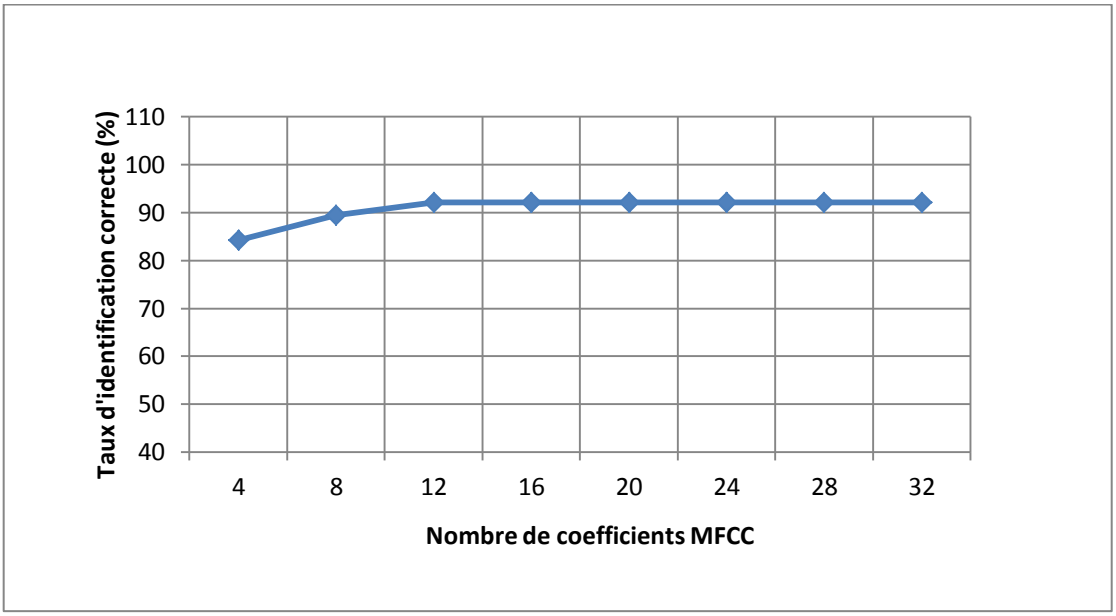


Fig. 4.12 Influence de la dimension des vecteurs acoustiques QV.

Le taux d'identification augmente avec le nombre de coefficients des vecteurs acoustiques (figure 4.12). Il devient pratiquement constant à partir de P=12. Les vecteurs acoustiques de dimension supérieure (P>12) n'apportent pas un plus d'informations remarquable sur l'identité des locuteurs

Conclusion

Dans un milieu non bruité (avec : Fe=16KHz), la quantification vectorielle est très performante. Cette performance s'améliore avec l'augmentation de L et P, mais les durées d'apprentissage et surtout du test risquent de devenir longues (augmentation du temps de calcul et de l'espace mémoire nécessaire pour le stockage des références). Pour remédier à ce problème, il faut ajuster les paramètres L et P de façon à garder une bonne performance, avoir une durée de test courte et réduire l'espace mémoire nécessaire.

4.2.2.3 Qualité des données d'apprentissage et du test

Nous introduisons une dégradation sur les données à utiliser. La fréquence d'échantillonnage devient égale à 8KHz. Nous faisons varier le rapport signal sur bruit SNR de 10dB à 100dB (nous travaillons avec les coefficients MFCC, L=32, P=12). Les résultats obtenus sont dans le tableau ci-dessous :

SNR	10	20	30	40	50	60	70	80	90	100
Ic	14.0	37.7	61.4	81.6	89.5	92.1	92.1	92.1	92.1	92.1

Tab. 4.9 Influence de la qualité des données QV.

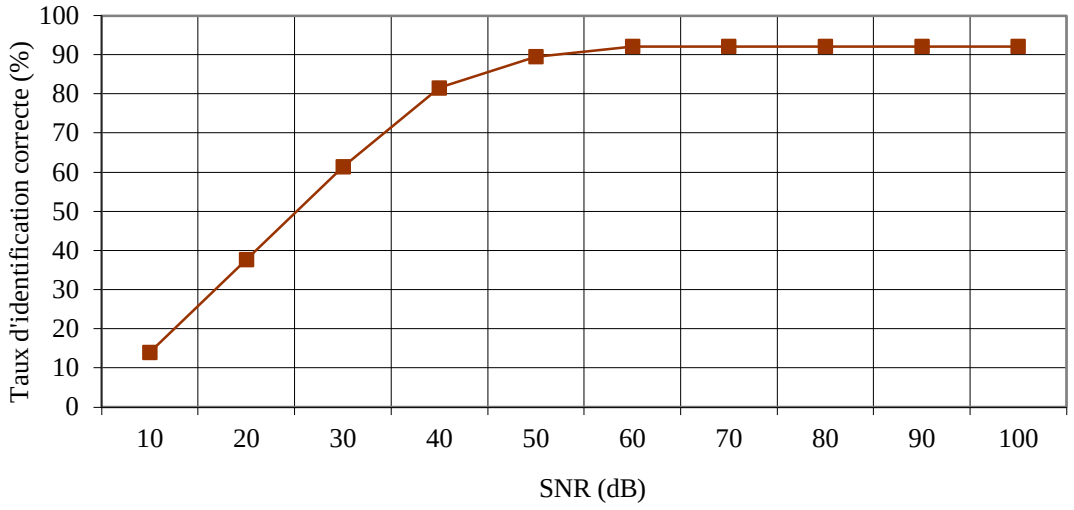


Fig. 4.13 Influence de la qualité des données QV.

D'après ces résultats (figure 4.13), la quantification vectorielle est sensible aux milieux fortement bruités (pour le RTC, la valeur typique du SNR est 40dB). La QV n'est pas robuste au bruit, cependant la performance du système s'améliore en améliorant la qualité des données. La QV est adaptée aux milieux faiblement bruités.

4.3 Etude comparative entre GMM et QV :

4.3.1 Influence de la dimension des vecteurs acoustiques sur Ic :

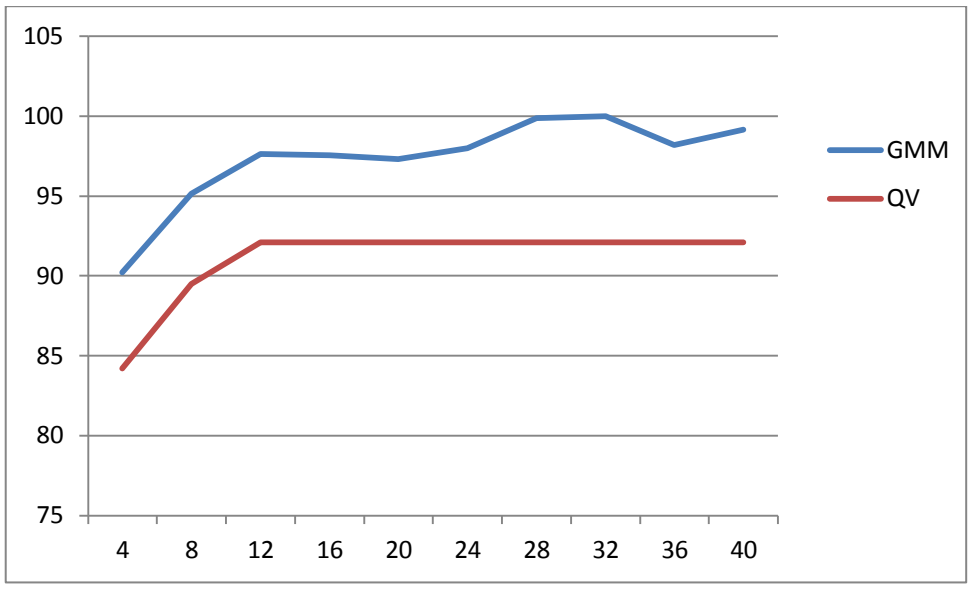


Fig. 4.14 Influence de la dimension des vecteurs acoustiques sur Ic.

On constate que la GMM offre des résultats largement meilleurs que ceux obtenus par la QV, mais la dernière atteint leur régime permanant plus rapide que la GMM.

4.3.2 Qualité des données d'apprentissage et du test

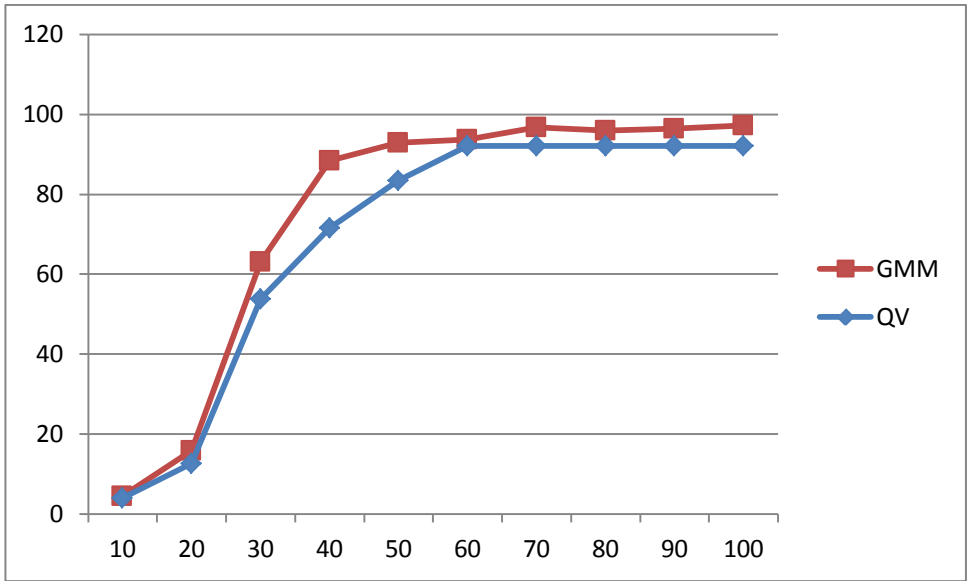


FIG 4.15 influence du SNR (dB) sur Ic%

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

Conclusions

La GMM

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

- Lorsqu' on dégrade la qualité des données de test et d'apprentissage, il faut augmenter le nombre de coefficients MFCC ainsi que l'ordre du modèle.
- Pour augmenter le nombre de composantes gaussiennes, il faut disposer de beaucoup de données d'apprentissage.

La QV

Dans un milieu non bruité (avec : Fe=16KHz), la quantification vectorielle est très performante. Cette performance s'améliore avec l'augmentation de L et P, mais les durées d'apprentissage et surtout du test risquent de devenir longues (augmentation du temps de calcul et de l'espace mémoire nécessaire pour le stockage des références). Pour remédier à ce problème, il faut ajuster les paramètres L et P de façon à garder une bonne performance, avoir une durée de test courte et réduire l'espace mémoire nécessaire.

Conclusion générale

Nous avons débuté cette thèse par un chapitre introductif expliquant les caractéristiques du signal parole ainsi que les outils utilisés pour effectuer la reconnaissance automatique du locuteur. Puis, nous avons décrit au chapitre 2 les paramètres qui caractérisent un locuteur. Nous proposons des solutions de base pour étudier quels sont les paramètres caractéristiques d'un locuteur qui définissent son identité. Puis, nous nous sommes intéressés à la modélisation des paramètres en utilisant deux approches: la GMM et la QV.

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

- La qualité et la quantité des données, ainsi que la taille du dictionnaire constituent le problème principal des systèmes d'identification du locuteur.
- La modélisation GMM fournit de bonnes performances. Néanmoins, elle nécessite beaucoup de données d'apprentissage, ce qui engendre des temps de calculs assez importants.
- La modélisation GMM c'est la meilleure solution pour les milieux fortement bruités.

L'introduction de l'algorithme de quantification vectorielle LBG pour l'apprentissage des modèles GMM permet une réduction significative des temps de calculs et la complexité relativement par rapport à l'algorithme EM avec une légère diminution des performances, ce qui le rend le choix idéal pour les applications grand public qui ne nécessitent pas un niveau de sécurité élevé.

- La QV :

D'après les résultats obtenus, la quantification vectorielle est sensible aux milieux fortement bruités (pour le RTC, la valeur typique du SNR est 40dB). La QV n'est pas robuste au bruit, cependant la performance du système s'améliore en améliorant la qualité des données. La QV est adaptée aux milieux faiblement bruités.

Cette performance s'améliore avec l'augmentation de L et P, mais les durées d'apprentissage et surtout du test risquent de devenir longues (augmentation du temps de calcul et de l'espace mémoire nécessaire pour le stockage des références). Pour remédier à ce problème, il faut ajuster les paramètres L et P de façon à garder une bonne performance, avoir une durée de test courte et réduire l'espace mémoire nécessaire.

Annexe A

1. Estimation du modèle autorégressif

1.1 Estimation du modèle AR de production de parole

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

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

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

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

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

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

alors on commis une erreur de prédiction :

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

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

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

1.1.1 Estimation des coefficients de prédiction

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

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

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

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

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

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

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

qui peut être mis sous forme matricielle suivante :

$$A \begin{bmatrix} a(1) \\ . \\ . \\ . \\ a(P) \end{bmatrix} = - \begin{bmatrix} \Phi_{xx}(1) \\ . \\ . \\ . \\ \Phi_{xx}(P) \end{bmatrix}$$

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

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

1.1.2 Algorithme de Levinson-Durbin

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

pour $\mathrm{m} = 1, 2, ..., \mathrm{p}$

pour $\mathrm{i} = 1, 2, ..., \mathrm{p}$

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

$\mathrm{a_i^{(m)}} = k_i$

pour $\mathrm{j} = 1, 2, ..., \mathrm{m}-1$

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

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

pour $\mathrm{j} = 1, 2, ..., \mathrm{p}$

$a_j = a_j^{(p)}$

1.1.3 Estimation du gain du modèle

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

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

On a aussi :

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

Ce qui implique :

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

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

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

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

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

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

2. Relation entre coefficients LPCC et les coefficients LPC

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

On peut en effet écrire :

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

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

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

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

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

On obtient donc la récurrence :

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

Annexe B

1 Distances et mesures de dissemblance dans l'espace acoustique

Il est possible d'utiliser toutes les distances classiques, en particulier les distances de Minkovski, parmi lesquelles la distance euclidienne, et la distance de Mahalanobis qui normalise les coefficients par leur matrice de covariance.

Dans un espace métrique, la distance entre deux vecteurs X et Y doit satisfaire les conditions suivantes :

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

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

La condition 2 peut être assurée en posant :

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

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

1.1 Distances usuelles

1. Distances de Minkovski

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

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

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

a. Distance euclidienne

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

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

b. Distance de Mahalanobis

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

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

et leur matrice de covariance se calcule par :

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

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

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

et permet de décorréler linéairement les coefficients.

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

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

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

BIBLIOGRAPHIE

[1] Lauwrence Rabiner et Bing-Hwang Juang. Fundamentals of speech recognition.signal processing. Prentice Hall, Englewood Cliffs, NJ, 1993.

[2] G. Doddington. Speaker recognition-identifying people from their voices. IEEE, 73(11):1651–1664, 1985.

[3] Mehdi Homayounpour. Vérification vocale d'identité Dépendante et indépendante du texte. PhD thesis, Université PARIS-SUD, centre d'Orsay, May 1995.

[4] A.E. Rosenberg, C.H. Lee, et S. Gokoen. Connected word talker verification using whole word hidden markov model. In ICASSP-91, pages 381–384, 1991.

[5] Gilbert Saporta. Probabilités, analyse des données et statistique, volume I. Editions Technip, 1990.

[6] J. Makhoul. Linear prediction: A tutorial review. Proceedings of the IEEE, 63(4):561–580, April 1975.

[7] Richard J. Mammone, Xiaoyu Zhang, et Ravi P. Ramachandran. Robust speaker recognition. IEEE signal processing magazine, pages 58–71, september 1996.

[8] S. Davis et P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Acoustics, Speech and Signal Processing 28(4), 357–366.

[9] Josef W. Picone. Signal modeling techniques in speech recognition. Proceedings of the IEEE, 81(9):1215–1247, September 1993.

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

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

[12] Augustine H. Gray et John D. Markel. Distance measure for speech processing. IEEE Trans. on acc. speech. and sig. proc.,ASSP, 24(5):380–391, October 1976.

[13] H. Sakoe et Chiba. Dynamic programing algorithm optimization for spoken word recognition. IEEE Trans. on ASSP, 26(1):43–49, 1978.

[14] C.S. Myers et L.R. Rabiner. Connected digit recognition using level building dtw algorithm. IEEE trans. on ASSP, 29(3):351–363, June 1981.

[15] L.L. Scharf. Statistical Signal Processing. Detection, Estimation and Time Series Analysis. Addison-Wesley Publishing Company, 1991.

[16] Guillaume Gravier. Vérification du locuteur par modèles de markov cachés gauche-droite. Rapport de stage dea, IDIAP, CH-1920 Martigny, 1995.

[17] T. Matsui et S. Furui. Likelihood normalization for speaker verification using a phoneme and speaker-independent model. Speech Communication, Elsevier, 17:109– 116, 1995.

[18] Frédéric Bimbot, M. Blomberg, et al. Sv algorithms improvements and evaluation, deliverable 4.2. Technical report, Telematics European Project LE-1930: Caller Verification in Banking and Telecommunications (CAVE), 1997.

[19] Douglas A. Reynolds. A Gaussian mixture modeling approach to textindependent speaker identification. PhD thesis, Georgia Institute of Technology, 1992.

[20] Younès Bennani et Patrick Gallinari. Connexionist approaches for automatic speaker recognition. In ESCA [1994], pages 95–102.

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

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

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

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

[25] L.RABINER, B-H.JUANG, "Fundamentals of speech recognition", Prentice Hall, New Jersey, 1993.

[26] G.BLANCHET, M.CHARBIT, "Signaux et images sous Matlab", HERMES science publications, Paris, 2001.