وزارة التعليم العالي و البحث العلمي

BADJI MOKHTAR – ANNABA UNIVERSITY
UNIVERSITE BADJI MOKHTAR - ANNABA

جامعة باجي مختار – عنابة

Faculté des SCIENCES de L'INGENIEUR Année: 2006-2007

Département d'INFORMATIQUE

THESE

Présentée en vue de l'obtention du diplôme de DOCTORAT

Système Neuro- Markovien basé sur la fusion de données floues et génétiques : Application pour la Reconnaissance automatique de la parole

Option
INTELLIGENCE ARTIFICIELLE

Par
Lilia LAZLI-BOUKHALFA

DIRECTEUR DE THESE Mohamed Tayeb LASKRI PR Univ. Annaba

DEVANT LE JURY

PRESIDENT	Zaïdi SAHNOUN	PR	Univ. Constantine
EXAMINATEURS	Mouloud KOUDIL	MC	I.N.I. Alger
	Mohamed El Bachir MENAI	MC	C.U. Tébessa
	Hayette MEROUANI	MC	Univ. Annaba

Résumé

L'Intelligence Artificielle (IA) continue de connaître des développements importants dans le domaine de la modélisation des processus cognitifs. Un des axes intéressant de ces développements est l'orientation vers les approches hybrides qui incorporent plusieurs paradigmes dans un même système.

Les modèles de Markov cachés (HMM – Hidden Markov Models) sont de nos jours l'approche la plus utilisée en reconnaissance de la parole. Pourtant, ils souffrent de nombreuses hypothèses contraignantes parmi lesquelles le fait que les vecteurs acoustiques sont supposés non corrélés ou encore l'hypothèse sur la distribution des densités de probabilités de chaque état HMM (distributions discrètes ou multi-gaussiennes). Ces hypothèses peuvent être contournées en utilisant un perceptron multi-couches (MLP – Multi Layer Perceptrons). Ce réseau de neurones estime les probabilités a posteriori utilisées par chaque état des HMM. Les modèles hybrides HMM/MLP ont déjà été utilisés avec succès pour l'Anglais britannique et américain aussi bien que pour le Français. Dans ce travail, nous rapportons les résultats obtenus des modèles hybrides HMM/MLP utilisés pour l'entraînement et la reconnaissance de la parole Arabe isolée sur une base personnelle qui permet de rejoindre les capacités discriminantes, la résistance au bruit des MLP et la souplesse des HMM afin d'obtenir de meilleures performances que les HMM classiques. De très bons résultats du système hybride ont été obtenus par rapport à un reconnaisseur classique utilisant les HMM discrets.
De nombreuses expériences ont déjà montré qu'une forte amélioration du taux de reconnaissance des systèmes HMM (Modèles de Markov cachés) traditionnels est observée lorsque plus de données d'entraînement sont utilisées. En revanche, l'augmentation du nombre de données d'entraînement pour les modèles hybrides HMM/ANN (Modèles de Markov cachés/Réseaux de neurones artificiels) s'accompagne d'une forte augmentation du temps nécessaire à l'entraînement des modèles mais pas ou peu des performances du système. Pour pallier cette limitation, nous rapportons dans ce travail les résultats obtenus avec une nouvelle méthode d'entraînement basée sur la fusion de données. Cette méthode a été appliquée dans un système de reconnaissance de la parole arabe. Ce dernier est basé d'une part, sur une segmentation floue (application de l'algorithme c-moyennes floues) et d'une autre part, sur une segmentation à base des algorithmes génétiques. L'intégration de ces algorithmes dans le reconnaisseur hybride a offert une amélioration significative de la performance du système hybride proposé.

Mots clefs - Reconnaissance de la parole arabe, segmentation floue, algorithmes génétiques, modèles de Markov cachés, réseaux de neurones artificiels, méthode de fusion de données.

Dédicace

Je dédie ce mémoire de DOCTORAT
à ma petite fille chérie MELISSA

Remerciements

Je voudrais exprimer ma profonde gratitude envers dieu tout puissant qui grâce à son aide, j'ai pu finir ce modeste travail.

Je tiens tout d'abord à remercier le Professeur Zaïdi SAHNOUN, le Dr. Mouloud KOUDIL, le Dr. Mohamed El Bachir MENAI et le Dr. Hayette MEROUANI pour l'intérêt qu'ils portent à mon travail en acceptant de participer à mon jury.

Je remercie le Professeur Mohamed Tayeb LASKRI qui m'a accueilli au sein de son groupe et qui a accepté la lourde tâche de rapporter sur mes travaux de recherche. Qu'il trouve dans cette thèse ma profonde gratitude et mon grand respect.

Je tiens également à exprimer mes plus vifs remerciements à mes parents, pour leur soutien moral indéfectible, leurs encouragements et pour leur précieuse collaboration depuis de longues années et sans lesquels ce travail n'aurait pas été possible, qu'ils trouvent dans cette thèse ma profonde gratitude.

Je n'oublie pas non plus ma soeur, parti bien trop tôt, mes frères et ma famille…

Je remercie chaleureusement mon mari Sans leur soutien dans les moments difficiles, je n'en serai pas là où j'en suis.

Enfin, merci à ma petite fille Mélissa, mon rayon de soleil, qui m'a souvent redonné le sourire et la rage de vaincre.

Table des matières

Partie II – Conception & Réalisation

4 Nouveaux algorithmes de partitionnement

PARTIE I

Etat de l'Art

Chapitre 1

Introduction générale

«Nous pouvons définir le son comme un coup donné par l'air à travers les oreilles au cerveau et au sang, et arrivant jusqu'à l'âme. Le mouvement qui s'ensuit, lequel commence à la tête et se termine dans la région du foie, est l'ouïe. »

Platon, 300 avant J.-C.

1 INTRODUCTION

La reconnaissance automatique de la parole par les machines est depuis longtemps un thème de recherche qui fascine le public, mais qui demeure un défi pour les spécialistes. Pour le grand public, un archétype de la communication homme-machine reste probablement le dialogue avec l'ordinateur HAL mis en scène par *Stanley Kubrick* dans le film *2001, l'Odyssée de l'Espace* (1968) : le dialogue en langage naturel avec une machine aussi intelligente que l'homme semble l'aboutissement normal des progrès technologiques. Dans le domaine de la recherche, il était possible d'imaginer à l'époque des progrès rapides des systèmes de Reconnaissance Automatique de la Parole (RAP), en dépit de l'opinion sceptique de quelques spécialistes [BRI82]. Mais le projet de *« compréhension de la parole »* lancé en 1971 par le département de la défense américaine (ARPA Speech Understanding Project) [BAU72] a contraint les chercheurs à tempérer leur optimisme : malgré des directions prometteuses, le sujet nécessite un effort à long terme.

Aujourd'hui, l'impact des systèmes de RAP est encore minime dans la vie courante, et la commande des ordinateurs ne s'effectue toujours pas par la voix, malgré les promesses de fabricants de logiciels ou de matériels informatique (Microsoft, Apple). L'annonce de la commercialisation du système de dictée vocal d'IBM pour les ordinateurs PC en 1994 a suscité de l'intérêt, mais aussi des réserves quant aux performances actuelles du système [Le Monde, 1994]. Pourtant les progrès réalisés depuis 25 ans en RAP sont très importants, grâce à un grand nombre de recherches traitant du problème sous tous ses aspects. Les limitations de la capacité des systèmes de reconnaissance, imposées à l'origine par la complexité de la tâche, sont progressivement repoussées, et des systèmes efficaces pour des applications spécialisées sont maintenant disponibles et commercialisés.

La reconnaissance automatique de la parole est un domaine d'étude actif depuis le début des années 50. Il est clair qu'un outil de reconnaissance de la parole efficace faciliterait l'interaction entre les êtres humains et les machines. Les applications possibles associées à un tel outil sont nombreuses et sont amenées à connaître un essor fabuleux dans les prochaines années. La plupart des applications en reconnaissance de la parole peuvent être regroupées en 4 catégories :

1. Commande et contrôle

   Contrôler à l'aide de la parole des équipements particuliers (machines, robots...) ou des programmes (ouvrir des fenêtres ou naviguer sous Windows, aide aux personnes handicapées).

2. Accès à des bases de données ou recherche d'informations

   Composition automatique de numéros de téléphone, serveurs vocaux, réservation d'un vol, guidage automatique (dans une voiture), remplir un questionnaire, effectuer un audit qualité...

3. Dictée vocale

Création de lettres, rapports et autres documents par l'intermédiaire de la parole.

4. Transcription automatique de la parole

Indexation de programmes télévision ou radio, sous-titrage et traduction automatique...

De nombreux progrès ont été réalisés ces dix dernières années dans ce domaine. Il existe d'ailleurs des logiciels vendus actuellement dans le commerce se vantant d'effectuer une reconnaissance de la parole continue pour un vocabulaire important. Néanmoins, les performances de ces systèmes sont encore largement inférieures à celles des êtres humains. Même si les progrès réalisés en moins de 50 ans sont énormes, il reste de nombreux problèmes à résoudre.

Les capacités d'un système de reconnaissance de la parole et la difficulté de son développement peuvent être mesurées en se posant quatre questions sur le système correspondant à quatre classes de variabilité du signal de parole :

- Qui peut utiliser le système ? N'importe quel locuteur (indépendamment du sexe, de l'âge...) peut-il parler à la machine avec des chances d'être compris ?

- Comment doit-on parler au système ? Quel est le mode d'élocution permis ?

- Que reconnaît le système ? Quel est le vocabulaire permis ?

- Quel est le taux de reconnaissance ? Autrement dit, quel est le pourcentage de mots reconnus ?

Les spécialistes de la reconnaissance de la parole répondent à ces questions en utilisant un vocabulaire particulier qu'il est utile de rappeler.

2 MODE DE FONCTIONNEMENT

Un système de reconnaissance peut être utilisé sous plusieurs modes :

- Dépendant du locuteur (monolocuteur)

Dans ce cas particulier, le système de reconnaissance a été configuré pour un locuteur spécifique[1].

1. C'est le cas de la plupart des logiciels de reconnaissance de la parole vendus sur le marché actuellement. Les principaux systèmes de dictée vocale actuels possèdent une phase d'entraînement recommandée avant toute utilisation (voire même une adaptation continue des paramètres au cours de l'utilisation du logiciel) afin d'effectuer une adaptation des paramètres à la voix de l'utilisateur.

- Pluri-locuteur (multi-locuteur)

Le système de reconnaissance a été élaboré pour un groupe restreint de personnes. Le passage d'un locuteur à un autre du même groupe se fait sans adaptation.

- Indépendant du locuteur

Tout locuteur peut utiliser le système de reconnaissance.

Il est évident que plus le nombre de locuteurs est élevé, plus l'adaptation du système de reconnaissance est difficile.

3 ELOCUTION

Le mode d'élocution caractérise la façon dont on peut parler au système. Il existe quatre modes d'élocution distincts.

- Mots isolés

Chaque mot doit être prononcé isolément, c'est à dire précédé et suivi d'une pause.

- Mots connectés

Le système reconnaît des séquences de quelques mots sans pause volontaire pour les séparer (exemple : Reconnaissance de chiffres connectés ou de nombres quelconques...).

- Parole continue lue

C'est le discours usuel, si ce n'est que les textes sont lus.

- Parole continue spontanée

C'est le discours usuel, sans aucune contrainte.

La reconnaissance de mots isolés fonctionne relativement bien de nos jours. De très bons résultats ont été publiés par de nombreux laboratoires. Généralement, de tels outils de reconnaissance de la parole sont utilisés pour un vocabulaire de commande correspondant à des actions simples (gestion de menus...).

Le mode « mots connectés » est utile dans des cas particuliers comme la reconnaissance de nombres quelconques. Les laboratoires de recherche ont publié de bons résultats pour un vocabulaire restreint. Le mode « parole continue » est naturel et beaucoup plus compliqué. Quelques laboratoires ont publié des premiers résultats impressionnant pour ce mode d'élocution avec un vocabulaire d'environ 65000 mots, mais il n'existe pas encore de système avec des performances comparables à celles des humains. La parole spontanée est beaucoup plus difficile à traiter, les dernières études rapportent des taux d'erreur de l'ordre de 50%.

4 VOCABULAIRE ET SYNTAXE

Le vocabulaire est l'ensemble des mots que le système est capable de reconnaître. La taille du vocabulaire peut varier de quelques mots à plusieurs dizaines de milliers de mots. Il est évident que plus le vocabulaire est grand plus le risque de se tromper augmente et donc moins le système est performant.

La syntaxe spécifie plutôt les contraintes imposées sur les suites de mots prononcés. Elle peut être inexistante (tout mot peut suivre n'importe quel autre mot), ou bien contraignante (après un mot donné seuls certains autres sont autorisés). L'utilisation d'une syntaxe facilite la tâche du système de reconnaissance en limitant le nombre de mots et d'hypothèses à traiter.

5 TAUX DE PERFORMANCE

Le taux de performance ou taux de reconnaissance permet de mesurer l'efficacité du système de reconnaissance testé. Ce taux varie fortement selon le type de canal de transmission utilisé (microphone, téléphone), la taille du vocabulaire, et le type d'élocution. Il existe différentes valeurs mesurant les performances d'un système de reconnaissance :

- Taux de reconnaissance : Pourcentage de mots ou de phrases reconnus correctement.

- Taux d'erreur : 1 - taux de reconnaissance (noté T.E. dans le reste du document).

- Taux de substitution : Pourcentage de mots pour lesquels le système a commis une erreur (noté Subs. Dans le reste du document).

- Taux de rejet : Pourcentage de mots que le système n'a pas compris.

- Taux d'omission : Pourcentage de mots non détectés (noté Sup. dans le reste du document).

- Taux d'insertion : Pourcentage de mots reconnus alors qu'aucun mot n'a été prononcé (noté Ins. Dans le reste du document).

- Réduction du taux d'erreur = $\frac{ancienT.E. - nouveauT.E.}{ancienT.E}$ (noté R.T.E. dans le reste du document). Ce taux permet de mesurer l'influence d'une méthode testée sur le système de reconnaissance.

6 OBJET DE LA THESE

Aujourd'hui, la reconnaissance automatique de la parole est un domaine de recherche bien établi, à partir duquel émergent des technologies permettant le développement d'applications réelles.

Le problème de la reconnaissance automatique de la parole consiste à extraire à l'aide d'un ordinateur l'information contenue dans le signal de parole. La technologie la plus utilisée depuis ces 20 dernières années est basée sur les modèles statistiques (modèle de Markov cachés – HMM) capables de modéliser simultanément les caractéristiques fréquentielles et temporelles du signal étudié. Dernièrement, une extension de ces modèles a été mise au point donnant naissance aux modèles hybrides. Ces derniers combinent la technologie des modèles de Markov cachés (HMM) et des réseaux de neurones artificiels (ANN: Artificial Neural Network) particulièrement les perceptrons multi-couches (MLP). De nombreuses publications [ANA95] ; [ATA91] ; [AUB93] ; [AVE87] ; [BEL57] ; [BAK74] ; [BAK75] ; [BAK76] ont déjà montré l'efficacité de ces modèles en reconnaissance de la parole (continue ou isolée) indépendamment du locuteur pour de petits et grands vocabulaires.

Notre étude s'intègre dans le cadre du développement d'un système de reconnaissance vocal indépendant de locuteur pour la langue arabe, néanmoins que le traitement de la parole arabe est encore à ses débuts auquel nous espérons apporter une contribution à travers ce travail. La reconnaissance par les méthodes les plus performantes de l'état de l'art reste insuffisante ; cette faiblesse est un facteur limitant des systèmes de RAP. Nous cherchons à améliorer les performances des systèmes de RAP en mettant comme objectif l'augmentation du taux de reconnaissance, appliquant ainsi une nouvelle technique qui combine efficacement les HMM et les MLP ayant comme principal but d'éliminer les limites et les faiblesses soulevées par chacun d'eux. Cette hybridation se base sur la complémentarité qui peut exister entre les HMM et les MLP et tente ainsi de les faire coopérer pour un système de reconnaissance vocal, où chacun d'entre eux prendra à sa charge le traitement d'une tâche qui s'accommode le mieux avec son style de raisonnement et où chacun d'entre eux viendrait pallier les inconvénients de l'autre sans pour autant en renforcer les faiblesses.

Les modèles de Markov cachés (HMM) qui se sont imposés comme l'un des modèles de référence pour leur lisibilité et surtout pour la simplicité et l'efficacité de ses principaux algorithmes, ont aussi des faiblesses qui concernent la modélisation de la durée inhérente à leur topologie, le coût en temps de calcul et en mémoire, ainsi que la mise en œuvre de ces modèles nécessite des hypothèses contraignantes.

Le système proposé permet de rejoindre les capacités discriminantes, la résistance au bruit des MLP et la souplesses des HMM afin d'obtenir de meilleures performances que les HMM classiques. Le MLP proposé permet d'évaluer les probabilités d'observation pour les états HMM qui seront entraînés par la suite selon une méthode itérative de type EM (Expectation - Maximization). Une attention particulière a été portée sur la construction de la chaîne de Markov et la sélection des états pertinents.

Le traitement de la parole arabe est encore à ses débuts, la raison pour laquelle, nous avons pensé à l'application des modèles hybrides HMM\ANN pour des bases de données allant des petits lexiques aux moyens lexiques, ayant comme objectif la reconnaissance de la parole indépendante du locuteur [LAZb02] ; [LAZc02]. Nous avons ainsi utilisé dans nos expériences un MLP pour estimer les probabilités a posteriori utilisées pour chaque état du HMM. Pour augmenter le taux de reconnaissance, et pour des fins d'une segmentation acoustique, nous avons proposé deux nouveaux algorithmes : (1) le premier repose sur des concepts de la logique floue : l'algorithme C-moyennes floues (FCM – Fuzzy C-Means) [LAZg03] ; [LAZh03] ; [LAZe03]. Nous avons pensé à cet algorithme, vu que l'algorithme classique c-moyennes (k-means en anglais) fournit des distributions discrètes assez dures, non probabilisées, qui ne transmet pas assez d'informations sur les observations discrètes. En revanche, l'algorithme FCM proposé permet de classer les données acoustiques en diverses classes selon un degré d'appartenance floue. (2) concernant le deuxième algorithme, nous nous sommes intéressés plus particulièrement, aux méthodes impliquant une classification supervisée par partition et nous avons retenu une comme base pour notre travail. Cette solution consiste à faire le choix d'une mesure que nous utilisons dans notre application. Cet algorithme cherche une "bonne" partition relativement à un critère qui mesure la qualité d'une partition. Nous sommes donc ramenés à un problème d'optimisation. Les propriétés de cet algorithme ne garantissent pas la convergence vers un optimum global, c'est pourquoi nous nous sommes intéressés à une heuristique de type Algorithmes Génétiques (AG), moins susceptibles d'être piégés par les minima locaux et désormais largement employés dans les problèmes d'optimisation.. Si sur un plan théorique, aucun résultat général ne prouve que cette méthode conduise à une solution optimale, en pratique la convergence globale est souvent constatée. Des expériences préliminaires au niveau du mot, utilisant des vocabulaires de tailles 1200 et 3900 mots sont rapportées. Nous comparons les résultats du système proposé avec ceux d'un système classique utilisant les HMM standards.

Nous avons constaté que ces modèles hybrides HMM\MLP souffrent de nombreux défauts parmi lesquelles le fait que le nombre de paramètres est en quelque sorte borné. En effet, aucune amélioration n'est généralement observée (comme habituellement pour les HMM continus) lorsque le nombre des données d'entraînement et/ou de paramètres est fortement augmenté. Pour pallier cette limitation des systèmes hybrides HMM/MLP, nous proposons dans cette thèse une nouvelle méthode visant à explorer ce problème. Cette méthode est basée sur des expériences qui ont déjà montré qu'il est possible d'améliorer sensiblement les performances des systèmes hybrides en combinant plusieurs modèles [LAZa04] ; [LAZb04]. A la base, l'hypothèse est que, si les modèles sont entraînés sur différentes parties du fichier d'entraînement, ils vont sélectionner des propriétés différentes des données, permettant ainsi une amélioration des résultats lorsque les sorties sont combinées. D'ailleurs c'est un peu dans cette optique que certains laboratoires travaillant en reconnaissance de la parole entraînent des modèles pour les hommes et pour les femmes. Lors de la reconnaissance, les modèles sont tous les deux utilisés et la sortie correspondant au meilleur score est sélectionnée [GAU94]. Ces modèles sont combinés selon plusieurs critères pour fournir le mot le plus probable.

7 STRUCTURE DE LA THESE

Après les quelques rappels annoncés ci-dessus, nous présentons dans le second chapitre d'une manière concise, mais abondante, le mécanisme de production et d'audition de la parole ainsi que les caractéristiques acoustiques et phonétiques des sons de la parole. Nous mettons ensuite, en évidence les principales difficultés rencontrées lors de sa modélisation en donnant un exemple d'un modèle Auto – Régressif pour la phonation.

Dans le troisième chapitre, nous décrivons l'architecture générale du système de reconnaissance de la parole et les différents éléments qui le composent. Nous mettons ensuite en œuvre les difficultés rencontrées pour la mise au point des systèmes de RAP. Les méthodes classiques employées en reconnaissance de la parole seront décrites par la suite. Nous résumerons ensuite, les principales techniques d'analyse pour l'extraction des paramètres acoustiques effectuée sur le signal vocal avant le processus de reconnaissance. Nous décrivons par la suite, les modèles les plus utilisés en reconnaissance de la parole : les modèles de Markov cachés (HMM), et nous terminerons par la description de quelques systèmes de reconnaissance de la parole isolée à grand vocabulaire.

Dans le quatrième chapitre, nous rappelons tout d'abord ce qu'est une classification, et sur quoi elle travaille, Nous présentons ensuite cinq propositions permettant de traiter d'une façon générale les données hétérogènes, puis nous limitons notre attention à un type particulier d'algorithmes : Algorithmes de partitionnement dont le principe sera décrit, ce qui nous permet d'introduire l'intérêt des algorithmes proposés dans le cadre d'une segmentation de la parole. Nous esquissons le cadre général du fonctionnement qui repose sur les concepts de la logique floue ainsi que le principe des algorithmes génétiques, nous montrerons éventuellement, les avantages de ces algorithmes par rapport à d'autres classiques utilisés pour les mêmes fins.

Le cinquième chapitre est consacré à la présentation du système hybride HMM/MLP développé pour la reconnaissance de la parole arabe. En premier, nous présentons les notions de base des réseaux de neurones artificiels : leur structure, leur apprentissage ainsi que leurs liens profonds avec les algorithmes stochastiques utilisés en traitement du signal. Nous montrerons par la suite que les modèles hybrides combinant la technologie des HMM et des MLP permettent de résoudre certains problèmes liés à l'utilisation des HMM. Nous montrerons que les taux de reconnaissance obtenus avec ces modèles hybrides sont aux moins aussi performants et même supérieurs aux modèles HMM classiques décrits dans le troisième chapitre en se basant sur les résultats obtenus et sur la description de quelques travaux qui ont montré l'avantage de combiner les réseaux de neurones et les modèles de Markov cachés en reconnaissance de la parole (isolée ou continue) indépendamment du locuteur pour de petits ou grands vocabulaires. Ils seront à la base de la plupart des expériences menées dans ce travail.

Nous définissons dans le sixième chapitre, une nouvelle méthode permettant de diviser en plusieurs parties l'ensemble d'entraînement et d'entraîner plusieurs MLP sur chacune de ces parties. Nous espérons ainsi tirer profit de l'entraînement des réseaux sur des données filtrées par la procédure de fusion mettant en exergue des propriétés différentes du signal. Différents types de combinaisons des systèmes ont été testés dans ce chapitre.

Dans le septième chapitre, nous résumerons en conclusion les différentes méthodes testées dans ce travail, les améliorations qu'elles ont engendrées et leurs possibles extensions

Chapitre 2

Généralités sur le Traitement de la parole

Le chapitre 2 présente un exposé condensé mais rigoureux au sujet des mécanismes indispensables à la bonne compréhension du traitement de la parole. Il s'agit de faire un rappel théorique sur les caractéristiques acoustiques, phonétiques et auditives de la parole. En se basera surtout sur l'aspect acoustique pour présenter la modélisation Auto-Regressive de l'organe phonatoire en l'éclaircissant par un modèle simplifié Auto-Regressif pour la phonation.

1 INTRODUCTION

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

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

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

Aucun de ces signaux, pourtant fort complexes, n'est cependant à la fois appelé à être *produit* et *perçu* instantanément par le cerveau, comme c'est le cas pour la parole. La parole est en effet produite par le conduit vocal, contrôlé en permanence par le cortex moteur. L'étude des mécanismes de phonation permettra donc de déterminer, dans une certaine mesure, ce qui est parole et ce qui n'en est pas. De même, l'étude des mécanismes d'audition et des propriétés perceptuelles qui s'y rattachent permettra de dire ce qui, dans le signal de parole, est réellement perçu. Mais l'essence même du signal de parole ne peut être cernée de façon réaliste que dans la mesure où l'on imagine, bien au-delà de la simple mise en commun des propriétés de production et de perception de la parole, les propriétés du signal dues à la mise en boucle de ces deux fonctions. Mieux encore, c'est non seulement la perception de la parole qui vient influer sur sa production par le biais de ce bouclage, mais aussi et surtout sa compréhension[1]. On ne parle que dans la mesure ou l'on s'entend et où l'on se comprend soi-même; la complexité du signal qui en résulte s'en ressent forcément[2].

S'il n'est pas en principe de parole sans cerveau humain pour la produire, l'entendre, et la comprendre, les techniques modernes de traitement de la parole tendent cependant à produire des systèmes automatiques qui se substituent à l'une ou l'autre de ces fonctions.

- Les analyseurs sont utilisés soit comme composant de base de systèmes de codage, de reconnaissance ou de synthèse (voir ci-dessous), soit en tant que tels pour des applications spécialisées, comme l'aide au diagnostic médical (pour les pathologies du larynx, par analyse du signal vocal) ou l'étude des langues.

---

1 Ce qui accroît encore la différence entre la parole et, par exemple, l'image : alors que la compréhension de l'image est en principe également accessible à tous, la compréhension de la parole est le résultat d'un apprentissage socioculturel lié à une communauté linguistique.

2 A cet égard, la discipline scientifique qui s'apparente le plus au traitement de la parole est sans doute le traitement automatique des caractères manuscrits.

- Les *analyseurs* de parole cherchent à mettre en évidence les caractéristiques du signal vocal tel qu'il est produit ou parfois tel qu'il est perçu (on parle alors d'*analyseur perceptuel*), mais jamais tel qu'il est compris, ce rôle étant réservé aux reconnaisseurs.

- Les *reconnaisseurs* ont pour mission de décoder l'information portée par le signal vocal à partir des données fournies par l'analyse. On distingue fondamentalement deux types de reconnaissance, en fonction de l'information que l'on cherche à extraire du signal vocal : la *reconnaissance du locuteur*, dont l'objectif est de reconnaître la personne qui parle, et la *reconnaissance de la parole*, où l'on s'attache plutôt à reconnaître ce qui est dit. On classe également les reconnaisseurs en fonction des hypothèses simplificatrices sous lesquelles ils sont appelés à fonctionner. Ainsi :

  - En reconnaissance du locuteur, on fait la différence entre *l'identification* et *la vérification* du locuteur, selon que le problème est de vérifier que la voix analysée correspond bien à la personne qui est sensée la produire ou qu'il s'agit de déterminer qui, parmi un nombre fini et préétabli de locuteurs, a produit le signal analysé.

  - On sépare reconnaissance du locuteur *dépendante du texte*, reconnaissance *avec texte dicté*, et reconnaissance *indépendante du texte*. Dans le premier cas, la phrase à prononcer pour être reconnue est fixée dès la conception du système; elle est fixée lors du test dans le deuxième cas, et n'est pas précisée dans le troisième.

  - On parle de reconnaisseurs de parole *monolocuteur, multi-locuteur* ou *indépendant du locuteur*, selon qu'il a été entraîné à reconnaître la voix d'une personne, d'un groupe fini de personnes, ou qu'il est en principe capable de reconnaître n'importe qui.

  - On distingue enfin reconnaisseurs de *mots isolés*, reconnaisseurs de *mots connectés*, et reconnaisseurs de *parole continue*, selon que le locuteur sépare chaque mot par un silence, qu'il prononce de façon continue une suite de mots prédéfinis ou qu'il prononce n'importe quelle suite de mots de façon continue.

- Les *synthétiseurs* ont quant à eux la fonction inverse de celle des analyseurs et des reconnaisseurs de parole : ils produisent de la parole artificielle. On distingue fondamentalement deux types de synthétiseurs : les *synthétiseurs de parole à partir d'une représentation numérique*, inverses des analyseurs, dont la mission est de produire de la parole à partir des caractéristiques numériques d'un signal vocal telles qu'obtenues par analyse, et les *synthétiseurs de parole à partir d'une représentation symbolique*¸ inverse des reconnaisseurs de parole et capables en principe de prononcer n'importe quelle phrase sans qu'il soit nécessaire de la faire prononcer par un locuteur humain au préalable. Dans cette seconde catégorie, on classe également les synthétiseurs en fonction de leur mode opératoire :

  - Les *synthétiseurs à partir du texte* reçoivent en entrée un texte orthographique et doivent en donner lecture.

- Les *synthétiseurs à partir de concepts*, appelés à être insérés dans des systèmes de dialogue homme-machine, reçoivent le texte à prononcer et sa structure linguistique, telle que produite par le système de dialogue.

- Enfin, le rôle des *codeurs* est de permettre la transmission ou le stockage de parole avec un débit réduit, ce qui passe tout naturellement par une prise en compte judicieuse des propriétés de production et de perception de la parole.

On comprend aisément que, pour obtenir de bons résultats dans chacune de ces tâches, il faut tenir compte des caractéristiques du signal étudié. Et, vu la complexité de ce signal, due en grande partie au couplage étroit entre production, perception, et compréhension, il n'est pas étonnant que les recherches menées par les spécialistes soient directement liées aux progrès obtenus dans de nombreuses autres disciplines scientifiques, progrès dont elles sont par ailleurs souvent à la fois les bénéficiaires et les instigatrices.

2 NIVEAUX DESCRIPTIFS DE LA PAROLE

La parole est le support de communication privilégie de l'homme. En tant que phénomène physique, la parole est le signal acoustique produit par le système vocal. Elle nécessite des processus de production et de perception permettant de transmettre des informations d'un être humain à un autre sous forme acoustique.

La parole transmet la pensée considérée comme information à travers le canal acoustique par l'intermédiaire de sons articulés différenciés, classés dans un dictionnaire fini qui dépend de la langue considérée.

Toute fois, les contraintes linguistiques (syntaxiques, morphologiques, sémantiques, pragmatiques) propres au langage introduisent une redandonce telle que même en présence de bruit important dans le canal de transmission, le cerveau peut saisir le sens d'un message.
L'information portée par le signal de parole peut être analysée de bien des façons. On en distingue généralement plusieurs niveaux de description non exclusifs : *Acoustique, phonétique, phonologique, morphologique, syntaxique, sémantique,* et *pragmatiqu*e.

2.1 Niveau acoustique

La parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire. La *phonétique acoustique*[3] étudie ce signal en le transformant dans un premier temps en signal électrique grâce au transducteur approprié : le microphone (lui-même associé à un préamplificateur). De nos jours, le signal électrique résultant est le plus souvent numérisé.

3. Dans la suite, nous présentons les niveaux acoustiques et phonétiques comme s'ils étaient indépendants bien que, stricto sensu, les aspects acoustiques de la parole sont du ressort d'une branche particulière de la phonétique : la phonétique acoustique (les autres étant la phonétique physiologique ou articulatoire, et la phonétique perceptive).

Il peut alors être soumis à un ensemble de traitements statistiques qui visent à en mettre en évidence les *traits acoustiques* : sa *fréquence fondamentale*, son *énergie*, et son *spectre*. Chaque trait acoustique est lui-même intimement lié à une grandeur perceptuelle : *pitch*, *intensité*, et *timbre*.

L'opération de numérisation, schématisée à la figure 2.1, requiert successivement : un *filtrage de garde*, un *échantillonnage*, et une *quantification*.

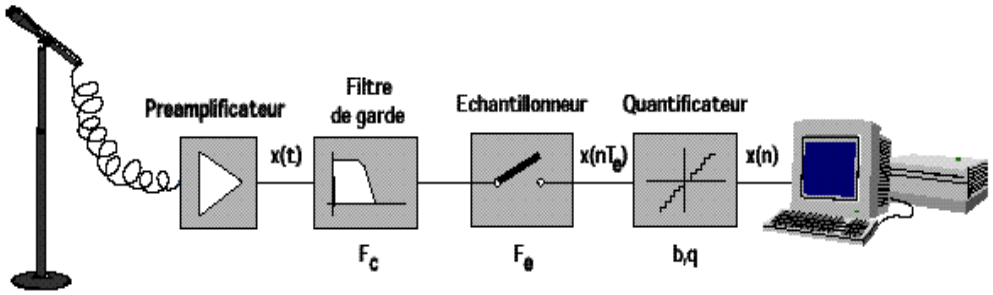


FIG 2.1 *Enregistrement numérique d'un signal acoustique. La fréquence de coupure du filtre de garde, la fréquence d'échantillonnage, le nombre de bits et le pas de quantification sont respectivement notés f c , f e , b, et q.*

2.1.1 Audiogramme

L'échantillonnage transforme le signal à temps continu *x(t)* en signal à temps discret $x(nT_e)$ défini aux instants d'échantillonnage, multiples entiers de la période d'échantillonnage T_e ; celle-ci est elle-même l'inverse de la fréquence d'échantillonnage f_e. Pour ce qui concerne le signal vocal, le choix de f_e résulte d'un compromis.

Son spectre peut s'étendre jusque 12 kHz. Il faut donc en principe choisir une fréquence f_e égale à 24 kHz au moins pour satisfaire raisonnablement au théorème de Shannon[4] [BOI87]. Cependant, le coût d'un traitement numérique, filtrage, transmission ou simplement enregistrement peut être réduit d'une façon notable si l'on accepte une limitation du spectre par un filtrage préalable. C'est le rôle du filtre de garde, dont la fréquence de coupure f_c est choisie en fonction de la fréquence d'échantillonnage retenue. Pour la téléphonie, on estime que le signal garde une qualité suffisante lorsque son spectre est limité à 3400 Hz et l'on choisit f_e = 8000 Hz. Pour les techniques d'analyse, de synthèse ou de reconnaissance de la parole, la fréquence peut varier de 6000 à 16000 Hz. Par contre pour le signal audio (parole et musique), on exige une bonne représentation du signal jusque 20 kHz et l'on utilise des fréquences d'échantillonnage de 44.1 ou 48 kHz. Pour les applications multimédia, les fréquences sous-multiples de 44.1 kHz sont de plus en plus utilisées: 22.5 kHz, 11.25 kHz.

4. Suivant le *théorème de Shannon*, les signaux doivent être échantillonnés à une fréquence d'échantillonnage supérieure ou égale à deux fois leur plus haute composante fréquentielle. Ils doivent être filtrés passe-bas dans le cas contraire.

Parmi le continuum des valeurs possibles pour les échantillons $x(nT_e)$, la quantification ne retient qu'un nombre fini $2b$ de valeurs (b étant le nombre de bits de la quantification), espacées du pas de quantification q. Le signal numérique résultant est noté $x(n)$. La quantification produit une erreur de quantification qui normalement se comporte comme un bruit blanc; le pas de quantification est donc imposé par le rapport signal à bruit à garantir. Si le pas de quantification est constant, ce rapport est fonction de l'amplitude du signal; les signaux de faible amplitude sont dès lors mal représentés. Aussi adopte-t-on pour la transmission téléphonique une loi de quantification logarithmique et chaque échantillon est représenté sur 8 bits (256 valeurs). Par contre, la quantification du signal musical exige en principe une quantification linéaire sur 16 bits (65536 valeurs).

Une caractéristique essentielle qui résulte du mode de représentation est le débit binaire, exprimé en bits par seconde (b/s), nécessaire pour une transmission ou un enregistrement du signal vocal. La transmission téléphonique classique exige un débit de 8 kHz x 8 bits = 64 Kb/s; la transmission ou l'enregistrement d'un signal audio exige en principe un débit de l'ordre de 48 kHz x 16 bits = 768 Kb/s (à multiplier par deux pour un signal stéréophonique)[5].

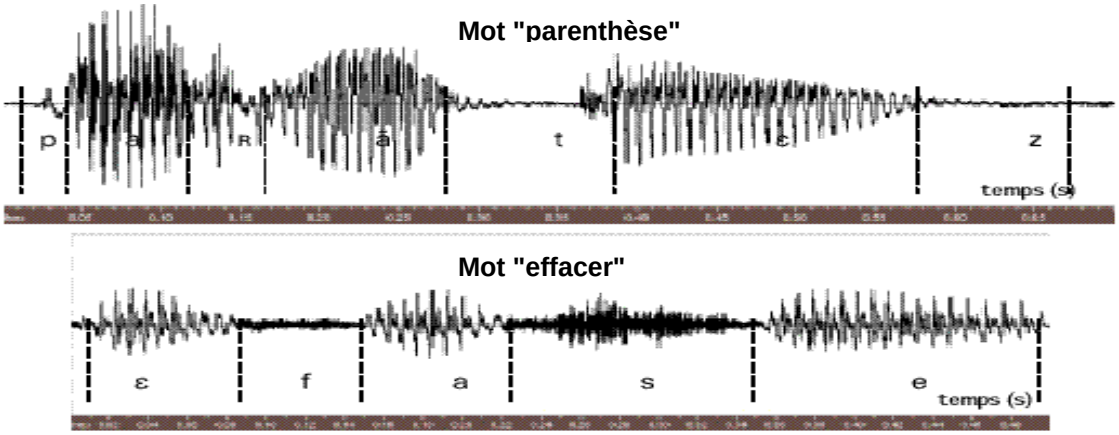


FIG 2.2 *Audiogramme de signaux de parole*

5. La redondance naturelle du signal vocal permet de réduire le débit binaire dans une très large mesure, au prix d'un traitement plus ou moins complexe et au risque d'une certaine dégradation de la qualité de la représentation.

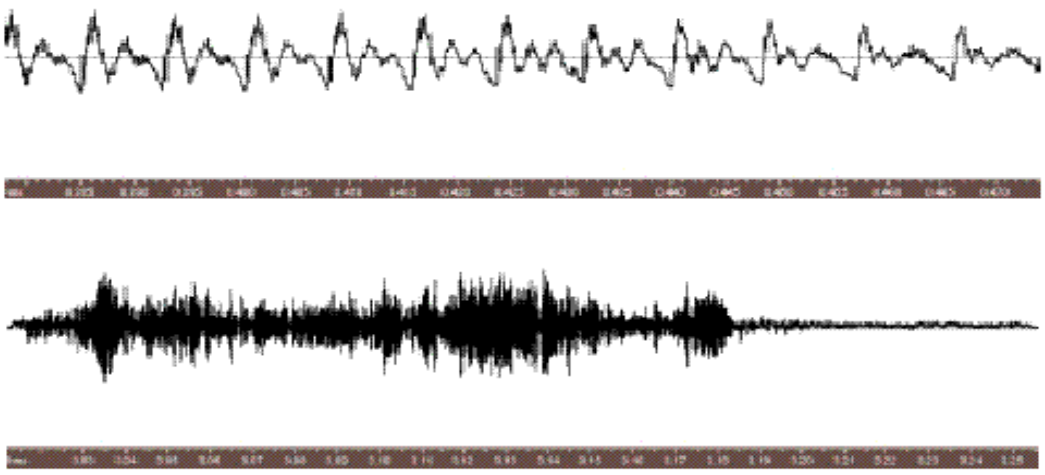


FIG 2.3 *Exemples de son voisé (haut) et non-voisé (bas)*

La figure 2.2 représente l'évolution temporelle ou *audiogramm*e, du signal vocal pour les mots 'parenthèse', et 'effacer'. On y constate une alternance de zones assez périodiques et de zones bruitées, appelées zones *voisées* et *non-voisées*. La figure 2.3 donne une représentation plus fine de tranches de signaux voisés et non voisés. L'évolution temporelle ne fournit cependant pas directement les traits acoustiques du signal. Il est nécessaire, pour les obtenir, de mener à bien un ensemble de calculs ad-hoc qui est l'objectif des paragraphes suivants.

2.1.2 Transformée de Fourier à court terme

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

La figure 2.4 illustre la transformée de Fourier d'une tranche voisée et celle d'une tranche non-voisée. Les parties voisées du signal apparaissant sous la forme de successions de pics spectraux marqués, dont les fréquences centrales sont multiples de la fréquence fondamentale. Par contre, le spectre d'un signal non voisé ne présente aucune structure particulière. La forme générale de ces spectres, appelée *enveloppe spectrale*, présente elle-même des pics et des creux qui correspondent aux résonances et aux anti-résonances du conduit vocal et sont appelés *formants* et *anti-formants*. L'évolution temporelle de leur fréquence centrale et de leur largeur de bande détermine le timbre du son. Il apparaît en pratique que l'enveloppe spectrale des sons voisés est de type passe-bas, avec environ un formant par kHz de bande passante, et dont seuls les trois ou quatre premiers contribuent de façon importante au timbre. Par contre, les sons non-voisés présentent souvent une accentuation vers les hautes fréquences.

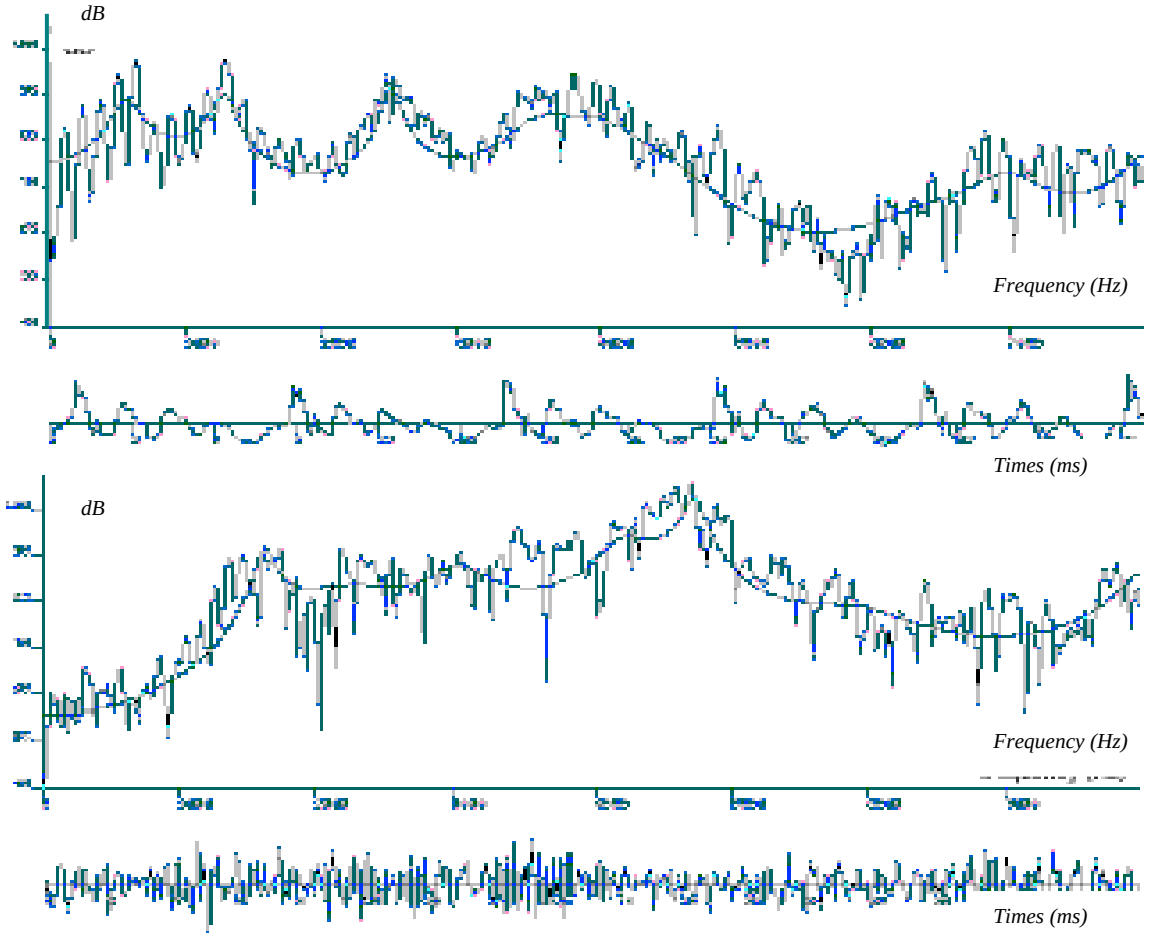


FIG 2.4 *Evolution temporelle (en haut) et transformée de Fourier discrète (en bas) du [a] et du [ʃ] de 'baluchon' (signaux pondérés par une fenêtre de Hamming de 30 ms).*

2.1.3 Spectrogramme

Il est souvent intéressant de représenter l'évolution temporelle du spectre à court terme d'un signal, sous la forme d'un *spectrogramm*e. L'amplitude du spectre y apparaît sous la forme de niveaux de gris dans un diagramme en deux dimensions temps-fréquence. On parle de spectrogramme *à large bande* ou *à bande étroite* selon la durée de la fenêtre de pondération (figure 2.5). Les spectrogrammes à bande large sont obtenus avec des fenêtres de pondération de faible durée (typiquement 10 ms); ils mettent en évidence l'enveloppe spectrale du signal, et permettent par conséquent de visualiser l'évolution temporelle des formants. Les périodes voisées y apparaissent sous la forme de bandes verticales plus sombres. Les spectrogrammes à bande étroite sont moins utilisés. Ils mettent plutôt la structure fine du spectre en évidence : les harmoniques du signal dans les zones voisées y apparaissent sous la forme de bandes horizontales.

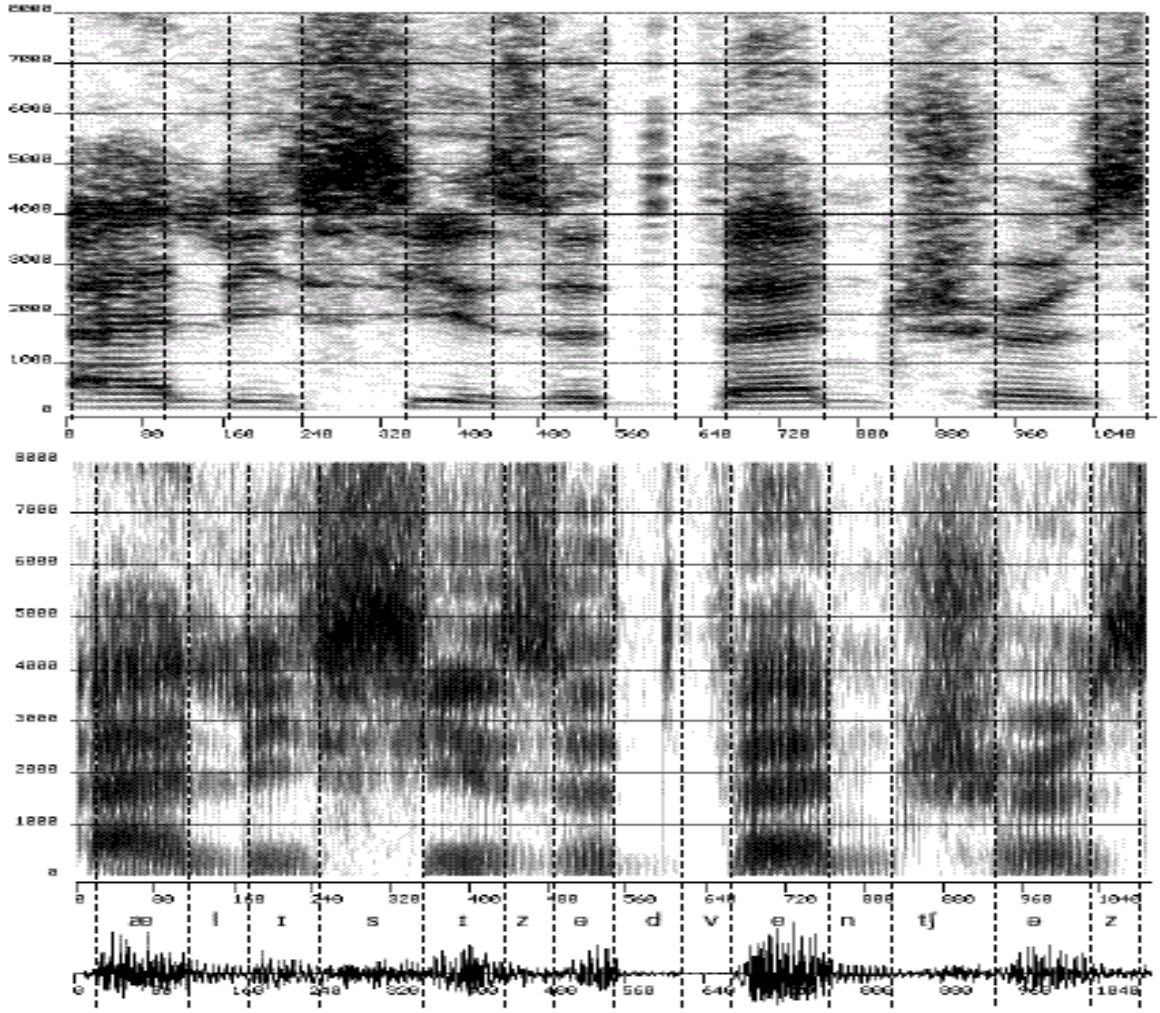


FIG 2.5 *Spectrogrammes à large bande (en bas), à bande étroite (en haut), et évolution temporelle de la phrase anglaise 'Alice's adventures', échantillonnée à 11.25 kHz (calcul avec fenêtres de Hamming de 10 et 30 ms respectivement).*

2.1.4 Fréquence fondamentale

Une analyse d'un signal de parole n'est pas complète tant qu'on n'a pas mesuré l'évolution temporelle de la fréquence fondamentale ou *pitch* :

- La fréquence fondamentale est la fréquence instantanée à laquelle vibrent les cordes vocales, elle fixe la hauteur d'une voix, ses variations contribuent à la perception de mélodie d'une phrase.

- Le pitch, couvre les 3 acceptations suivantes : Fréquence laryngienne si l'on veut faire référence au processus de génération articulatoire, fréquence fondamentale F_0 si

l'on se place plutôt dans le domaine acoustique, hauteur de la voix pour renvoyer au champ perceptif.

La figure 2.6 donne l 'évolution temporelle de la fréquence fondamentale de la phrase *"les techniques de traitement de la parole*". On constate qu'à l'intérieur des zones voisées la fréquence fondamentale évolue lentement dans le temps. Elle s'étend approximativement de 70 à 250 Hz chez les hommes, de 150 à 400 Hz chez les femmes, et de 200 à 600 Hz chez les enfants.

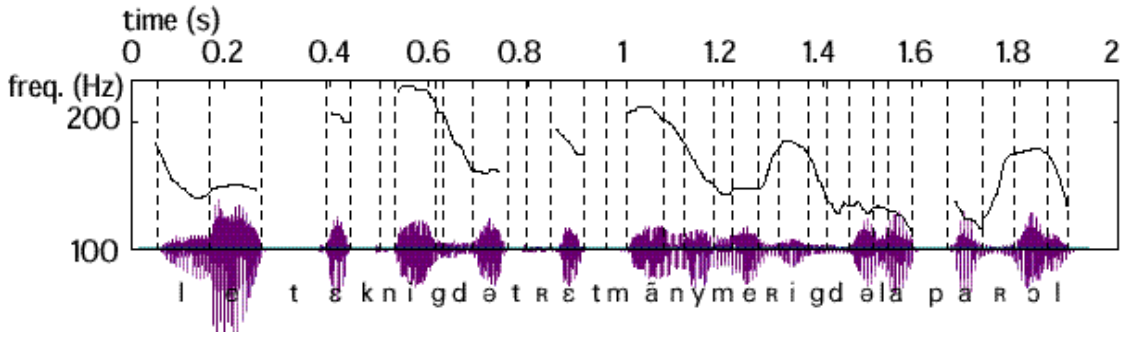


FIG 2.6 *Evolution de la fréquence de vibration des cordes vocales dans la phrase "les techniques de traitement numérique de la parole". La fréquence est donnée sur une échelle logarithmique; les sons non-voisés sont associés à une fréquence nulle.*

2.2 Niveau phonétique

Au contraire des acousticiens, ce n'est pas tant le signal qui intéresse les phonéticiens que la façon dont il est produit par le système articulatoire, présenté à la figure 2.7, et perçu par le système auditif.

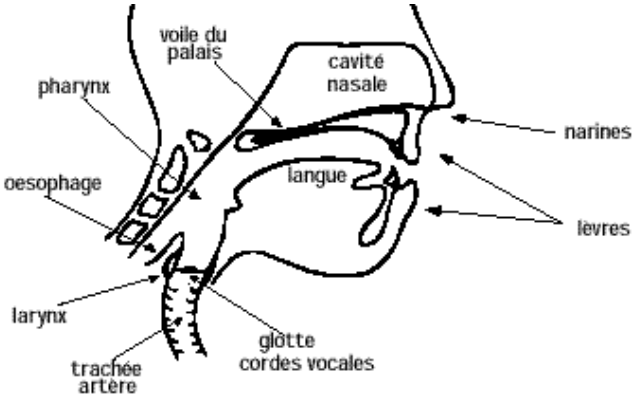


FIG 2.7 *L'appareil phonatoire*

2.2.1 Phonation

La parole peut être décrite comme le résultat de l'action volontaire et coordonnée d'un certain nombre de muscles. Cette action se déroule sous le contrôle du système nerveux central qui reçoit en permanence des informations par rétroaction auditive et par les sensations kinesthésiques. L'appareil respiratoire fournit l'énergie nécessaire à la production de sons, en poussant de l'air à travers la trachée-artère. Au sommet de celle-ci se trouve le *larynx* où la pression de l'air est modulée avant d'être appliquée au conduit vocal. Le larynx est un ensemble de muscles et de cartilages mobiles qui entourent une cavité située à la partie supérieure de la trachée (Figure 2.8). Les *cordes vocales* sont en fait deux lèvres symétriques placées en travers du larynx. Ces lèvres peuvent fermer complètement le larynx et, en s'écartant progressivement, déterminer une ouverture triangulaire appelée *glott*e. L'air y passe librement pendant la respiration et la voix chuchotée, ainsi que pendant la phonation des sons non-voisés (ou *sourds*[6]). Les sons voisés (ou *sonores*) résultent au contraire d'une vibration périodique des cordes vocales. Le larynx est d'abord complètement fermé, ce qui accroît la pression en amont des cordes vocales, et les forces à s'ouvrir, ce qui fait tomber la pression, et permet aux cordes vocales de se refermer; des impulsions périodiques de pression sont ainsi appliquées au conduit vocal, composé des cavités pharyngienne et buccale pour la plupart des sons. Lorsque la *luette* est en position basse, la cavité nasale vient s'y ajouter en dérivation. Notons pour terminer le rôle prépondérant de la langue dans le processus phonatoire. Sa hauteur détermine la hauteur du pharynx : plus la langue est basse, plus le pharynx est court. Elle détermine aussi le *lieu d'articulation*, région de rétrécissement maximal du canal buccal, ainsi que l'*aperture*, écartement des organes au point d'articulation [CAL89].

6 Les phonéticiens appellent *sourd* ou *sonore* ce que les ingénieurs qualifient de *voisé* ou *non voisé*.

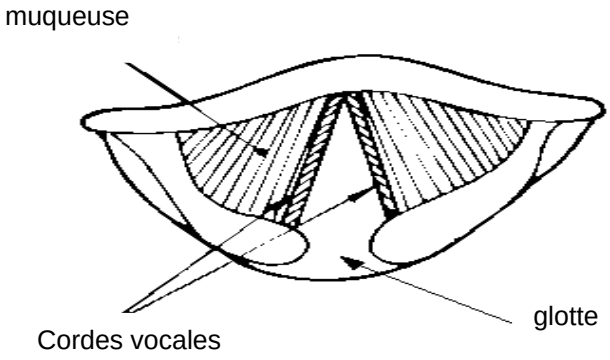


FIG 2.8 *Section du larynx, vu de haut*

2.2.2 Alphabet Phonétique International

L'Alphabet Phonétique International (IPA) associe des symboles phonétiques aux sons, de façon à permettre l'écriture compacte et universelle des prononciations (voir tableau 2.1 pour le français) [BOI87].

2.2.3 Phonétique articulatoire

La parole se distingue des autres sons par des caractéristiques acoustiques ayant leurs origines dans les mécanismes de production. Les sons de parole sont produit soit par les vibrations des cordes vocales (sources de voisements), soit par l'écoulement turbulent de l'air dans le conduit vocal, soit lors du relâchement d'une occlusion de ce conduit (source de bruit).

Il est intéressant de grouper les sons de parole en classes phonétiques, en fonction de leur *mode articulatoire*. On distingue généralement trois classes principales : les *voyelles*, les *semi-voyelles* et les *liquides*, et les *consonnes* [GAL90] ; [JUN90].

2.2.3.1 Caractéristiques phonétiques du Français

1. Voyelles

Les voyelles [i, e, ɛ, ∂, a, ɔ, o, y, u, ϕ, oe, e, $\tilde{\varepsilon}$, $\tilde{\partial}$, $\tilde{\jmath}$, o$\tilde{e}$] diffèrent de tous les autres sons par le degré d'ouverture du conduit vocal (et non, comme on l'entend souvent dire, par le degré d'activité des cordes vocales, déjà mentionné sous le terme de *voisement*). Si le conduit vocal est suffisamment ouvert pour que l'air poussé par les poumons le traverse sans obstacle, il y a production d'une voyelle. Le rôle de la bouche se réduit alors à une modification du timbre vocalique. Si, au contraire, le passage se rétrécit par endroit, ou même s'il se ferme temporairement, le passage forcé de l'air donne naissance à un bruit : Une consonne est produite. La bouche est dans ce cas un organe de production à part entière.

TAB 2.1 *Alphabet phonétique international pour le Français*

Phonèmes																	
Voyelles				**Semi-consonnes**		**Consonnes**											
Orales		**Nasales**				**Liquides**		**Nasales**		**Fricatives**				**Occlusives**			
										Voisées		***Non voisées***		***Voisées***		***Non voisées***	
[i]	Lit	**[ɛ̃]**	Lin	**[j]**	Yéyé	**[l]**	Lait	**[m]**	Mais	**[v]**	Vais	**[f]**	Fait	**[b]**	Baie	**[p]**	Paie
[e]	Les	**[ɑ̃]**	Lent	**[w]**	Wais	**[R]**	Raie	**[n]**	Nez	**[z]**	Zéros	**[s]**	Sait	**[d]**	Dais	**[t]**	Taie
[ɛ]	Lait	**[ɔ̃]**	Long	**[y]**	Huer			**[h]**	Gagner	**[z]**	Jeux	**[ʃ]**	Chez	**[g]**	Gai	**[k]**	Quai
[a]	Là	**[œ̃]**	Un														
[u]	Loup																
[o]	Lot																
[y]	Lu																
[Ø]	Leu																
[œ]	Leur																
[∂]	Le																

Les voyelles se différencient principalement les unes des autres par leur *lieu d'articulation*, leur *aperture*, et leur *nasalisation*. On distingue ainsi, selon la localisation de la masse de la langue, les voyelles *antérieures*, les voyelles *moyennes*, et les voyelles *postérieures*, et, selon l'écartement entre l'organe et le lieu d'articulation, les voyelles *fermées* et *ouvertes*. Les voyelles *nasales* [$\tilde{\varepsilon}$, $\tilde{\partial}$, $\tilde{\varsigma}$, oe~] diffèrent des voyelles *orales* [i, e, ɛ, ∂, a, ɔ, o, y, u, ɸ, oe, e] en ceci que le voile du palais est abaissé pour leur prononciation, ce qui met en parallèle les cavités nasales et buccales. Notons que, dans un contexte plus général que celui de la seule langue française, d'autres critères peuvent être nécessaires pour différencier les voyelles, comme leur *labialisation*, leur *durée*, leur *tension*, leur *stabilité*, leur *glottalisation*¸ voire même la *direction du mouvement de l'air*.

2. Semi- voyelles

Les semi-voyelles [j, w, y], quant à elles, combinent certaines caractéristiques des voyelles et des consonnes. Comme les voyelles, leur position centrale est assez ouverte, mais le relâchement soudain de cette position produit une friction qui est typique des consonnes.

3. Liquides

Les liquides [l, R] sont assez difficiles à classer. L'articulation de [l] ressemble à celle d'une voyelle, mais la position de la langue conduit à une fermeture partielle du conduit vocal. Le son [R], quant à lui, admet plusieurs réalisations fort différentes.

4. Consonnes

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

En français, la distinction de mode d'articulation conduit à deux classes : les *fricatives* (ou *constrictives*) et les *occlusives* (ou *plosives*).

- Les fricatives sont créées par une constriction du conduit vocal au niveau du lieu d'articulation, qui peut être le palais [ʃ , z], les dents [s, z], ou les lèvres [f, v]. Les fricatives non-voisées sont caractérisées par un écoulement d'air turbulent à travers la glotte, tandis que les fricatives voisées combinent des composantes d'excitation périodique et turbulente : les cordes vocales s'ouvrent et se ferment d'une façon périodique, mais la fermeture n'est jamais complète.

- Les occlusives correspondent quant à elles à des sons essentiellement dynamiques. Une forte pression est créée en amont d'une occlusion maintenue en un certain point du conduit vocal (qui peut ici aussi être le palais [k, g], les dents [t, d] ou les lèvres [p, b]), puis relâché brusquement. La période d'occlusion est

appelée la phase de tenue. Pour les occlusives voisées [p, d, g] un son à basse fréquence est émis par vibration des cordes vocales pendant la phase de tenue; pour les occlusives non voisées [p ,t , k], la tenue est un silence.

- Les nasales [m, n, ŋ] font intervenir les cavités nasales par abaissement du voile du palais.

Les traits acoustiques du signal de parole sont évidemment liés à sa production. L'intensité du son est liée à la pression de l'air en amont du larynx. Sa fréquence, qui n'est rien d'autre que la fréquence du cycle d'ouverture/fermeture des cordes vocales, est déterminée par la tension de muscles qui les contrôlent. Son spectre résulte du filtrage dynamique du signal glottique (impulsions, bruit ou combinaison des deux) par le conduit vocal, qui peut être considéré comme une succession de tubes ou de cavités acoustiques de sections diverses. Ainsi, par exemple, on peut approximativement représenter les voyelles dans le plan des deux premiers formants (Fig.2.9). On observe en pratique un certain recouvrement dans les zones formantiques correspondant à chaque voyelle (un affichage en trois dimensions figurant les trois premiers formants permettrait une meilleure séparation) [CAL89].

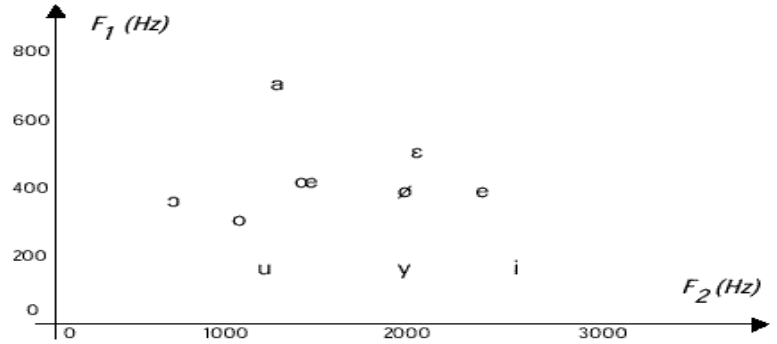


FIG 2.9 *Représentation des voyelles dans le plan F1-F2*

Nous présentons dans le paragraphe au-dessous, une classification des lettres arabes faite par des spécialistes en linguistique qui sera la base de l'analyse d'un ensemble de phonèmes dans le but d'extraire les paramètres pertinents utilisés ultérieurement pour notre système de reconnaissance.

2.2.3.2 Caractéristiques phonétiques de l'Arabe

La langue arabe utilisée dans l'enseignement, les médias et les manifestations culturelles dans les pays arabes, constitue un domaine d'investigation intéressant sur plusieurs plans, du fait du nombre élevé des consonnes. Cependant, le traitement de la parole arabe n'a pas été l'objet de beaucoup de travaux contrairement à l'Anglais, le Français,... etc. nous pouvons signaler : Les travaux de : Al-Ani(1970), Ghazali(1983), Rajouani(1986), Youssef- El Imam(1989), Ahmed- El Shafi(1989), Znagui(1992-1993).

La langue arabe qui s'écrit et se lit de droite à gauche, se compose d'un certain nombre de symboles ou graphismes : 28 consonnes, 3 voyelles courtes, 2 semi- voyelles, 3 voyelles longues, les chiffres et les symboles de ponctuation.

La langue arabe est définit comme étant une langue consonantique oú il y a beaucoup de consonnes et peu de voyelles. La classification des sons de l'Arabe est basée selon [TAB97]:

- Le lieu d'articulation qui représente la zone du conduit vocal participant à la formation du son (bilabiale, labiodentale, inter- dentale, dentale,... etc.) (voire tableau 2. 3).

- Le mode d'articulation qui est lié aux diverses sources d'excitation du conduit vocal (orale, nasale, semi- voyelle, liquide, fricative, occlusive) (voire tableau 2.4) .

- Le voisement, c'est la vibration ou non des cordes vocales (voisé ou sonore, non voisé ou sourd) (voire tableau 2.5) .

1. Consonnes

La plupart des consonnes de la langue arabe prennent 4 formes, en apparence différente, selon qu'elles sont isolées, initiales, médiales ou finales. Cela tient à ce qu'elles se composent d'une forme essentielle et d'un appendice. Cette appendice disparaît lorsque la lettre doit être liée à celle qui la suit (voire tableau 2.2).

2. Voyelles

- Voyelles courtes

En Arabe, il y a 3 voyelles courtes classées d'après la position des organes de phonation qui concourent à leur émission, lèvres et langue.

- La première voyelle se prononce en contractant la langue au fond de la bouche et en avançant les lèvres qui s'arrondissent jusqu'à presque se joindre. Elle est représentée par le signe « ُ » placé au-dessus de la consonne, appelé ***damma***. Elle est prononcée comme « ***ou*** » français au voisinage des consonnes **[ب,ن,م,ت]**. Elle est prononcée comme « ***u*** » français (entre « ***ou*** » et « ***o*** » fermé) au voisinage des consonnes **[ص,ض,ط,ظ,خ,غ,ف,ح,ع]**.

- La deuxième voyelle se prononce en ouvrant largement la bouche et en conservant la langue dans une position horizontale. Elle est représentée par le signe « َ » placé au-dessus de la consonne, appelé ***fatha***. Elle est prononcée comme « ***a*** » moyen français, au voisinage des consonnes **[ح,ف,ط,ص]**. Elle est prononcée entre « ***a*** » moyen et « ***e*** » ouvert français, au voisinage des consonnes **[ن,م,ت,ب]**.

- La troisième voyelle se prononce en portant le devant de la langue en avant et en l'étale largement tandis que l'arrière frôle presque le palais et que les commissures des lèvres s'étirent. Elle est représentée par le signe « ِ » placé au-dessous de la consonne, appelé ***kasra***. Elle est prononcée comme « ***i*** » fermé français au voisinage des consonnes **[ن,م,ت,ب]**. Elle est prononcée entre « ***i*** » et « ***é*** » fermé français au voisinage des consonnes **[ح,ف,ط,ص]**.

- Voyelles longues

En Arabe, il y a 3 voyelles longues :

- La consonne [و – ***waw***] dépourvue de voyelle et précédée d'une ***damma*** cesse d'être consonne et devient une voyelle longue :[**بو – *bu***, **تو – *tu***, **نو – *nu***].

- La consonne [أ – ***alif***] dépourvue de [ء – ***hamza***] sert à allonger la voyelle courte [أ] qui le précède, ce qui nous donne la voyelle longue [آ – ***aa***] : [**با – *baa***, **تا – *taa***, **ما – *maa***].

- La consonne [ي – ***ya***] dépourvue de voyelle et précédée d'une ***kasra*** cesse d'être une consonne et devient la voyelle longue [إي – ***ii***] : [**بي – *b ii***, **تي – *tii***, **مي – *mii***].

- Semi- voyelles

Dans l'alphabet arabe, il existe 2 semi -voyelles :

- [و w, ولد – ***w****aled*].
- [ي y, بيت – *bay**t***].

3. Autres graphismes

- Sukune

C'est le signe « ° » placé au-dessus de la consonne pour indiquer que cette consonne n'est pas menue de voyelle : [**من *quiconque***, **عن *loin de***, **سل *interroge***].

- Tanwin

Les 3 signes qui représentent les voyelles sont quelques fois redoublés à la fin des noms et les voyelles finales se lisent alors comme si elles étaient suivies du son « ***n*** ». Nous appelons ce phénomène ***tanwin***, il indique l'indétermination :

- Cas du ***marfu*** : [**كتابٌ** – *kitab**un***].
- Cas du ***mansub*** : [**كتاباً** – *kita**ban***].
- Cas du ***majzum*** : [**كتابٍ** – *kita**bin***].

- Tachdid

Il est présenté par le signe « ّ » placé au-dessus d'une consonne pour indiquer la gémination. Une consonne géminée surmontée d'un ***tachdid*** se prononce comme si elle était écrite 2 fois, la première portant le sukune et terminant la syllabe précédente, la deuxième portant la voyelle qui accompagne le tachdid et commençant une syllabe : [مسّ équivalent à مسْسَ – ***mas-sa***].

TAB 2.5 *Classification des sons arabes*

Voisement					
Voisés - sonores		**Non voisés - sourds**		**Non sonores - Non sourds**	
ـَ	Fatha	تـ	Ta	ءـ	Hamza
ـُ	Damma	طـ	Tâ		
ـِ	Kasra	كـ	Kaf		
آ	Aa	قـ	Qaf		
أو	Ui	فـ	Fa		
إي	Ii	ثـ	Tha		
جـ	Djim	سـ	Sin		
وـ	Waw	شـ	Chin		
يـ	Ya	خـ	Kha		
بـ	Ba	هـ	Ha		
ضـ	Dhad	حـ	Hâ		
ر	Ra				
لـ	Lam				
مـ	Mim				
نـ	Nun				
ذ	Dhal				
ظـ	Dâ				
زـ	Zay				
غـ	Ghayn				

TAB 2.2 *Les consonnes de l'alphabet arabe*

Figure	Nom	IPA	Prononciation
أ	Alif	a	***a, français***
ب	Ba	b	***b, français***
ت	Ta	t	***t, français***
ث	Tha	th	***the dur, anglais***
ج	Djim	dj	***giorgio, italien***
ح	Hâ	h	-
خ	Kha	kh	***nach, allemand***
د	Dal	d	***d, français***
ذ	Dhal	dh	***th doux, anglais***
ر	Ra	r	***r roulé, français***
ز	Zay	z	***z, français***
س	Sin	s	***s, français***
ش	Chin	ch	***chat, français***
ص	Sad	ŝ	***s emphase, français***
ض	Dhad	d˙	-
ط	Tâ	t˙	***t emphase, français***
ظ	Dâ	z˙	-
ع	Ayn	-	-
غ	Ghayn	gh, rh	***r grasseyé, français***
ف	Fa	f	***f , français***
ق	Qâf	q	-
ك	Kâf	k	***k , français***
ل	Lam	l	***l, français***
م	Mim	m	***m, français***
ن	Nun	n	***n, français***
ه	Ha	h	***h, anglais***
و	Waw	û, w	***w, anglais***
ي	Ya	î, y	***yougoslavie, français***

TAB 2.3 *Classification des sons arabes selon le lieu d'articulation*

Lieu d'articulation										
Uvulaire	**Pharyngale**	**Laryngale**	**Bilabiale**	**Labiodentale**	**Interdentale**	**Dentale**	**Alvéolaire**	**Palatale**	**Vélaire**	**Postpalatale**
ق	ح	ء	ب	ف	ذ	د	ر	ـِ	ـ	ـُ
	ع	ه	م		ث	ت	ز	إي	أ	أو
			و		ظ	ض	س	ي	ك	و
						ط	ص	ش		خ
						ل	ن	ج		غ

TAB 2.4 *Classification des sons arabes selon le mode d'articulation*

Mode d'articulation							
Voyelle orale	**Semi voyelle**	**Fricative**	**Son combiné**	**Occlusive**	**Nasale**	**Littérale liquide**	**Vibrant**
ـ	و	ف	ج	ب	م	ل	د
ـُ	ي	د		ض	ن		
ـِ		ث		ر			
آ		ظ		ل			
أو		ز		ت			
إي		س		ط			
		ص		ك			
		ش		ق			
		خ					
		غ					
		ح					
		ع					
		ه					

2.3 Audition - perception

Dans le cadre du traitement de la parole, une bonne connaissance des mécanismes de l'audition et des propriétés perceptuelles de l'oreille est aussi importante qu'une maîtrise des mécanismes de production. En effet, tout ce qui peut être mesuré acoustiquement ou observé par la phonétique articulatoire n'est pas nécessairement perçu.

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

L'appareil auditif comprend l'*oreille externe*, l'*oreille moyenne*, et l'*oreille interne* (figure 2.10). Le conduit auditif relie le pavillon au tympan : c'est un tube acoustique de section uniforme fermé à une extrémité, son premier mode de résonance est situé vers 3000 Hz, ce qui accroît la sensibilité du système auditif dans cette gamme de fréquences. Le mécanisme de l'oreille interne (*marteau, étrier, enclume*) permet une adaptation d'impédance entre l'air et le milieu liquide de l'oreille interne. Les vibrations de l'étrier sont transmises au liquide de la *cochlée*. Celle-ci contient la *membrane basilaire* qui transforme les vibrations mécaniques en impulsions nerveuses. La membrane s'élargit et s'épaissit au fur et à mesure que l'on se rapproche de l'apex de la cochlée; elle est le support de l'*organe de Corti* qui est constitué par environ 25000 *cellules ciliées* raccordées au nerf auditif. La réponse en fréquence du conduit au droit de chaque cellule est esquissée à la figure 2.11. La fréquence de résonance dépend de la position occupée par la cellule sur la membrane; au-delà de cette fréquence, la fonction de réponse s'atténue très vite. Les fibres nerveuses aboutissent à une région de l'écorce cérébrale appelée *aire de projection auditive* et située dans le lobe temporal. En cas de lésion de cette aire, on peut observer des troubles auditifs. Les fibres nerveuses auditives afférentes (de l'oreille au cerveau) et efférentes (du cerveau vers l'oreille) sont partiellement croisées : chaque moitié du cerveau est mise en relation avec les deux oreilles internes [LAN77].

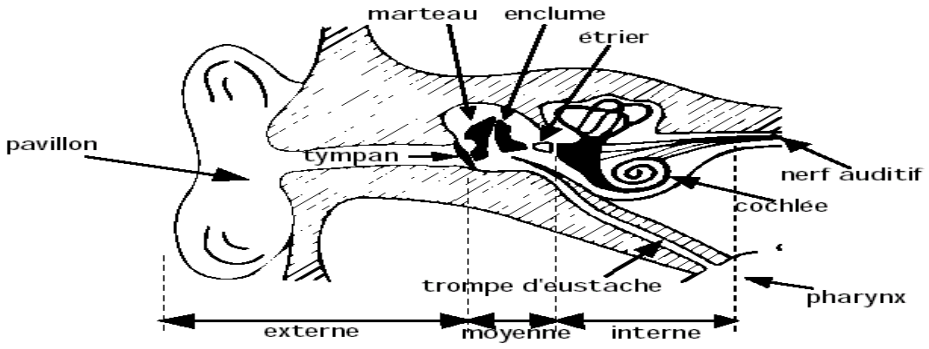


FIG 2.10 *Le système auditif*

Il reste très difficile de nos jours de dire comment l'information auditive est traitée par le cerveau. Les chercheurs ont pu par contre étudier comment elle était finalement perçue, dans le cadre d'une science spécifique appelée *psychoacoustique*. Il est intéressant d'en connaître les résultats les plus marquants de la contribution majeure des psychoacousticiens dans l'étude de la parole.

Ainsi, l'oreille ne répond pas également à toutes les fréquences. La figure 2.12 présente le champ auditif humain, délimité par la courbe de *seuil de l'audition* et celle du *seuil de la douleur*. Sa limite supérieure en fréquence (≈16000 Hz, variable selon les individus) fixe la fréquence d'échantillonnage maximale utile pour un signal auditif (≈ 32000 Hz).

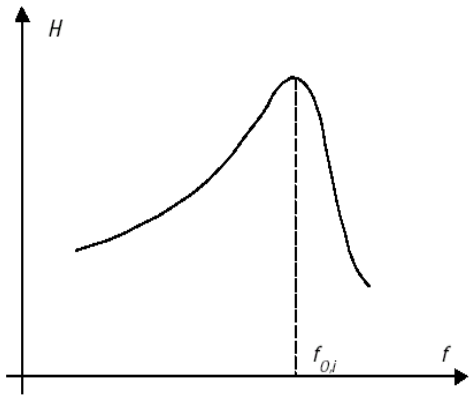


FIG 2.11 *Réponse en fréquence Humaine d'une cellule ciliée*

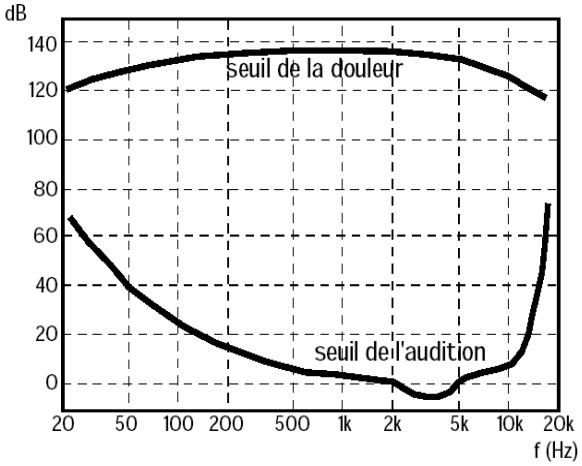


FIG 2.12 *Le champ auditif*

A l'intérieur de son domaine d'audition, l'oreille ne présente pas une sensibilité identique à toutes les fréquences. La figure 2.13. a fait apparaître les courbes d'égale impression de puissance auditive (aussi appelée *sonie*, exprimée en *sones*) en fonction de la fréquence. Elles révèlent un maximum de sensibilité dans la plage [500 Hz, 10 KHZ], en dehors de laquelle les sons doivent être plus intenses pour être perçus.

Enfin, un son peut en cacher un autre. Cette propriété psychoacoustique, appelée *phénomène de masquage*, peut être visualisée sous la forme de courbes de masquage (Figure 2.13.b), qui mettent en évidence la modification locale du seuil d'audition en fonction de la présence d'un signal déterminé (un bruit à bande étroite centré sur 1 kHz dans le cas de la figure 2.13.b). Une modélisation efficiente des propriétés de masquage de l'oreille permet de réduire le débit binaire nécessaire au stockage ou à la transmission d'un signal acoustique, en éliminant les composantes inaudibles.

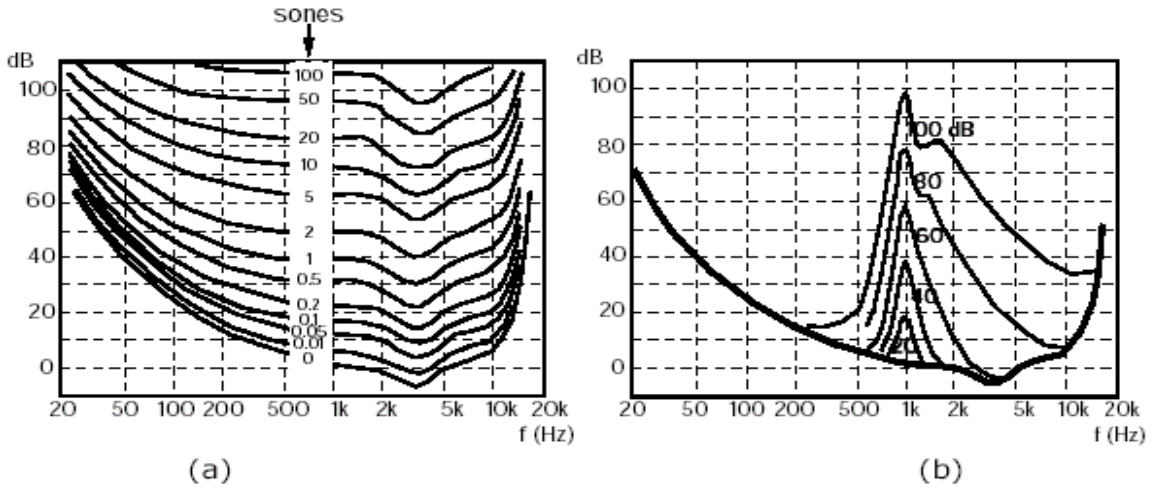


FIG 2.13 *(a) : Courbes isosoniques en champ ouvert.* (b) : Masquage auditif par un bruit à bande étroite : limite d'audibilité en fonction de la *puissance du bruit masquant*

Remarquons que ce qui est perçu n'est pas nécessairement *compris*. Une connaissance de la langue interfère naturellement avec les propriétés psychoacoustiques de l'oreille. En effet, les sons ne sont jamais prononcés isolément, et le contexte phonétique dans lequel ils apparaissent est lui aussi mis à contribution par le cerveau pour la compréhension du message. Ainsi, certains sons portent plus d'information que d'autres, dans la mesure où leur probabilité d'apparition à un endroit donné de la chaîne parlée est plus faible, de sorte qu'ils réduisent l'espace de recherche pour les sons voisins. Les sons sont organisés en unités plus larges, comme les mots, qui obéissent eux-mêmes à une syntaxe et constituent une phrase porteuse de sens. Par conséquent, c'est tout notre savoir linguistique qui est mis à contribution lors du décodage acoustico-phonétique[7].

3 MODELISATION DE LA PAROLE

L'analyse de la parole est une étape indispensable à toute application de synthèse, de codage ou de reconnaissance. Elle repose en général sur un *modèle*. Celui-ci possède un ensemble de *paramètres* numériques, dont les plages de variation définissent l'ensemble des signaux couverts par le modèle. Pour un signal et un modèle donné, l'*analyse* consiste en l'*estimation* des paramètres du modèle dans le but de lui faire correspondre le signal analysé. Pour ce faire, on met en oeuvre un *algorithme d'analyse*, qui cherche généralement à minimiser la différence, appelée *erreur de modélisation*, entre le signal original et celui qui serait produit par le modèle s'il était utilisé en tant que synthétiseur (figure 2.14).

7 Cette influence marquée de la langue sur la perception fait-elle aussi l'objet d'une étude spécifique, dans le cadre de la *psycholinguistique*.

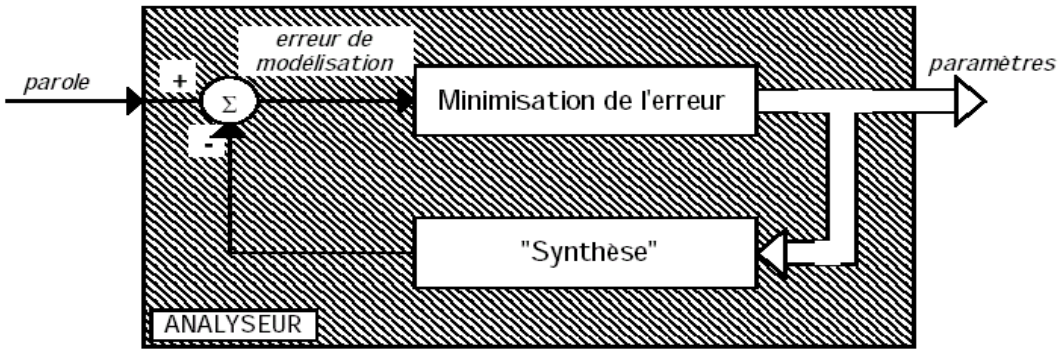


FIG 2.14 *Schéma de principe d'un analyseur de parole. En pratique, l'étape de synthèse peut être implicite*

3.1 Modèle électrique de la phonation : Le modèle Auto-Régressif

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

Un signal voisé peut être modélisé par le passage d'un train d'impulsions *u(n)* à travers un filtre numérique récursif de type *tous pôles*. On montre que cette modélisation reste valable dans le cas de sons non-voisés, à condition que *u(n)* soit cette fois un bruit blanc. Le modèle final est illustré à la figure 2.15. Il est souvent appelé *modèle auto-régressif*, parce qu'il correspond dans le domaine temporel à une régression linéaire de la forme :

$$x(n)=\delta u(n)+\sum_{i=1}^{\Gamma}-a_i x(n-i)$$

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

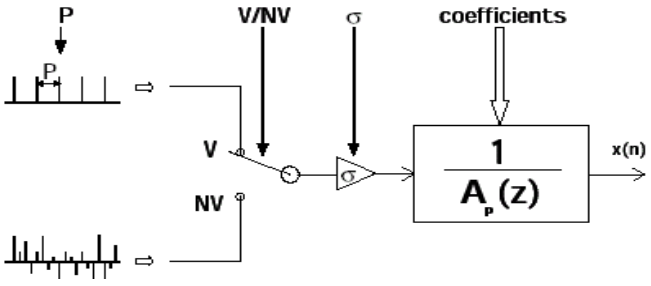


FIG 2.15 *Le modèle auto-régressif*

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

Le problème de l'estimation d'un modèle AR, souvent appelée *analyse LPC* (pour 'Linear Prediction Coding'[8]) revient à déterminer les coefficients d'un filtre tous pôles dont on connaît le signal de sortie, mais pas l'entrée. Il est par conséquent nécessaire d'adopter un critère, afin de faire un choix parmi l'infinité de solutions possibles. Le critère classiquement utilisé est celui de la *minimisation de l'énergie de l'erreur de prédiction*. La mise en équation de ce problème conduit aux équations dites de *Yule-Walker* :

$$\Phi a = -\phi, \quad \phi = \left[\phi_{xx}(1), \phi_{xx}(2), ..., \phi_{xx}(p)\right]^T, \quad a = \left[a_1, a_2, ..., a_p\right]^T$$

avec :

$$\Phi = \begin{Bmatrix} \phi_{xx}(0) & \phi_{xx}(1) & ... \phi_{xx}(p-1) \\ \phi_{xx}(1) & \phi_{xx}(0) & ... \phi_{xx}(p-2) \\ \phi_{xx}(2) & \phi_{xx}(1) & ... \phi_{xx}(p-3) \\ . & . & \\ . & . & \\ \phi_{xx}(p-1) & \phi_{xx}(p-2) & ... \phi_{xx}(0) \end{Bmatrix}$$

On constate en passant que la matrice Φ est symétrique et que les diagonales parallèles à la diagonale principale contiennent des éléments égaux. Une telle matrice est dite *Toeplitz*. Il existe dans ce cas des algorithmes rapides pour la résolution, appelés algorithmes de Levinson et de Schur.

3.2 Considérations pratiques

Pour mener à bien une analyse LPC, il faut pouvoir choisir [BOI87]:

- La fréquence d'échantillonnage *fe*;
- La méthode d'analyse et l'algorithme correspondant;
- L'ordre *p* de l'analyse LPC;
- Le nombre d'échantillons par tranche *N* et le décalage entre tranches successives *L*;

qualité du signal à analyser. On choisira plutôt 8 kHz pour les signaux téléphoniques, 10 kHz pour les applications de reconnaissance, et 16 kHz pour les applications de synthèse. Dans le cadre d'applications multimédia, on préférera les fréquences normalisées de 11.25 et 22.5 kHz, sous-multiples des 44.1 kHz du Compact Disk.

8. La prédiction linéaire a été initialement utilisée pour le codage de la parole.

Le choix de la fréquence d'échantillonnage est fonction de l'application visée et de la L'ordre d'analyse conditionne le nombre de formants que l'analyse est capable de prendre en compte. On estime en général que la parole présente un formant par kHz de bande passante, ce qui correspond à une paire de pôles pour $Ap(z)$. Si on y ajoute une paire de pôles pour la modélisation de l'excitation glottique, on obtient les valeurs classiques de p=10, 12, et 18 pour fe=8, 10 et 16 kHz respectivement. Elles trouvent d'ailleurs une justification expérimentale dans le fait que l'énergie de l'erreur de prédiction diminue rapidement lorsqu'on augmente p à partir de 1, pour tendre vers une asymptote autour de ces valeurs : il devient inutile d' augmenter encore l'ordre, puisqu'on ne prédit rien de plus.

La durée des tranches d'analyse et leur décalage sont souvent fixées à 30 et 10 ms respectivement. Ces valeurs ont été choisies empiriquement; elles sont liées au caractère quasi-stationnaire du signal de parole.

Enfin, pour compenser les effets de bord, on multiplie en général préalablement chaque tranche d'analyse par une fenêtre de pondération *w(n)* de type *fenêtre de Hamming* :

$w(n) = 0.54 + 0.46 \cos(2\Pi \frac{n}{N})$, pour n = 0 .. N-1

On retiendra donc que l'analyse LPC d'un signal de parole implique la résolution d'un système de (l'ordre de) **10 d'équations à 10 inconnues toutes les 10 ms.**

3.3 Un exemple complet

La figure 2.16. donne la représentation AR du mot *'parenthèse'* prononcé en Arabe (Fe= 8 kHz, p = 10), telle qu'obtenue après analyse par prédiction linéaire (LPC : Linear Prediction Coding). L'analyse est menée sur des tranches de 30 ms (240 échantillons), à raison d'une analyse toutes les 10 ms.

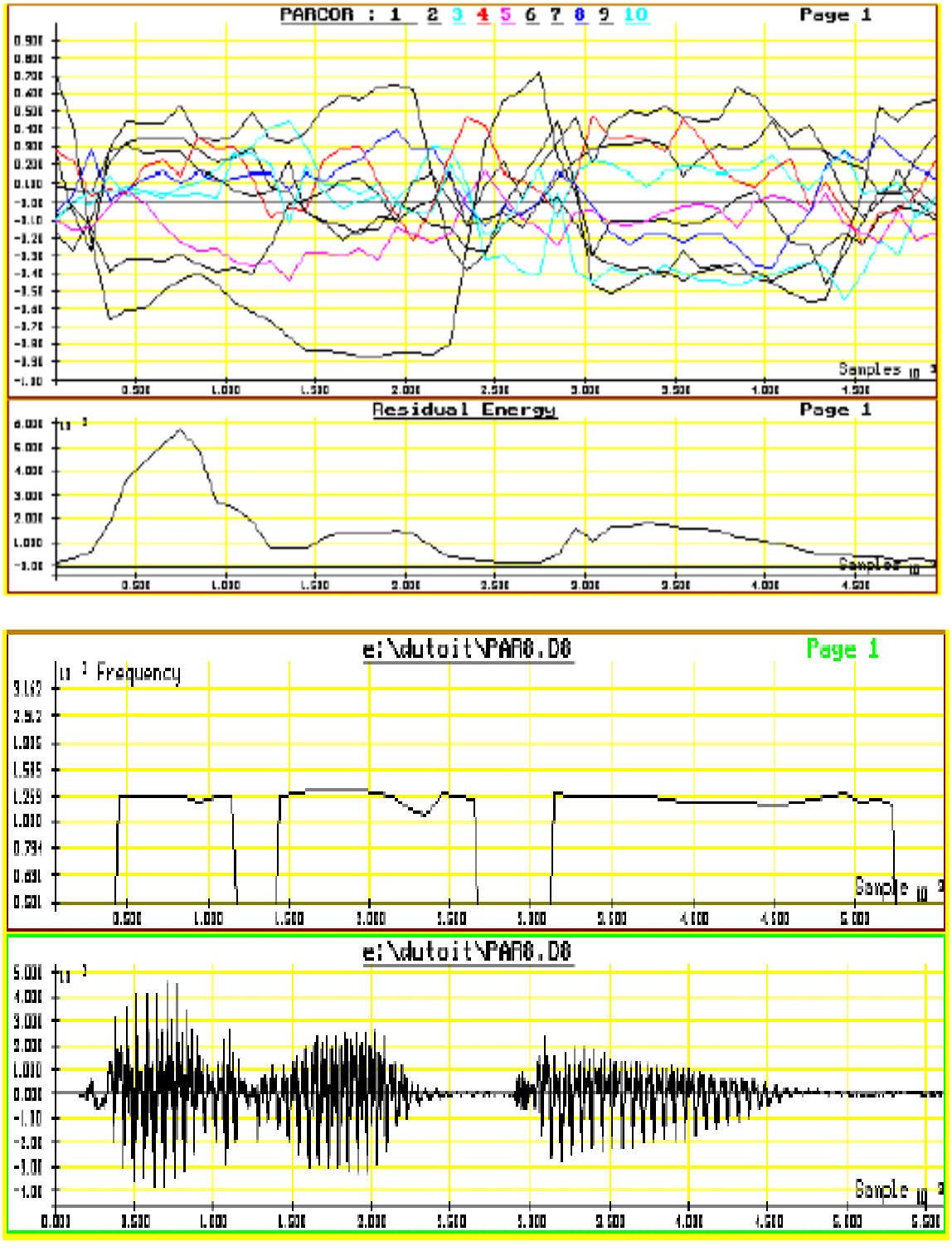


FIG 2.16 *Analyse LPC du mot 'parenthèse'. De haut en bas : les coefficients PARCOR, le gain s, et le pitch P. La décision V/NV apparaît implicitement dans la représentation du pitch*

4 CONCLUSION

La parole est la faculté de communiquer la pensée par un système de sons articulés ; c'est le moyen de communication privilégie entre les humains qui sont les seuls être vivants à utiliser un tel système structuré. Il nous a paru nécessaire de donner avant tout un exposé condensé mais rigoureux sur les formalismes indispensables à la bonne compréhension du traitement de la parole. Il s'agit de faire un rappel théorique concis mais suffisamment complet, à savoir, une brève description de l'appareil phonatoire et auditif humain qui permettront de mieux comprendre les phénomènes de la production de la parole. On se basera surtout sur l'aspect acoustique et phonétique, pour présenter un modèle Auto – Régressif simplifié de l'organe phonatoire, pourront justifier l'utilisation des différentes méthodes de modélisation paramétrique du signal vocal, en vu des aspects d'analyse et de la reconnaissance de parole qui seront abordés dans le chapitre suivant.

Chapitre 3

Reconnaissance de la Parole

Le chapitre 3 décrit l'architecture générale du système de reconnaissance de la parole ainsi que les différents éléments qui le compose. Nous présentons ensuite, les caractéristiques les plus notoires du signal de parole qui peuvent dégrader d'une façon significative la qualité du système de reconnaissance. Nous résumerons par la suite, les approches de reconnaissance les plus performantes qui peuvent être employées dans le système de reconnaissance de la parole. Nous mettons en évidence après, les propriétés de quelques types de coefficients actuellement utilisés pour l'extraction de paramètres effectuée sur le signal vocal avant d'entamer le processus de reconnaissance. Nous décrivons par la suite, le principe des algorithmes de quantification vectorielle qui sont les plus utilisés dans le traitement de la parole. Dans la section qui suit, nous présentons, les modèles les plus utilisés en reconnaissance de la parole : Les modèles de Markov cachés (HMM) et nous terminons par examiner quelques types de systèmes déjà connus de reconnaissance en mots isolés et grand vocabulaire.

1 INTRODUCTION

La conception d'un système de reconnaissance automatique de la parole est rendue difficile par la complexité du signal de parole. La production de la parole est un processus continu, et l'identification univoque d'unités symboliques dans ce flux n'est pas toujours possible. Différentes approches ont été développées pour réaliser la reconnaissance de parole. Les méthodes les plus performantes actuellement sont des méthodes statistiques utilisant le formalisme des modèles de Markov cachés. Elles rendent concevable la reconnaissance de parole continue à grand vocabulaire et indépendamment du locuteur. La figure 3.1 présente les différentes étapes dans les processus d'entraînement et de reconnaissance d'un système de reconnaissance. La ligne pointillée marque la séparation entre les deux processus. Lors de l'entraînement, il est nécessaire d'entraîner des modèles statistiques particuliers, les modèles de Markov cachés HMM [BAU72] ; [BAK76] ; [BAH83] ; [BOU94] ; [JEL76] ; [RAB86] ; [RAB93] sur des données acoustiques. Outre les données acoustiques, l'entraînement de ces modèles nécessite une définition précise des unités lexicales de base utilisées et un dictionnaire décrivant la liste des mots qui pourront être reconnus. Lors de la reconnaissance après extraction des paramètres acoustiques, un décodage est effectué et le système de reconnaissance fournit en sortie le son le plus probable étant donnée les modèles HMM. Chacun des éléments présents sur cette figure sera décrit dans ce chapitre.

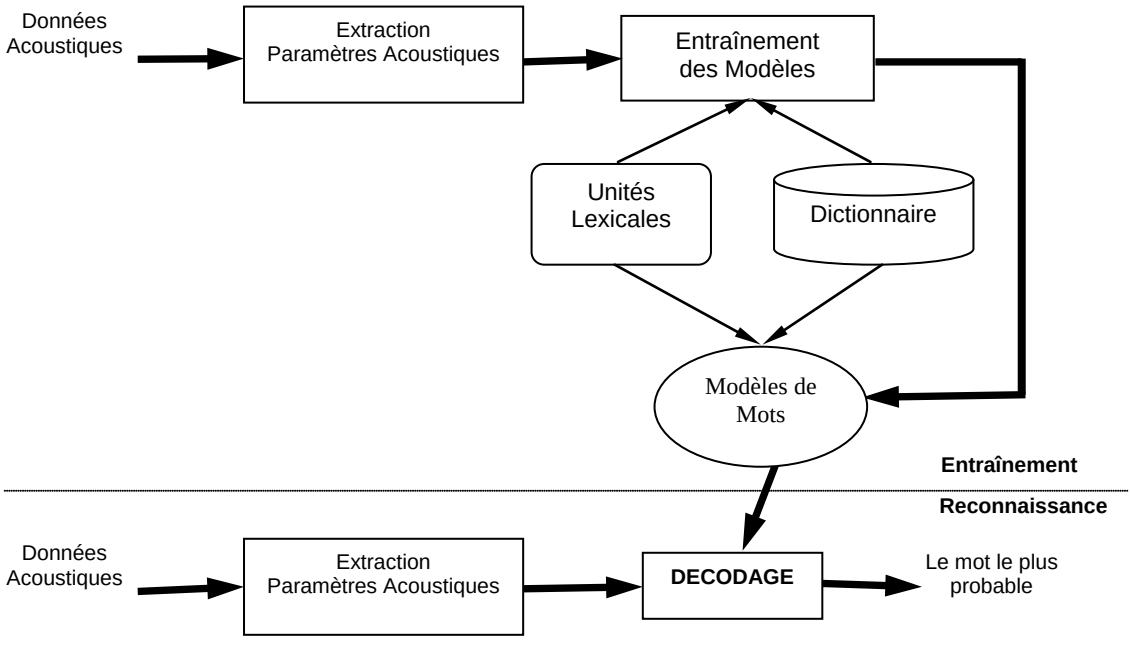


FIG 3.1 *Schéma général d'un système de reconnaissance de la parole*

2 COMPLEXITE DU SIGNAL DE PAROLE

Le signal de parole n'est pas un signal ordinaire; il est le vecteur d'un phénomène extrêmement complexe : la communication parlée. La reconnaissance de la parole pose de nombreux problèmes aux chercheurs depuis 1950. D'un point de vue mathématique, il est difficile de modéliser le signal de parole, car ses propriétés statistiques varient au cours du temps. Nous allons ici tenter de mettre en évidence quelques caractéristiques notoires de ce signal non-stationnaire afin de faire ressortir les problèmes posés lors de son traitement.

2.1 Redondance

Le signal de parole est extrêmement redondant. Cette grande redondance lui confère une robustesse à certains types de bruits. De nombreuses recherches sont menées afin de rendre les systèmes de reconnaissance robustes aux bruits [DUP96] ; mais les performances humaines sont encore loin d'être atteintes. Rappelons que les expériences décrites dans ce travail ont été réalisées sur des corpus de parole claire (sans bruits).

2.2 Variabilité

Le signal de parole possède une très grande variabilité. Une même personne ne prononce jamais un mot deux fois de façon identique. La vitesse d'élocution peut varier, la durée du signal est alors modifiée. Toute altération de l'appareil phonatoire peut modifier la qualité de l'émission (exemple : rhume, fatigue...). De plus, la diction évolue dans le temps. La voix est modifiée au cours des étapes de la vie d'un être humain (enfance, adolescence, âge adulte...). La variabilité interlocuteur est encore plus évidente. La hauteur de la voix, l'intonation, l'accent différent selon le sexe, l'origine sociale, régionale ou nationale. Enfin, la parole est un moyen de communication où de nombreux éléments entrent en jeu, tels le lieu, l'émotion du locuteur, la relation qui s'établit entre les locuteurs (stressante ou amicale). Ces facteurs influencent la forme et le contenu du message. L'acoustique du lieu (milieu protégé ou environnement bruité), la qualité du microphone ou de la ligne téléphonique, les bruits de bouche, les hésitations, les mots hors vocabulaire sont autant d'interférences supplémentaires sur le signal de parole que le système doit compenser.

2.3 Effets de coarticulation

La production parfaite d'un son suppose un positionnement précis des organes phonatoires. Le déplacement de ces organes est limité par une certaine inertie mécanique. Les sons émis subissent alors l'influence de ceux qui les précèdent ou les suivent. Ces effets de coarticulation sont des interférences sur le signal de parole. Ils entraînent l'altération des formes sonores en fonction du contexte. L'effet de coarticulation est un facteur de variabilité supplémentaire important du signal de parole.

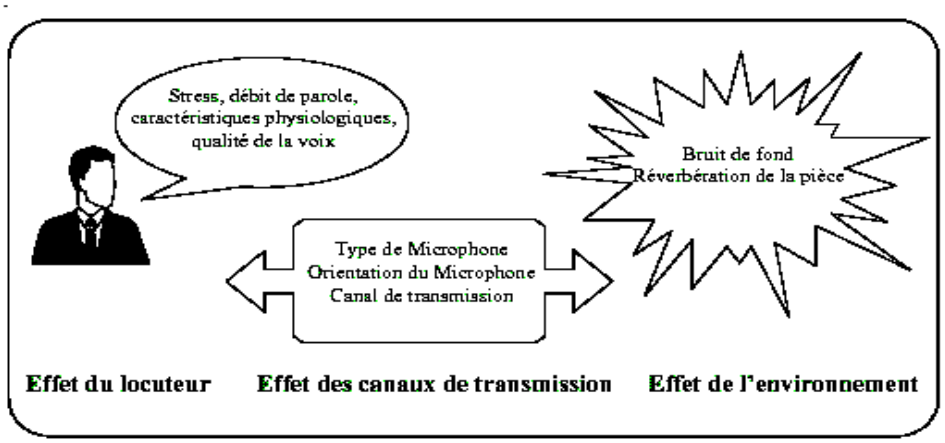


FIG 3.2 *Représentation schématique de différentes sources de variabilités*

3 TACHE DE RECONNAISSANCE

Plutôt que d'affronter simultanément toutes ces difficultés, il est préférable de simplifier le problème de la RAP en se limitant à des sous-problèmes. Les difficultés de mise au point d'un système de reconnaissance de la parole dépendent des conditions d'utilisation du système, qui sont caractérisées par leur degré de liberté, du plus contraint au plus libre, dans les domaines suivants:

- Le nombre d'utilisateurs du système: Celui-ci peut être mono-locuteur, multi- locuteurs ou indépendant du locuteur;

- La taille du vocabulaire: Petit vocabulaire (moins de mille mots), grand vocabulaire (moins de cent mille mots) ou très grand vocabulaire (plus de cent mille mots);

- La complexité du langage utilisé: Langage contraint par une syntaxe artificielle ou langage naturel;

- Le mode d'élocution: Mots isolés ou parole continue;

- La robustesse aux conditions d'enregistrement: Système nécessitant de la parole de bonne qualité ou fonctionnant en milieu bruité.

Avec des méthodes statistiques à base de modèles de Markov cachés, il est concevable de réaliser une reconnaissance de la parole continue indépendamment du locuteur, en grand vocabulaire, pour un enregistrement de bonne qualité et un langage artificiel [LEE88]. En qualité téléphonique, les performances ne permettent pour le moment que des applications avec de petits vocabulaires [GAG90].

Enfin, la "machine à dicter" doit permettre la reconnaissance de parole continue en langage naturel et pour tout locuteur. Cet objectif encore ambitieux fait l'objet d'appels d'offres de la part d'organismes officiels (par exemple de l'AUPELF-UREF en 1994). Le traitement de très grands vocabulaires impose l'utilisation d'unités acoustiques sub-lexicales, et la définition d'un modèle de langage. Différentes méthodes de reconnaissance peuvent être employées, les plus performantes étant les méthodes statistiques.

4 METHODES DE RECONNAISSANCE

On distingue usuellement en reconnaissance de la parole l'approche analytique et l'approche globale. La première approche cherche à traiter la parole continue en décomposant le problème, le plus souvent en procédant à un décodage acoustico-phonétique exploité par des modules de niveau linguistique. La seconde consiste à identifier globalement un mot ou une phrase en les comparant avec des références enregistrées. La distinction entre global et analytique a perdu de sa pertinence avec l'introduction des méthodes statistiques à base de modèles de Markov pour la reconnaissance de la parole continue et le traitement de grands vocabulaires; il s'agit de méthodes globales qui peuvent exploiter des unités acoustiques sub-lexicales.

4.1 Approche analytique

L'approche analytique cherche à résoudre le problème de la parole continue en isolant des unités acoustiques courtes comme les phonèmes, les diphonèmes ou les syllabes. Un exemple classique de cette approche est l'analyse par traits: des indices acoustiques sont calculés à partir du signal de parole; ils permettent de faire des hypothèses locales sur certains traits phonétiques, comme le voisement, la nasalisation, le lieu d'articulation ou le degré d'ouverture du conduit vocal. En fonction de ces traits, le signal acoustique est segmenté et une identification phonétique des segments est réalisée. Le décodage acoustico-phonétique ainsi obtenu est exploité par des modules d'ordre linguistique. Les niveaux lexical, syntaxique ou sémantique utilisent des sources de connaissances spécialisées et sont organisés avec le module acoustique dans des architectures montantes ou descendantes [HAT91]. Les systèmes analytiques, conçus avec des objectifs ambitieux, sont restés au stade expérimental. Leur faiblesse provient d'un processus de décision trop précoce, à savoir une segmentation préalable à l'identification ou une identification phonétique sans prise en compte des niveaux linguistiques. Les méthodes globales, développées pour la reconnaissance de mots isolés, ne font pas d'hypothèse sur la structure phonétique des mots, ce qui évite une erreur pénalisante au début du traitement.

4.2 Approche globale

Les méthodes globales identifient un mot ou une phrase en les considérant comme des entités élémentaires et en les comparant avec des références enregistrées. Leur essor en reconnaissance de parole est dû à l'exploitation de critères de comparaison performants, comme l'alignement temporel dynamique des formes acoustiques, et à leur application à des représentations adaptées du signal, qu'il s'agisse de l'analyse spectrale ou de la prédiction linéaire.

Disposant d'une représentation du signal de parole, la reconnaissance de mots isolés est un problème classique de reconnaissance des formes. L'ensemble des n_m mots du vocabulaire est noté $E_m = \{m_k\}_{1 \le k \le n_m}$ et chaque mots m_k est représenté par une ou plusieurs formes acoustiques de référence R_{mk} , par exemple les paramètres spectraux calculés de manière périodique sur le signal. Une forme de test observée O, qui est la suite des spectres d'un mot inconnu, est comparée à chacune des références. Le mot inconnu est identifié au mot de référence $\tilde{m}$ dont il est le plus proche au sens d'une certaine distance D:

$$\tilde{m} = \arg \min_{m \in E_m} D(O, R_m) \tag{3.1}$$

Le calcul de la distance nécessite la mise en correspondance d'une forme de référence et de la forme inconnue. Or, la durée d'un même mot est variable d'une prononciation à l'autre, et de plus les déformations ne sont pas linéaires en fonction du temps. La distance D est donc calculée sur l'alignement temporel qui rapproche le mieux les deux formes. Mais une recherche exhaustive de toutes les déformations possibles est exclue en raison de l'explosion combinatoire.

L'alignement temporel dynamique (Dynamic Time Warping ou DTW en anglais) résout efficacement ce problème, en exploitant le principe d'optimalité de Bellman [BEL57]. La construction de l'alignement optimal entre les formes de référence et de test est réalisée par récursivité sur l'indice du temps, en exploitant le fait que le chemin optimal est l'extension d'un sous-chemin lui-même optimal. La complexité de l'alignement est de ce fait considérablement réduite puisqu'elle passe d'un ordre exponentiel à un ordre polynomial. Les premières applications de la programmation dynamique en parole sont développées en URSS en 1968 [VIN68] ; [VEL70], puis au Japon à partir de 1970 [SAK78]. La méthode est efficace pour la reconnaissance mono-locuteur à petit vocabulaire et en mots isolés. Des extensions de l'alignement temporel dynamique ont été proposées pour la reconnaissance indépendante du locuteur [RAB79] ou la reconnaissance de mots enchaînés [BRI82] ; [MYE81] ; [SAK79]. Cependant, l'approche statistique propose un formalisme plus général et permet la reconnaissance de grands vocabulaires en parole continue de manière plus efficace que par DTW en intégrant la modélisation des niveaux linguistiques.

4.3 Approche statistique

F. Jelinek a proposé une formalisation statistique simple issue de la théorie de l'information et qui est aujourd'hui classique pour décomposer le problème de la reconnaissance de la parole continue [JEL76] . Soit O une suite d'observations acoustiques, et M une suite de mots prononcés. Connaissant les observations O, on cherche la suite de mots $\tilde{M}$ la plus probable parmi toutes les suites possibles $E_M = E_m^*$.

L'approche statistique permet ainsi d'intégrer les niveaux acoustiques et linguistiques dans un seul processus de décision. Ces niveaux sont classiquement représentés par des modèles de Markov cachés (Hidden Markov Models ou HMM). Les unités acoustiques

modélisées peuvent être des mots comme dans l'approche globale ou des unités plus courtes telles que le phonème comme dans l'approche analytique. La modélisation markovienne est plus générale que l'alignement temporel dynamique et tient compte non seulement de la non linéarité temporelle du processus mais aussi de la variabilité acoustique de la production de la parole. Son application à la reconnaissance de la parole continue a été rendue possible par l'augmentation continue de la puissance des ordinateurs et de la taille des bases de données disponibles.

Les chercheurs de CMU (Carnegie Mellon University) et d'IBM sont les premiers à avoir introduit le formalisme des modèles de Markov cachés en reconnaissance de la parole [BAK74] ; [KLA77]. Au cours des dix dernières années, les systèmes des plus grands laboratoires internationaux travaillant en RAP, comme les systèmes SPHINX de CMU [LEE88], BYBLOS de BBN (Bolt Beranek and Newman Inc.) [CHO87], TANGORA d'IBM [AVE87], ou ceux développés à AT&T [WIL93], ont été conçus avec une approche statistique markovienne. Cette approche a aussi été appliquée avec succès en France au CNET ou au LIMSI qui obtient des performances équivalentes à celles des meilleurs systèmes actuels [GAU94].

5 EXTRACTION DES PARAMETRES

Pour résoudre les problèmes liés à la complexité de la parole, il est possible de calculer des coefficients représentatifs du signal traité. Ces coefficients sont calculés à intervalles temporels réguliers. En simplifiant les choses, le signal de parole est transformé en une série de vecteurs de coefficients.

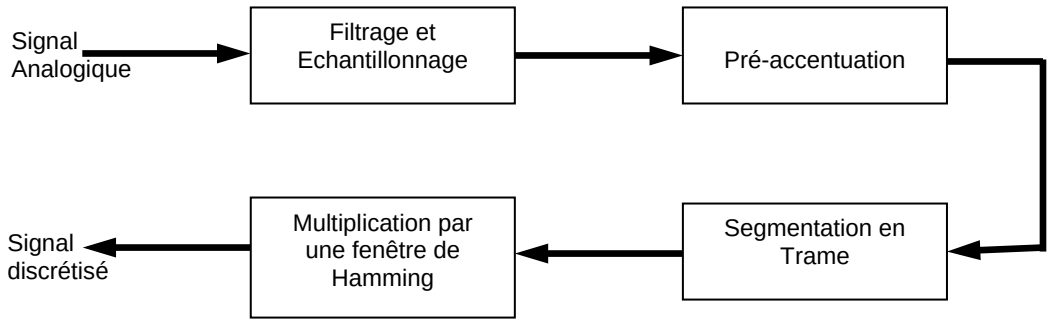


FIG 3.3 *Mise en forme du signal*

Ces coefficients doivent représenter au mieux le signal qu'ils sont censés modéliser, et extraire le maximum d'informations utiles pour la reconnaissance. Nous étudierons dans ce paragraphe les coefficients les plus utilisés en reconnaissance de la parole. Nous commencerons par les coefficients cepstraux aussi appelés cepstres. Par la suite, nous mettrons en évidence les propriétés d'autres coefficients tels que les coefficients PLP et RASTA-PLP. Nous parlerons enfin des coefficients MFCC, BACC, CMS et LDA.

Avant tout calcul, il est nécessaire de mettre en forme le signal de parole. Pour cela, quelques opérations sont effectuées avant tout traitement. La figure 3.3 illustre l'ensemble de ces opérations. Le signal est tout d'abord filtré puis échantillonné à une

fréquence donnée. Une pré-accentuation est effectuée afin de relever les hautes fréquences. Puis le signal est segmenté en trames. Chaque trame est constituée d'un nombre N fixe d'échantillons de parole. En général, N est fixé de telle manière que chaque trame corresponde à environ 30 ms de parole (durée pendant laquelle la parole peut être considérée comme stationnaire). Enfin, une multiplication par une fenêtre de "Hamming" [1] est effectuée, afin de réduire les effets de bords. Ce traitement implique une hypothèse importante du fait des limitations postérieures qu'elle occasionne :

Le signal vocal est supposé stationnaire sur une courte période.

Après cette mise en forme du signal (commune à la plupart des méthodes d'analyse de la parole), une transformée de Fourier (DFT –Transformée de Fourier Discrète, en particulier FFT – Transformée de Fourier Rapide) est appliquée pour passer dans le domaine fréquentiel. Le spectre de puissance à court terme P (w) est calculé selon :

$$P(w) = Re[X(w)]^2 + Im[X(w)]^2 = |X(w)|^2 \qquad (3.2)$$

X(w) est le spectre du signal temporel et w représente la fréquence angulaire en $rad.s^{-1}$.

Ensuite, chaque méthode et type de coefficients ont des particularités que nous nous proposons de décrire dans les paragraphes suivants.

5.1 Coefficients cepstraux

Il existe deux méthodes de calcul des coefficients cepstraux couramment utilisées en reconnaissance de la parole :

- Analyse spectrale.
- Analyse paramétrique.

5.1.1 Analyse spectrale

Le signal de parole est modélisé en utilisant le spectre. Il est possible de calculer directement les coefficients cepstraux en suivant le processus décrit à la figure 3.4.

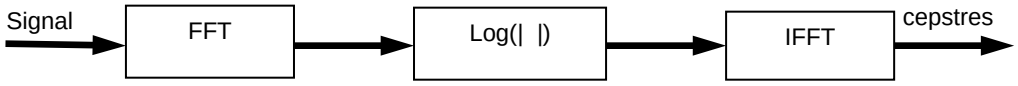


FIG 3.4 *Calcul des cepstres par analyse spectrale*

1. $w(n) = 0.54 + 0.46.\cos(2\pi\frac{n}{N-1})$

Outre ce type de fenêtre couramment utilisée en reconnaissance de la parole. Il existe plusieurs autre types de fenêtre telle que rectangulaire, Hanning, Blackman, … La fenêtre de Hamming est un cas particulier de la fenêtre de Hanning.

mieux à la perception humaine. Une de ces transformations perceptuelles se fait selon Il suffit pour obtenir les coefficients cepstraux de prendre le spectre de puissance à court terme défini au début du paragraphe (c'est à dire la FFT – Transformée de Fourier Rapide du signal, et sa norme[2]), puis de passer dans le domaine logarithmique et enfin de prendre la fonction inverse FFT – IFFT. Ce type de calcul des cepstres est généralement moins utilisé que la méthode paramétrique décrite ci-après du fait de la charge de calcul importante associée au calcul de la FFT et de la FFT inverse. L'échelle de Mel[3] [DAV80] et les coefficients ainsi obtenus sont appelés MFCC (Mel-scale Frequency Cepstral Coefficients).

La sélectivité de l'oreille diminue avec l'accroissement des fréquences. Par analogie, on utilise généralement des bancs de filtres dont la répartition reproduit cette sélectivité (échelles de Mel). Pour ce type de banc de filtres, la largeur de chacun d'eux augmente avec la fréquence centrale.

Ceci permet, avec peu de filtres et tout en couvrant la bande passante, d'obtenir une bonne résolution dans les basses fréquences (là où se trouvent les premiers formants) et de garder suffisamment d'informations dans les hautes fréquences. Cette dernière représentation est actuellement considérée comme une des plus performantes en reconnaissance de la parole (cf. Figure 3.5). Certains chercheurs calculent des paramètres BACC (Bark Auditory Cepstral Coefficients) en échelle Bark; mais les différences entre les deux échelles sont peu importantes.

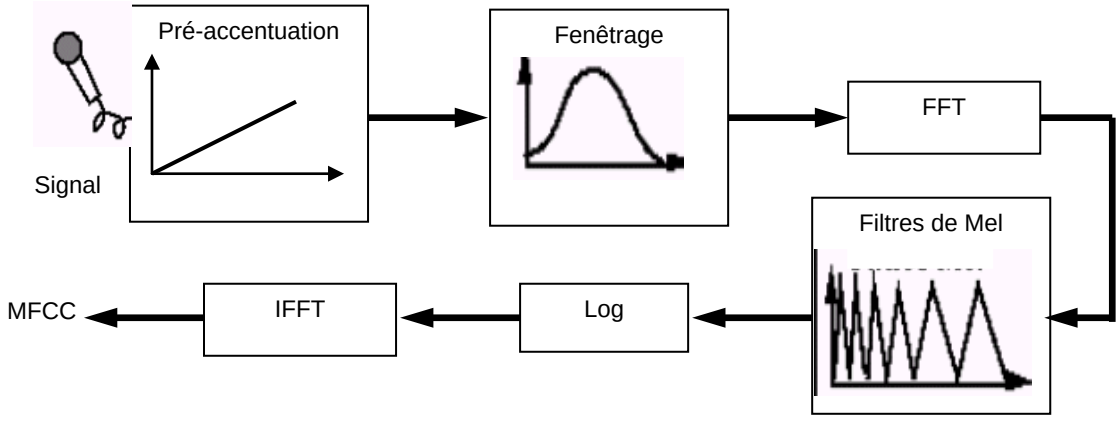


FIG 3.5 *Chaîne d'analyse du signal produisant les coefficients MFCC*

5.1.2 Analyse paramétrique

Cette analyse permet de déterminer les coefficients cepstraux des signaux de la parole. L'étude du mécanisme de phonation montre qu'il est possible de modéliser le signal vocal par le passage d'un signal d'excitation u(n) au travers d'un filtre dont la transmittance est $\sigma / A(z)$.

2 Seul le module de la FFT est retenu, la phase de la transformée de Fourier du signal de parole ne contient pas d'information utilisables pour la reconnaissance de la parole.

3 Echelle de Mel : $Mel(f) = b.\log_{10}(1 + f/c)$ avec b=2600 et c=700, f représente la fréquence.

Cette hypothèse implique que le signal vocal peut alors être considéré comme un signal auto régressif. On appelle signal auto-régressif d'ordre p, un signal qui exprime qu'un échantillon quelconque x(n) est une combinaison linéaire des p échantillons qui le précèdent plus un terme d'excitation [BOI87].

$$x(n) + \sum_{i=1}^{P} a(i).x(n-i) = \sigma.u(n) \qquad (3.3)$$

Ainsi le signal vocal peut alors être considéré comme engendré par un filtre dont il faut trouver les coefficients a(i). Ce modèle de production du signal de parole est appelé AR (auto-régressif).

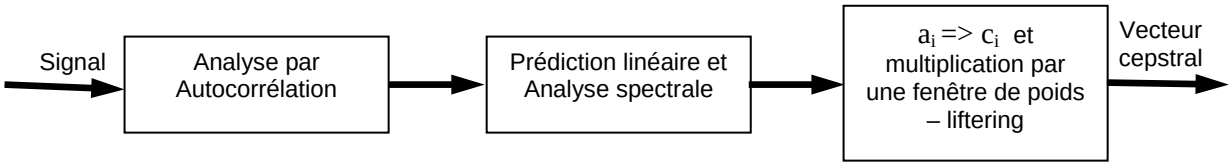


FIG 3.6 *Processus de calcul des cepstres par la méthode de prédiction linéaire*

L'obtention des coefficients a(i) du modèle AR se fait en utilisant des méthodes classiques de prédiction linéaire[4] (p désigne toujours l'ordre du filtre). La méthode de calcul des cepstres est basée sur une relation de récurrence liant les cepstres et les coefficients a(i) (cf. Figure 3.6) [ATA91]. Ensuite un liftrage "liftering" est effectué pour augmenter la robustesse des coefficients cepstraux [RAB93]. Ce liftrage consiste en une multiplication par une fenêtre de poids (cf. Formule 3.4) des coefficients cepstraux augmentant l'amplitude des coefficients connus pour être moins sensibles au canal de transmission et au locuteur[5].

$$\forall n \in [1,Q] \quad W(n) = 1 + \frac{Q}{2}.\sin\left(\frac{\pi.n}{Q}\right) \qquad (3.4)$$

Où Q est le nombre de coefficients. Cette méthode de prédiction linéaire est beaucoup plus utilisée en reconnaissance de la parole que celle de l'analyse spectrale. En effet, il existe de nombreux algorithmes (Schur, ou Leroux et Gueguen) qui peuvent être utilisés en temps réel, alors que l'analyse spectrale demande beaucoup plus de temps de calcul (calcul d'une FFT et d'une FFT inverse). Ainsi, il existe les paramètres LPCC (Linear Prediction Cepstral Coefficients) qui sont calculés à partir d'une modélisation auto-régressive du signal.

4 Analyse LPC, méthode de l'autocorrélation ou de la covariance (cf. § 2.3).

5 Ce type de technique s'accompagne d'une amélioration sensible des taux de reconnaissance pour les HMM à distributions discrètes.

Si un modèle auto-régressif A = (1, $a_1 ... a_p$)^t d'ordre p a été estimé sur une trame du signal, les d premiers coefficients cepstraux C_k sont obtenus par:

$$C_k = a_k - \sum_{i=1}^{k-1} \frac{i}{k} C_i a_{k-i}, \quad 1 \le k \le d \tag{3.5}$$

Ces coefficients sont utilisés à AT&T [RAB89], et à CMU [LEE89].

Il existe un autre modèle de production (transmittance du filtre de la forme : $\sigma.C(z)/A(z)$. Le modèle défini par cette transmittance est appelé ARMA (Auto Régressive Moving Average) pour lequel chaque échantillon est constitué par une combinaison linéaire de p échantillons passés et de (q+1) échantillons présents et passés de l'excitation u.

$$x(n) + \sum_{i=1}^{p} a(i).x(n-i) = \sigma. \sum_{i=0}^{q} c(i).u(n-i) \tag{3.6}$$

L'estimation des modèles ARMA étant plus délicate que celle des modèles AR, on préfère utiliser les modèles auto-régressifs dans la plupart des applications. Ainsi pour l'analyse spectrale comme pour l'analyse paramétrique, le signal de parole a été transformé en une série de vecteurs cepstraux calculés pour chaque trame (correspondant à un nombre d'échantillons du signal d'entrée). Ces coefficients jouent un rôle capital dans les méthodes utilisées pour reconnaître la parole.

5.2 Soustraction cepstrale

La parole peut être modélisée par la convolution d'un signal de parole x(n) et d'un terme relatif au canal de transmission h(n). Soit en notant * l'opérateur convolution :

$$y(n) = x(n) * h(n) \tag{3.7}$$

Dans ce cas, la transformée de Fourier discrète du signal y(n) s'écrit[6] :

$$Y(f) = FFT(y(n)) = X(f).H(f) \tag{3.8}$$

Ce qui peut encore s'écrire en prenant la norme et le log :

$$\ln|Y(f)| = \ln|X(f)| + \ln|H(f)| \tag{3.9}$$

6 Avec $f = n.f_e/N$ pour $n \in [0, N-1]$, f_e est la fréquence d'échantillonnage.

Reprenons maintenant la transformée de Fourier inverse pour calculer les coefficients cepstraux, il vient alors :

$$\tilde{y}(n) = \tilde{x}(n) + \tilde{h}(n) \tag{3.10}$$

Ce terme $\tilde{h}(n)$ est souvent appelé bruit convolutif ou distorsion du canal de transmission et influence fortement le processus de reconnaissance (puisqu'il a une influence sur les paramètres acoustiques). La soustraction cepstrale[7] (Cepstral Mean Substraction [FUR86]) est en fait une technique de normalisation des paramètres acoustiques afin de les rendre moins sensibles à des facteurs extérieurs (bruits convolutifs).

Cette technique de normalisation consiste à calculer pour chaque locuteur (ou pour chaque phrase) la moyenne des vecteurs acoustiques et à effectuer la soustraction de cette moyenne pour obtenir de nouveaux paramètres acoustiques. On élimine ainsi l'influence de ce fameux bruit convolutif $\tilde{h}(n)$. Il va de soi que les taux de reconnaissance généralement observés avec cette technique de normalisation sont largement supérieurs à ceux correspondant aux vecteurs acoustiques classiques. En effet, on utilise une information qui n'est a priori pas accessible pour un système de démonstration (la moyenne des vecteurs acoustiques pour un locuteur ou une phrase donnée).

Il est possible de calculer cette moyenne des vecteurs acoustiques en ligne, c'est-à-dire au fur et à mesure de l'arrivée des vecteurs acoustiques en initialisant cette moyenne à la moyenne globale des vecteurs acoustiques sur la base d'entraînement, et en recalculant cette moyenne pour chaque nouveau vecteur acoustique. Les taux de reconnaissance observés sont alors moins bons que pour le type de normalisation décrit précédemment, mais ce type de technique peut être utilisé pour un système de démonstration. Néanmoins, il est nécessaire d'attendre quelques secondes de parole avant que le processus de normalisation ne se stabilise. Les résultats obtenus avec le type de normalisation "offline" des paramètres acoustiques sont largement supérieurs à ceux obtenus avec les paramètres acoustiques classiques[8] , mais ils ne reflètent pas les réelles performances du système. De plus cette technique de normalisation produit une distorsion en présence de bruit additif.

5.3 Coefficients PLP (Perceptual Linear Predictive)

L'estimation par prédiction linéaire (LP) des coefficients du modèle auto- régressif (tout pôle) du spectre de la parole est très utilisée en reconnaissance de la parole. Néanmoins, un des principaux défauts de cette méthode est que le filtre « identifie » uniformément le spectre sur toutes les fréquences de la bande d'analyse. Or cette propriété est loin d'être vérifiée pour l'oreille humaine, car il a été établi que l'oreille humaine est plus sensible aux fréquences situées au milieu de la bande d'analyse du spectre (100 à 3000 Hz). Ainsi, il est possible que certains détails importants du spectre ne soient pas pris en compte lors de l'analyse LP. L'analyse PLP permet de résoudre ce problème [HER90].

---

7 Notée CMS dans le reste du document.
8 Notamment pour des canaux de transmission du type lignes téléphoniques.

Le but de cette analyse est d'estimer les paramètres d'un filtre auto-régressif tout pôle, modélisant au mieux le spectre auditif [HER91]. Le schéma général de cette méthode est visible sur la figure 3.7. L'application de l'analyse PLP sur une trame d'analyse est schématisée dans la figure 3.8.

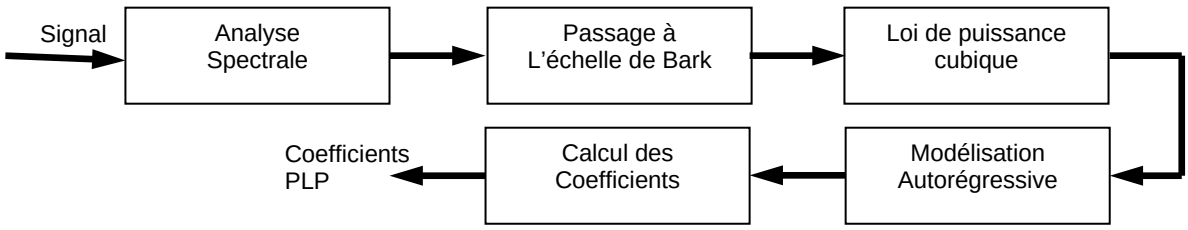


FIG 3.7 *Méthode de calcul des coefficients PLP*

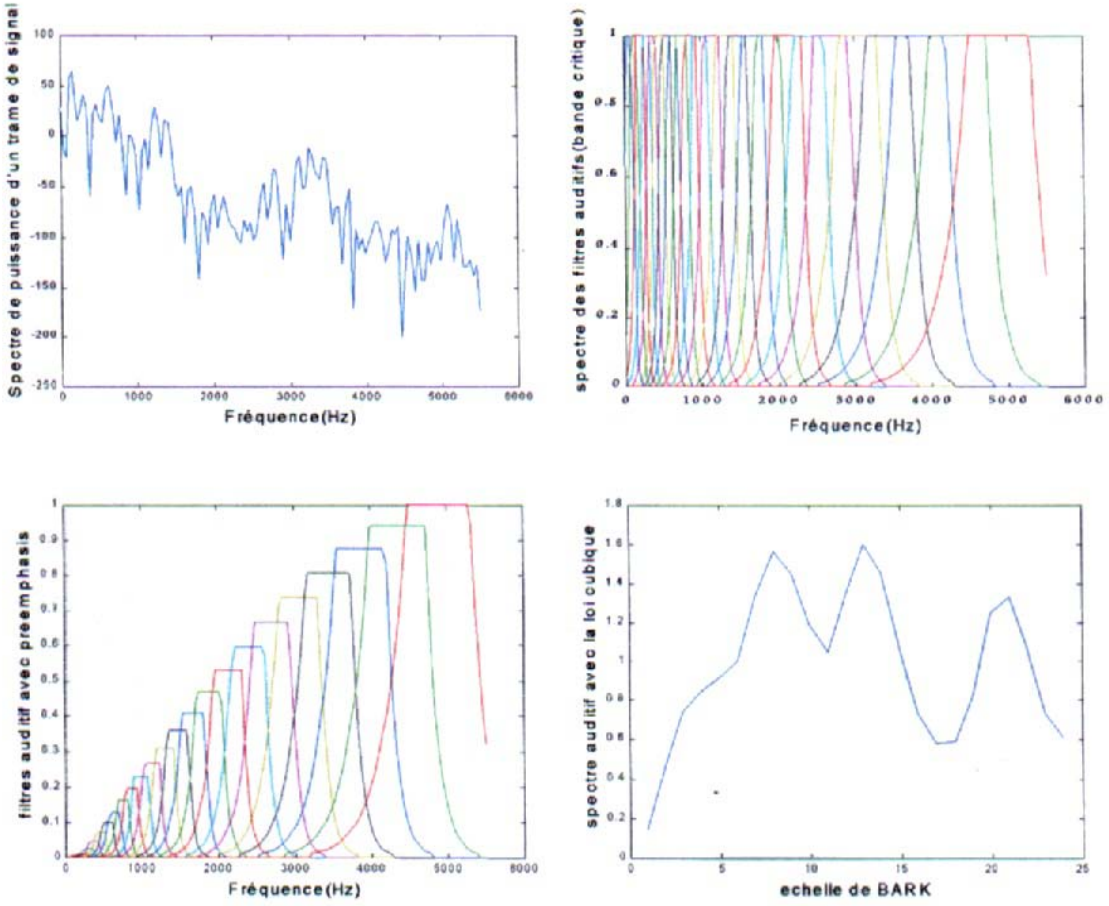


FIG 3.8 *Application de l'analyse PLP sur une trame d'analyse*

Observons plus en détail les différents blocs utilisés pour calculer les coefficients PLP. Comme défini auparavant, le spectre de puissance court terme P(w) est calculé. Ensuite, un passage de l'échelle de fréquence usuelle à l'échelle de Bark est effectué. Ce passage à l'échelle de Bark, permet d'approximer de manière grossière ce que nous savons de la forme des filtres auditifs. La forme des filtres auditifs est approximativement constante le long de l'échelle de Bark. Le spectre d'énergie court terme dans l'échelle de Bark est convolué avec le spectre de puissance de la courbe de bande critique. w représente la fréquence angulaire en rad.s^{-1} et Ω la fréquence de Bark[9]. Cette transformation de l'échelle fréquentielle a été mise au point par *Schroeder* [SCH77].

$$\Theta(\Omega_t) = \sum_{\Omega=-1.3}^{\Omega=2.3} P(\Omega - \Omega_t).\Psi(\Omega) \tag{3.11}$$

où $\Psi(\Omega)$ est la courbe de masquage :

$$\Psi(\Omega) = \begin{cases} 0 & \text{si } \Omega \prec -1.3 \\ 10^{2.5(\Omega+0.5)} & \text{si } -1.3 \le \Omega \le -0.5 \\ 1 & \text{si } -0.5 \le \Omega \le 0.5 \\ 10^{-1.0(\Omega-0.5} & \text{si } 0.5 \le \Omega \le 2.5 \\ 0 & \text{si } \Omega \succ 2.5 \end{cases}$$

On essaye ensuite d'approximer la sensibilité de l'oreille humaine à différentes fréquences par l'intermédiaire d'une fonction de transfert E(w). Le spectre de puissance est multiplié par cette fonction de transfert.

$$E(\Omega) = E(w).\Theta(\Omega) \tag{3.12}$$

La non-linéarité entre l'intensité d'un son et sa force de perception par l'oreille est ensuite approximée par une loi de puissance :

$$\Gamma(\Omega) = E(\Omega)^{1/3}$$

Enfin, on effectue une modélisation auto-régressive classique du spectre du modèle auditif tout pôle, en calculant les coefficients auto-régressifs du filtre. Hermansky et al. [HER90] ont montré que les coefficients PLP introduisent une meilleure robustesse au bruit que les coefficients cepstraux classiques. Une extension de cette méthode a été développée et a donné de bons résultats : les coefficients RASTA-PLP [HER94]. Le filtrage RASTA (RelAtive SpecTrAl processing) consiste à supprimer les facteurs constants dans chaque composante spectrale du spectre auditif court terme avant l'estimation du modèle tout pôle.

---

[9] $\Omega(w) = 6.\ln\left[\frac{w}{1200.\pi} + \left(\left(\frac{w}{1200.\pi}\right)^2 + 1\right)^{1/2}\right]$.

En effet, supposons que la parole soit corrompue par un bruit convolutif (exemple d'un changement des caractéristiques fréquentielles d'un canal de transmission causé par l'utilisation d'un nouveau microphone).
Une telle distorsion apparaît sous la forme d'une constante additive dans le logarithme du spectre de la parole. La plupart des paramètres acoustiques et donc des systèmes de reconnaissance sont affectés par de tels phénomènes.

Le filtrage RASTA supprime les composantes spectrales (ou log spectrales) qui varient plus lentement ou plus rapidement que le signal de parole et donc leurs influences sur les systèmes de reconnaissance de la parole. Une extension de cet algorithme nommée Adaptative lin-log RASTA ou J-RASTA permet d'effectuer une compensation pour les bruits à la fois convolutifs et additifs. Une transformation non linéaire dans le domaine spectrale est effectuée :

$$Y(f) = \ln[1 + J.X(f)] \tag{3.13}$$

où J est une constante dépendante du rapport signal sur bruit (SNR). Cette transformation est du type linéaire pour de petites valeurs de J et logarithmique pour des valeurs élevées. Ce coefficient J peut être ajusté en le rendant inversement proportionnel à l'énergie moyenne du bruit. En utilisant ce type d'ajustement, Hermansky [HER94] a montré que les résultats obtenus avec ce type de coefficients sont comparables à ceux obtenus avec un système entraîné et testé dans les mêmes conditions de bruits (pour peu qu'il y ait suffisamment de données pour estimer l'énergie du bruit). Le filtrage log-RASTA étant moins sensible aux variations du canal de transmission (changement de microphone), il est le plus souvent à la base des systèmes de démonstration, c'est pourquoi nous l'utiliserons très souvent dans ce travail.

5.4 Coefficients LDA

Plusieurs groupes de chercheurs [AUB93] ; [HAE92] ; [SIO95] ont montré que l'analyse discriminante linéaire permettait d'améliorer les performances des systèmes de reconnaissance ainsi que leur robustesse à certains types de bruits. L'objet de l'analyse discriminante linéaire est de déterminer des paramètres adaptés au problème de classification. Les paramètres discriminants sont obtenus en calculant une transformation linéaire des vecteurs d'entrée x (de dimension n) vers des vecteurs y (de dimension m, m < n) de telle manière que la séparation des classes auxquelles sont assignés ces vecteurs lors de la classification soit maximale. Les critères d'optimisation le plus souvent rencontrés sont :

$$J_1 = \mathrm{tr}(S_2^{-1} S_1) \tag{3.14}$$

$$J_2 = \det(S_2^{-1} S_1) \tag{3.15}$$

Où tr(A) représente la trace de la matrice A et det(A) son déterminant. Les matrices S_1 et S_2 sont deux matrices choisies parmi les matrices de covariance à l'intérieur des classes, entre les classes et mélange de classes [KLA77]. On peut montrer que l'optimisation des critères J_1 et J_2 conduit aux même paramètres discriminants et que cette optimisation est indépendante du choix des matrices de variance. Fontaine & *al.*

ont aussi montré dans [FON97] qu'il était possible d'utiliser des réseaux de neurones pour effectuer une analyse discriminante non linéaire. Les expériences préliminaires ont effectivement montré que les réseaux étaient aptes à fournir des paramètres discriminants très efficaces pour les tâches de reconnaissance. Ce type de coefficients ne sera pas utilisé pour les expériences présentées dans cette étude.

5.5 Etude comparative de représentations

L'étude classique de *S. Davis* et *P. Mermelstein* [DAV80] compare plusieurs représentations du signal: cepstre en sortie d'un banc de filtres en échelle Mel (MFCC) ou en échelle linéaire (LFCC), coefficients de prédiction linéaire (LPC) ou de réflexion (RC), cepstre calculé à partir des coefficients auto-régressifs (LPCC). Ces représentations sont associées à la distance euclidienne ou à la mesure d'Itakura en ce qui concerne les coefficients LPC. Les tests consistent en un alignement temporel dynamique entre des paires de mots minimales.

B. Hanson et *T. Applebaum* s'intéressent aux conséquences de l'effet Lombard et du bruit sur la reconnaissance de mots isolés [HAN90]. Ils remarquent la nette supériorité d'une distance cepstrale pondérée sur une distance euclidienne. De plus, le pré-traitement perceptif des coefficients PLP améliore les résultats de l'analyse par prédiction linéaire. *J.-C. Junqua, H. Wakita* et *H. Hermansky* aboutissent aux mêmes conclusions [JUN93]. *B.Chigier* et *H.Leung* ont utilisé des représentations spectrales et cepstrales classiques ou dérivées du modèles d'audition de *S. Seneff* [CHI92]. Leurs expériences sont réalisées en enregistrement normal avec la base TIMIT et sur du signal de qualité téléphonique avec la base NTIMIT. Elles consistent en une identification phonétique par un classifieur gaussien ou un perceptron multi-couches. Le meilleur résultat est obtenu pour le perceptron avec des coefficients PLP, et pour le classifieur gaussien avec des coefficients en échelle Bark. Les critères de classification testés ne semblent pas adaptés au modèle d'audition de *S. Seneff*.

C. Jankovski, *H. Vo* et *R. Lippmann* ont comparé plusieurs pré-traitements et représentations lors d'expériences de reconnaissance de mots isolés par modèles de Markov pour diverses conditions d'enregistrement [JUA85]. Des coefficients MFCC servent de référence, et sont comparés avec des coefficients LPCC et des pseudo coefficients cepstraux obtenus à partir du modèle d'audition de *S.Seneff* et d'un modèle auditif proposé par *O. Ghitza* [GHI94]. Le critère de comparaison acoustique des modèles est similaire à une distance de Mahalanobis diagonale. En milieu très bruité, les coefficients basés sur des modèles d'audition sont plus performants que les MFCC, mais l'écart devient très faible pour des enregistrements en condition normale. De plus, les coefficients LPCC sont moins performants que les coefficients MFCC, même avec une analyse d'ordre élevée et une transformation bilinéaire permettant de se rapprocher de l'échelle Mel [LEE91]. l'écart devient considérable en milieu bruité. Le tableau 3.1 présente les caractéristiques de quelques systèmes de reconnaissance utilisant ces représentations.

TAB 3.1 *Caractéristiques des modules de pré-traitements pour quelques systèmes de reconnaissance*

Laboratoire	Bp	A	F	P	Analyse	Coefficients	TΔ	Catégorie
	5		36	6	20 Mel	10 MFCC		DTW+dE
NTT	4		32	8	AR	10 LPCC+ 10Δ+ ΔE	72	DTW+dE
ATT	3	0.95	45	15	9 AR	8 LPCC+8 Δ	105	DTW+dM
CUED-REPN	8		32	16	30 Bark	20 BASC		NN
IBM-Tangora	10		26	10	30 Mel	20 MFSC		HMMd
BBN-Byblos	10		20	10	TFD	14 MFCC		HMMd
CMU-Sphinx	8	0.97	20	10	14 AR, BL	12 LPCC+E+13 Δ	40	HMMd
CMU-SphinxII	8	0.97	20	10	14 AR	12 LPCC+12 Δ+ ΔE+13 Δ Δ	40/80	HMMc
ATT	3.8	0.95	30	10	10 AR, L	12 LPCC+12 Δ+ ΔE+13 Δ Δ	50/70	HMMc
ATT	3	Oui	45	15	8 AR, L	12 LPCC+12 Δ+ ΔE+13 Δ Δ	75	HMMc
Panasonic	5		20	10	8 PLP	8 PLP+E+9 Δ	50	HMMc
CUED-HTK	8	Oui	16	10	Mel	12 MFCC+E+13 Δ		HMMc
CUED-HTK95	8				Mel	12 MFCC++13 Δ+13 Δ Δ		HMMc
LIMSI	8		10	10	15 Bark	15 BACC+E+16 Δ+16 Δ Δ		HMMc
CNET-Phil90	4	Oui	16	16	24 Bark	8 MFCC+E+9 Δ+9 Δ Δ		HMMc
ATT	4		10	10	AR	9 LPCC+9 Δ+6 Δ Δ+ ΔE+ Δ ΔE		HMMc

Légende

BP	Largeur de bande d'analyse (kHz)
α	Facteur de pré-accentuation ($1- \alpha z^{-1}$)
F	Taille de la fenêtre d'analyse (ms)
P	Pas d'analyse (ms)
Analyse	9 AR = Analyse auto-régressive d'ordre 9 20 Mel / Bark = Bande de 20 filtres en échelle Mel ou Bark BL = Transformation bilinéaire L = Filtrage cepstral par fenêtre sinusoïdale
Coefficients	LP/MF/BA (Linear Prediction / Mel Frequency / Bark Auditory) CC/SC (Cepstrum / Spectrum Coefficients) PLP (Perceptual Linear Prediction) E (Logarithme de l'énergie) Δ et Δ Δ (Coefficients différentiels du 1^{er} et 2^{eme} ordre)
TΔ	Durée de la régression pour le calcul des coefficients différentiels du 1^{er} ordre (ms)
Catégorie	DTW + dE / dM (Alignement dynamique + distance euclidienne ou de Mahalanobis) NN (Réseaux de neurones) HMMd/sc/c (Modèles de Markov cachés discrets / semi-continus / continus)

6 QUANTIFICATION VECTORIELLE

Lorsque l'on a extrait un jeu assez réduit de paramètres, il peut être intéressant d'utiliser un algorithme de quantification vectorielle. L'idée consiste à stocker des valeurs typiques des vecteurs de paramètres du signal dans un dictionnaire, et de considérer ensuite le signal comme une suite de mots de ce dictionnaire.

On utilise donc un algorithme de clustering sur des données d'apprentissage pour déterminer une partition de l'espace vectoriel considéré (cf. Figure 3.9). Les descriptions de chaque cluster sont alors stockées dans le dictionnaire. Pour coder une nouvelle suite de vecteurs, on remplace chaque vecteur par le numéro du cluster auquel il appartient.

Utiliser un tel codage apporte des simplifications sensibles des algorithmes de reconnaissance. Elles permettent en particulier de transformer des modèles de Markov continus en modèles de Markov discrets. Il existe dans la littérature plusieurs algorithmes de clustering, nous citons dans ce paragraphe les plus utilisés et les plus classiques dans le traitement de la parole.

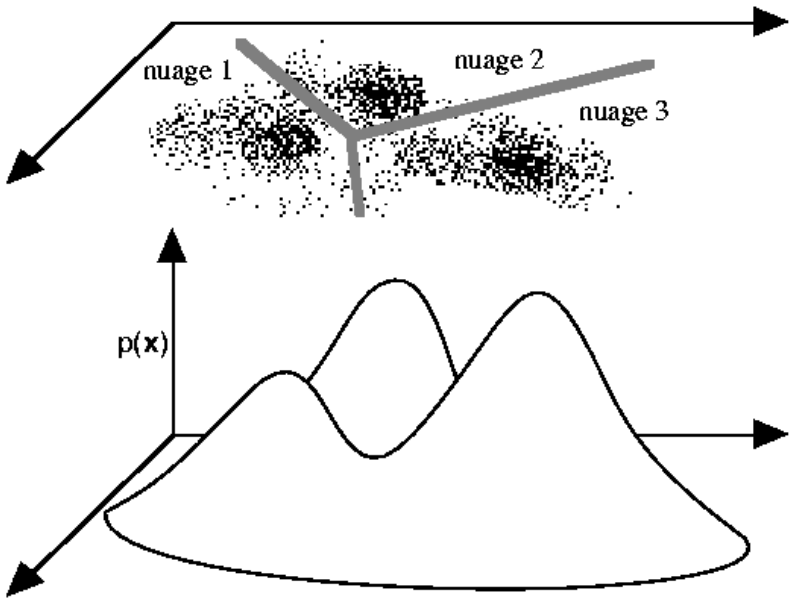


FIG 3.9 *Distribution de probabilités, un échantillon de points associés, et un découpage en nuages (clusters*)

6.1 Algorithme de K-Means [BOI87]; [CAL89]

On dispose d'un ensemble de L vecteurs x que l'on désire positionner en M classes. On désignera par :

- x_j^i : Les vecteurs appartenant à la classe i.
- y_i : Le centroïde de la classe i .
- L_i : Le nombre de vecteurs de la classe i.
- $d(x_j^i, y_i)$: La distance ou mesure de distorsion entre x_j^i et y_i.
- D_i : La distorsion totale de la classe i, $D_i = \sum_j d(x_j^i, y_i)$.
- D : La distorsion pour l'ensemble des vecteurs, $D = \sum_{i=1}^{M} D_i$.

Un nombre M de classes étant imposé a priori, le problème consiste à trouver la partition et les centroïdes de façon à minimiser la distorsion totale D. Une procédure itérative peut être basée sur les deux observations suivantes :

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

La variante LBG de l'algorithme de K-Means décrit ci-dessous permet l'optimisation de la partition du code-book. Le comportement de cet algorithme est influencé par le nombre des centroïdes initiaux.

ALGORITHME DE K-Means

Données

- Vecteur d'apprentissage x_i.

1 Initialisation

- Choisir un code-book initial $C = \{y_i ; 1 \le i \le M\}$.
- Mettre $m = 0; D_m = D_0$ (valeur max).

2 Construction des classes

- Partitionner les vecteurs d'apprentissage x_j en M selon l'hypothèse suivante : $x_j \in C_i$ si $d(x_j, y_i) \le d(x_j, y_k)$ pour tout $1 \le k \le M$, où d représente la norme euclidienne.
- Mettre $m = m + 1$ et calculer : $D = \sum_{i=1}^{M} \sum d(x_j^i, y_i)$, avec $x_j \in C_i$.

3 Mise à jour des centroïdes

- Mettre à jour chaque centroïde y_i par : $y_i = \frac{1}{|C_i|} \sum x_j$; avec $x_j \in C_i$.

4 Test d'arrêt

- Si $\frac{D_m - D_{m-1}}{D_m} \succ \zeta$ alors aller à l'étape 2 ; où la valeur de ζ est fixée a priori.

6.2 Algorithme de K- Plus Proches Voisins [BEN92]

L'algorithme de K- Plus proches voisins est lié à la notion de *proximité* ou de *ressemblance*. L'idée de cet algorithme est relativement simple. Elle consiste, étant donné un point $x \in \Re^n$ représentant la forme à reconnaître, à déterminer la classe des k points les plus proches de x parmi l'ensemble des formes d'apprentissage et à retenir pour la décision, la classe la plus représentée. Si $k = 1$, le point x est donc simplement attribué à la classe de son plus proche voisin.

Une variante de cet algorithme consiste à fixer un seuil s où $k/2 \le s \le k$ et à décider que x appartient à la classe w si au moins s parmi les k plus proches voisins de x appartient à w. Si aucune classe ne vérifie cette propriété, une décision de rejet est prise. Cet algorithme possède des propriétés de convergence remarquables quand le nombre d'échantillons d'apprentissage tend vers l'infini.

Néanmoins, cet algorithme a la réputation d'être lent en phase de décision. Il nécessite en effet, pour un espace d'apprentissage de N échantillons, de calculer N distances dans un espace à M dimensions pour prendre une décision. Cependant des variantes sub-optimales ont été proposées, nécessitant moins de calcul fondées sur différentes idées qui conduisent à une réduction de l'espace de recherche en limitant le nombre de points à consulter et donc le temps de calcul.

7 MODELES DE MARKOV CACHES

Depuis leur introduction en traitement de la parole [BAK75] ; [JEL76], les modèles de Markov cachés (Hidden Markov Models ou HMM) ont pris une importance considérable, au point que la quasi-totalité des systèmes de RAP utilisent cette modélisation. Les modèles de Markov cachés supposent que le phénomène modélisé est un processus aléatoire et inobservable qui se manifeste par des émissions elles-mêmes aléatoires. Ces deux niveaux donnent à l'approche markovienne une flexibilité qui est séduisante pour modéliser un phénomène aussi complexe que la production de la parole. De nombreuses présentations théoriques des HMM existent dans la littérature; nous reprenons en partie les notations de *L. Rabiner* [RAB89].

7.1 Définition

Un modèle de Markov caché est un automate stochastique particulier capable, après avoir été entraîné, d'estimer la probabilité qu'une séquence d'observations ait été générée par ce modèle. Idéalement, il faudrait pouvoir associer à chaque phrase possible un modèle. Il va de soi que ceci est irréalisable en pratique car le nombre de modèles serait beaucoup trop élevé. Des sous-unités lexicales comme le mot, la syllabe ou le phonème sont utilisées afin de réduire le nombre de paramètres à entraîner. A chacune de ces unités est associé un modèle de Markov caché constitué d'un nombre fini d'états prédéterminés. Formellement, un modèle de Markov caché peut être défini par l'ensemble des paramètres λ [RAB89]:

$$\lambda = L, A, B, \Pi$$

1. L est le nombre d'états du modèle,
2. $A = a_{ij} = p(q_j / q_i)$ la matrice de transition sur l'ensemble des états du modèle (on définit ainsi une topologie par l'intermédiaire de cette matrice A),
3. $B = b_{j(x_n)} = p(x_n / q_j)$ l'ensemble des probabilités d'émission de l'observation x_n dans l'état q_j .
   - Dans le cas d'entrées discrètes (après quantification vectorielle des observations), les probabilités d'émission peuvent être décrites comme des fonctions de densité de probabilité discrète $p(y_i / q_j)$.
   - Dans le cas d'entrées continues $(x_n \in \Re^d), p(x_n / q_j)$ est :
     - Supposée être de la forme d'une distribution multi-variables gaussienne, entièrement définie par le vecteur moyenne et la matrice de covariance, ou d'une distribution de type multi-gaussienne (somme pondérée de gaussiennes).
     - Supposée être représentée par un réseau de neurones (cf. § 5).
4. Π la distribution initiale des états, $\forall j \in [1, L]\ p(q_j / q_l)$ [10] .

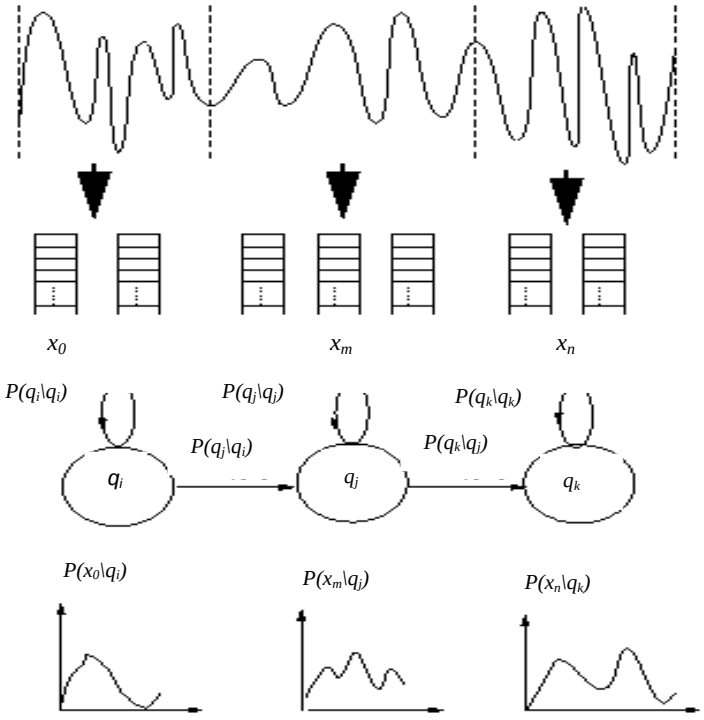
10 q_l représente l'état initial du modèle HMM, il ne peut émettre de vecteurs acoustiques.

En reconnaissance de la parole, des modèles de Markov gauche-droite d'ordre 1 sont le plus souvent utilisés du fait de l'aspect séquentiel du signal de la parole [BAK76]. Un modèle de Markov à 3 états est visible sur la figure 3.10[11] .

Les modèles de Markov cachés (HMM) supposent que la séquence de vecteurs acoustiques représentative du signal de parole soit une succession de segments stationnaires. Ainsi la parole est modélisée par une succession d'états, avec des transitions instantanées possibles entre ces états $p(q_j/q_k)$. Chaque observation est supposée être une fonction probabiliste de l'état. Deux processus stochastiques concurrents sont observés :

- Un premier processus qui est la séquence d'observations $X = X_1 , ..., X_N$.
- Un second processus, la séquence d'états non directement observable.

C'est pourquoi ces modèles sont dits ``cachés". La séquence d'états n'est pas directement observable.



par une distribution de probabilités pour chaque état associé à une observation et par des probabilités de transitions entre les états.

11 D'autres topologies sont aussi utilisées. Citons par exemple les modèles de Markov avec modélisation de la durée **[1 ; 100]** qui améliorent sensiblement les taux de reconnaissance ou les modèles de Bakis.

7.2 Problèmes à résoudre

Soit M le modèle de Markov caché associé à la parole X et constitué d'une concaténation de sous-unités lexicales. La reconnaissance de la séquence de vecteurs acoustiques X s'effectue en trouvant le modèle M qui maximise la probabilité $P(M/X,\lambda)$ (probabilité qu'un modèle M génère une séquence de vecteurs acoustiques X étant donné une série de paramètres λ). Cette probabilité est aussi appelée probabilité a posteriori. Malheureusement, il n'est pas possible d'accéder directement à cette probabilité par le processus d'entraînement des modèles de Markov, mais seulement à la probabilité qu'un modèle donné générera une certaine séquence de vecteurs acoustiques P (X/M).

En utilisant la loi de Bayes (3.16), il est possible de lier ces deux probabilités selon :

$$P(M/X) = \frac{P(X/M).P(M)}{P(X)} \tag{3.16}$$

Où

- P (X/M) est la vraisemblance de la séquence d'observations X étant donné le modèle M,
- P (M) est la probabilité a priori du modèle,
- P (X) la probabilité a priori de la séquence de vecteurs acoustiques.

Nous verrons un peu plus tard qu'il est nécessaire de choisir un critère pour l'entraînement des paramètres $\lambda = \{A,B,\Pi\}$:

- Critère MAP (Maximum A Posteriori probability) : Maximum a posteriori.
- Critère MLE (Maximum Likelihood Estimation) : Maximum de vraisemblance.

Hypothèses :

- **H1** On suppose que P (M) peut être calculée indépendamment des observations. Cette probabilité est en effet indépendante de X et peut être estimée à partir du modèle de langage.
- **H2** Pour une séquence d'observations connue, P (X) peut être considérée constant, puisqu'elle est indépendante du modèle, si les paramètres de ces modèles sont fixés. Ainsi maximiser $P(X/M).P(M)/P(X)$ revient à maximiser $P(X/M).P(M)$.

Il faut alors résoudre 3 problèmes liés à ces modèles [BOU94] ; [RAB89] ; [RAB93].

1. **L'estimation des probabilités :** comment calculer P(X/M) et quelles sont les hypothèses nécessaires à propos du modèle pour se définir une série de paramètres utiles pour la reconnaissance?

2. **L'entraînement :** étant donné une séquence d'observations X_j associée à leurs modèles de Markov respectifs, comment déterminer les paramètres des modèles afin que chacun ait la probabilité la plus grande possible de générer les séquences d'observations associées? Comment trouver l'ensemble des paramètres – qui maximiseront $P(M/X,\lambda)$ pour l'ensemble des séquences de vecteurs acoustiques X associé au modèle M[12] ? Cette probabilité n'étant pas directement accessible, on préfère maximiser $P(X/M,\lambda)$ (soit utiliser le critère MLE plutôt que MAP)[13].

3. **Le décodage :** étant donné une séquence de modèles de Markov avec leurs paramètres entraînés et une séquence d'observations X, comment trouver la meilleure séquence M_k de modèles de Markov élémentaires pour maximiser la probabilité que M_k ait généré les observations?

7.2.1 Problème 1 : Estimation des probabilités.

Le problème de l'estimation des probabilités peut être énoncé de la façon suivante : étant donné un modèle de Markov M, comment calculer la probabilité P(X/M) qu'il génère la séquence de vecteurs acoustiques X?

De manière générale, le calcul de cette probabilité s'effectue selon :

$$P(X/M) = \sum_{\Gamma} P(C,X/M) = \sum_{\Gamma} P(Q_1^N, X/M)$$

Où

- Γ = ensemble des chemins C dans le modèle M,
- $Q_1^N = \{q_I = q^0, q^1, ..., q^N, q^{N+1} = q_F\}$ est une séquence ordonnée de N états[14],
- q^n représentant l'état du HMM à l'instant n,
- q_k^n signifie que l'état q_k est visité à l'instant n.

Il existe deux procédures récurrentes de calcul de cette probabilité que nous nous proposons de décrire :

- L'algorithme Forward-Backward qui fournit une solution exacte à ce problème faisant intervenir tous les chemins dans le modèle HMM.

12 Ce critère est discriminant car il minimise le taux d'erreur.

13 Le critère MLE n'est plus discriminant ce qui constitue une des premières limitations des modèles de Markov cachés.

14 Les états finaux q_F et initiaux q_I sont exclus car ils ne peuvent pas émettre de vecteurs acoustiques. Ils sont généralement utilisés pour initialiser les équations récurrentes.

Algorithme Avant-Arrière (Forward-Backward)

L'algorithme *Forward-Backward* [BEL57] peut être utilisé pour calculer P(X/M) de manière récursive en posant :

$$\forall l \in [1,L]\ \alpha_n(l/M) = P(q_l^n, X_1^n / M) \tag{3.17}$$

Il vient alors :

$$\forall l \in [1,L]\ \alpha_{n+1}(l/M) = [\sum_{k=1}^{L} \alpha_n(k/M).p(q_l / q_k, M)].p(x_{n+1} / q_l) \tag{3.18}$$

Avec $\alpha_1(l) = p(x_1 / q_l).p(q_l / q_1)$

De même posons :

$$\forall l \in [1,L]\ \beta_n(l/M) = P(X_{n+1}^N / q_l^n, X_1^n, M) \tag{3.19}$$

Il vient alors :

$$\forall l \in [1,L]\ \beta_n(l/M) = \sum_{k=1}^{L} p(q_k / q_l, M).p(x_{n+1} / q_k).\beta_{n+1}(k/M) \tag{3.20}$$

Avec $\forall l \in [1,L]\ \beta_N(l) = 1$ ou 0

Et en utilisant ces deux dernières formulations, il est possible de calculer P(X/M) selon :

$$\forall n \in [1,N]\ P(X/M) = \sum_{i=1}^{L} \alpha_n(l/M).\beta_n(l/M) \tag{3.21}$$

Où les termes $\alpha_n(l/M)$ et $\beta_n(l/M)$ peuvent être calculés de manière récurrente. De plus, en posant n = N, q_F état final et q_I état initial :

$$P(X/M) = \sum_{\forall l \in \Im} \alpha_N(l/M) = \alpha_{N+1}(q_F / M) = \beta_0(q_I / M) \tag{3.22}$$

Remarque : $\Im$ représente l'ensemble des états finaux.

C'est souvent cette dernière formulation qui est utilisée en reconnaissance de la parole, car elle ne nécessite que le calcul des termes α pour déterminer la probabilité P(X/M)[15]. Il faut noter que les équations décrites ici ont nécessité une série d'hypothèses simplificatrices et limitatrices. Les équations présentées ici sont basées sur le critère du maximum de vraisemblance. Il est possible d'utiliser un autre critère permettant de réduire sensiblement la charge de calcul : le critère de Viterbi [FOR73] ; [VIT67].

---

15 Bien que le calcul des α et β permet aussi de calculer la probabilité suivante qui est utilisée (nous le verrons plus loin) lors de l'entraînement des modèles HMM :

$$P(q_{k,X/M}^n) = \alpha_n(k).\beta_n(k) \tag{3.23}$$

Algorithme de Viterbi

Au lieu de prendre en compte tous les chemins autorisés, seul le plus probable est gardé. Ainsi, il suffit de remplacer dans les équations précédentes l'opérateur Σ par max. Ce critère est largement utilisé en reconnaissance de la parole du fait du faible coût qui lui est associé[16] .

L'algorithme de Viterbi est donc une simplification de la récurrence avant qui devient :

$$P(q_l^n, X_1^n / M_j) = \max_k [P(q_k^{n-1}, X_1^{n-1} / M_j).p(q_l^n / q_k^{n-1}, M_j)].p(x_n / q_l^n, M_j) \tag{3.24}$$

Soit en passant au log :

$$-\log(P(q_l^n, X_1^n / M_k)) = \min_k[-\log(P(q_k^{n-1}, X_1^{n-1} / M_j)) - \log(p(q_l^n / q_k^{n-1}, M_j))] - \log(p(x_n / q_l^n, M_j)) \tag{3.25}$$

Cette dernière équation montre que le calcul de la probabilité P (X/M) est très semblable à une récurrence du type DTW (Dynamic Time Warping).

Hypothèses simplificatrices

L'utilisation pratique de ces modèles ne peut se faire qu'après avoir fixé quelques hypothèses qu'il est important de rappeler.

- **H1** On a supposé que P (M) peut être calculée indépendamment. Cette probabilité est en effet indépendante de X et peut être estimée à partir du modèle de langage.

- **H2** Pour une séquence d'observations connue, P(X) peut être considéré constante, puisqu'elle est indépendante du modèle, si les paramètres de ces modèles sont fixés.

- **H3** Les modèles de Markov sont supposés du premier ordre; ainsi la probabilité que la chaîne de Markov soit dans l'état q_l au temps n dépend uniquement de l'état de la chaîne de Markov au temps n-1 et est indépendante du passé.

- **H4** Les vecteurs acoustiques ne sont pas corrélés, la probabilité qu'un vecteur acoustique soit émis au temps n dépend uniquement de la transition de l'état q_k^{n-1} à q_l^n et est indépendante du passé.

16 n effet, il est évident que l'opération max est moins coûteuse en temps de calcul que l'opération Σ sur tous les états.

- **H5** La probabilité d'émission est supposée dépendante uniquement de l'état courant pour réduire le nombre de paramètres.

Pour calculer la probabilité d'émission, chaque état q_l doit être associé à une densité de probabilité $p(x_n / q_l)$. Il faut donc émettre des hypothèses supplémentaires à propos de cette densité de probabilité.

- **H6** Dans le cas d'entrées continues, $p(x_n / q_l)$ est supposée être de la forme d'une distribution multi-variables gaussienne [JUA85]. Ainsi la probabilité peut être exprimée selon :

$$p(x / q_l) = \sum_{j=1}^{N} c_{ij} . N_{ij}(x) \qquad (3.26)$$

où $N_{ij}(x)$ désigne la valeur au point x (trame acoustique) de la gaussienne dont les paramètres sont décrits ci-après.

$$N_{ij}(x) = \frac{1}{\sqrt{(2\Pi)^d . \det(\sum ij)}} . e^{-\frac{1}{2}.(x-\mu_{ij})^T . \Sigma_{ij}^{-1} . (x-\mu_{ij})} \qquad (3.27)$$

Avec :

- μ_{ij} : Vecteur moyen pour la gaussienne j de l'état q_l ,
- Σ_{ij} : Matrice de covariance pour la gaussienne j de l'état q_l ,
- c_{ij} : Coefficients de pondération,
- d : Dimension de l'espace dans lequel sont mesurées les trames acoustiques.

Les systèmes basés sur ce type de modélisation sont appelés systèmes à densités d'observations continues ou systèmes multi-gaussiens [BAH87] ; [BOU90]. L'utilisation de plusieurs gaussiennes permet une meilleure modélisation de la probabilité $p(x_n / q_l)$. De plus, en général, les différentes composantes du vecteur x_n sont supposées non corrélées[17] . Le nombre de paramètres utilisés est alors beaucoup moins important car les matrices de covariances des distributions sont diagonales. De nombreux systèmes de reconnaissance sont basés sur ce type d'estimation des probabilités d'émissions. Malheureusement, ces modèles nécessitent un temps de calcul élevé pour estimer les probabilités (donc pour l'entraînement et la reconnaissance). Il est en effet nécessaire de calculer un grand nombre de densités gaussiennes pour chaque trame acoustique et chaque état HMM. Une solution intermédiaire consiste à utiliser des modèles de Markov semi-continus [BEL90]. Ces modèles utilisent un dictionnaire de gaussiennes partagées par toutes les distributions associées à chaque état. Ce type de modèle a permis d'obtenir de très bonnes performances (du point de vue du taux de reconnaissance et du temps de calcul) pour la reconnaissance de la parole continue "grand vocabulaire" en combinant les avantages des deux méthodes (modèles continus et discrets) présentées dans ce paragraphe.

17 est encore une hypothèse supplémentaire à ajouter.

- **H7** Dans le cas d'entrées discrètes, l'hypothèse H6 n'est plus nécessaire. La séquence de vecteurs acoustiques X est quantifiée. Chaque vecteur acoustique x_n est remplacé par un centroïde y_i (le plus proche au sens d'une distance prédéfinie[18]) sélectionné dans un dictionnaire prédéterminé (codebook[19]).

Ainsi les probabilités d'émission peuvent être décrites comme des fonctions de densité de probabilité discrète $p(y_i / q_l)$. Ce type de système est couramment appelé système discret. Le temps de calcul nécessaire à l'utilisation de ces modèles est beaucoup moins important que pour les systèmes continus, mais les performances restent limitées.

7.2.2 Problème 2 : Estimation des paramètres et entraînement des modèles

Le but de l'entraînement des modèles acoustiques est de trouver l'ensemble des paramètres λ maximisant sur l'ensemble des données d'entraînement X_j la vraisemblance des données étant donné les modèles associés M_j , soit :

$$\arg\max_{\lambda} \prod_{j=1}^{J} P(X_j / M_j, \lambda_j) \qquad (3.28)$$

Il est nécessaire d'estimer deux ensembles de paramètres :

- Les probabilités de transitions entre les états : a_{ij} .
- Les probabilités d'émission des observations pour chaque état : $b_j(x_n)$.

Les approches les plus utilisées sont basées sur des adaptations de l'algorithme EM (Expectation-Maximization) [DEM77] appelées :

- Algorithme de Baum-Welch : P(X/M) est estimée en tenant compte de tous les chemins possibles (voir paragraphe précédent).

- Algorithme de Viterbi : P(X/M) est estimée en tenant compte du meilleur chemin uniquement (approximation de l'algorithme EM).

Entraînement Baum-Welch

L'algorithme de Baum-Welch est un processus itératif où, à chaque itération, de nouvelles valeurs des paramètres λ des modèles sont estimées à partir des anciennes valeurs. L'entraînement des modèles est effectué à partir de l'estimation de P (X/M) en tenant compte de tous les chemins possibles. On utilise les formules de récurrence "avant" et "arrière" définies dans ce chapitre. Nous avons déjà montré en posant :

18 Euclidienne par exemple.

19 Le dictionnaire est obtenu par une technique consistant à regrouper l'ensemble des vecteurs acoustiques en classes. Chaque classe contient des vecteurs relativement proches les uns des autres au sens de la distance choisie. Le centroïde d'une classe est alors défini comme la moyenne de l'ensemble des vecteurs de cette classe et est utilisé comme vecteur prototype .

$$\forall l \in [1,L]\,\alpha_n(l/M) = P(q_l^n, X_1^n / M)$$

$$\forall l \in [1,L]\,\beta_n(l/M) = P(X_{n+1}^N / q_l^n, X_1^n, M)$$

$$\forall n \in [1,N]\,P(X/M) = \sum_{l=1}^{L} \alpha_n(l/M).\beta_n(l/M)$$

et

$$\gamma_n(k/M) = P(q_k^n / X, M) = \frac{\alpha_n(k/M).\beta_n(k/M)}{P(X/M)}$$

Dans le cas d'une distribution mono-gaussienne, les formules de réestimations des paramètres sont données par les équations suivantes :

$$\mu_i = \frac{\sum_{n=1}^{N} \gamma_n(i/M).X_n}{\sum_{n=1}^{N} \gamma_n(i/M)} \tag{3.29}$$

$$\Sigma_i = \frac{\sum_{n=1}^{N} \gamma_n(i/M).(X_n - \mu_i).(X_n - \mu_i)^T}{\sum_{n=1}^{N} \gamma_n(i/M)} \tag{3.30}$$

Où μ_i et Σ_i sont la moyenne et la matrice de covariance de l'état i. L'extension au cas multi-gaussienne se fait aisément.

Le processus itératif mis en oeuvre consiste donc en deux phases :

- Une phase d'estimation où les récurrences avant et arrière nous permettent d'obtenir $p(q_k^n / X, M, \lambda)$ à partir d'un ensemble de paramètres λ.

- Une étape de maximisation où les paramètres λ sont mis à jour en utilisant les formules décrites dans ce paragraphe.

Entraînement Viterbi

Les paramètres sont optimisés de façon à maximiser la vraisemblance du meilleur chemin. Cela revient à supposer que les probabilités $\gamma_n(k) = P(q_k^n / X, M)$ sont égales à 0 ou 1. Comme pour l'algorithme EM, on part d'un ensemble de paramètres initiaux λ^0 et les paramètres optimaux λ sont obtenus de manière itérative. Le processus d'entraînement est composé d'une étape d'estimation (E) qui sert à trouver la segmentation qui maximise la vraisemblance à partir des paramètres, et d'une étape de maximisation (M), qui effectue une mise à jour des paramètres étant donnée cette segmentation. L'ensemble des paramètres initiaux λ^0 peut être estimé à partir d'une segmentation initiale (linéaire ou autre ou à partir de modèles déjà entraînés (par exemple par l'intermédiaire d'un corpus déjà segmenté comme TIMIT [ZUE90]). Il est ensuite possible à partir de la segmentation optimale trouvée de calculer les paramètres

des fonctions de vraisemblance en considérant tous les vecteurs associés à chacune des classes. Ce processus de réalignement des données acoustiques à l'aide d'un modèle et de réentraînement d'un nouveau modèle[20] est effectué jusqu'à ce qu'une certaine convergence soit atteinte (la segmentation ne varie plus ou l'accroissement relatif de la vraisemblance pour l'ensemble des données d'entraînement est inférieur à un seuil fixé).

7.2.3 Problème 3 : Décodage

Après l'entraînement des paramètres des modèles HMM, la reconnaissance d'une séquence de vecteur acoustique X correspondant à une parole s'effectue par le calcul de la probabilité $P(X/M_j)$ pour tous les modèles (ou séquence de modèles) M_j. Ce calcul peut être effectué en utilisant les récurrences déjà décrites dans ce chapitre. Il faut toutefois noter que dans le cas de grands vocabulaires, ces récurrences sont coûteuses en temps de calcul. De plus, dans le cas de la parole continue, le nombre de combinaisons possibles des modèles (séquences de M_j) est très important. Ainsi, des techniques d'élagage ("pruning") ont été développées pour permettre l'utilisation de ces récurrences. La procédure de reconnaissance consiste à trouver le modèle (ou la séquence de modèles) M_k pour lequel :

$$k = \arg\max_{\forall j} P(M_j / X) = \arg\max_{\forall j} P(X / M_j).P(M_j) \qquad (3.31)$$

Pour une reconnaissance de la parole continue, le score du modèle acoustique $P(X/M_j)$ doit être multiplié par la probabilité de la séquence de mots associée $P(M_j)$.

8 RECONNAISSANCE EN MOTS ISOLES

Lors d'une tâche de reconnaissance en mots isolés, on demande au locuteur de faire une petite pause entre chaque mot. Le signal se présente alors sous la forme illustrée à la figure 3.11. Or puisqu'il existe des algorithmes relativement efficaces qui permettent de distinguer silence et parole, le problème de la reconnaissance en mots isolés, consiste simplement à faire correspondre un mot du vocabulaire à chaque îlot de parole. Néanmoins, pour de très grands vocabulaires, il n'est pas réaliste d'accomplir cette tâche de façon exhaustive, c'est-à-dire en appliquant les algorithmes « avant » ou de Viterbi à chaque mot du vocabulaire, d'où une variété d'approches qui tentent de réduire l'espace de recherche et, par conséquent, le temps d'exécution de la reconnaissance.

FIG 3.11 *Exemple d'un signal acoustique résultant d'une dictée en mots isolés*

20 Comme pour l'entraînement Baum-Welch décrit précédemment.

8.1 Systèmes de reconnaissance en mots isolés et grand vocabulaire

Dans cette section les approches utilisées dans plusieurs systèmes de reconnaissance en mots isolés et grands vocabulaires. Ces systèmes ne font plus l'objet de recherches actives. Par contre, les idées qu'ils ont permis de développer sont à la base de certaines applications expérimentales ou commerciales. On peut entre autres citer Dragon Dictate et StockTalk [LEN92]). La grande majorité de ces systèmes ont été développés pour l'anglais.

8.1.1 Système du CSELT

Le système du CSELT (Centro Studi E Laboratori Telecomunicazioni) a été développé pour l'Italien [FIS88] ; [FIS89] ; [LAF87] ; [LAF88]. Son lexique est d'environ 8000 entrées. La reconnaissance se compose de deux étapes: un pairage rapide et un pairage exact . Lors de la première étape, le signal est étiqueté en terme de traits acoustiques généraux: silence, fricative, voyelle haute etc. Par la suite, lors de la deuxième étape, l'algorithme de Viterbi n'est appliqué qu'aux seuls mots dont la séquence phonétique correspond (plus ou moins exactement c'est-à-dire selon un algorithme d'alignement dynamique) à la séquence de traits attribuée au signal. Ces ensembles de mots auxquels correspond une même séquence de traits sont connus sous le nom de cohortes. Puisque les traits acoustiques généraux sont fixes et connus d'avance, ces cohortes peuvent être pré-calculées, ce qui rend très rapide la première partie de l'algorithme.

Par contre, il est aussi très important que le moins d'erreurs possibles ne soient introduit lors de cette première étape, c'est-à-dire que le moins de mots possibles auxquels l'algorithme de Viterbi aurait attribué le meilleur score, ne soient omis de la liste de mots pré-sélectionnés. Or, si les traits acoustiques plus sont précis, plus les cohortes extraites sont petites, plus malheureusement on peut confondre deux traits et extraire une mauvaise cohorte, d'où la nécessité de bien choisir ces traits (voir à ce sujet, [LAG85] et [VER89]). Par exemple, si l'on choisit de distinguer les plosives voisées que sont b, d, et g des plosives correspondantes non voisées p, t, k, on peut alors distinguer les paires de mots bedon/peton, dans/tant et gué/quai, si cette distinction n'est pas faite, les paires citées auraient été incluses dans une même cohorte. Le voisement étant un trait relativement facile à détecter de façon fiable, cette distinction est un bon choix. Par contre, la distinction entre les diverses plosives voisées étant d'un point de vue acoustique plus ardue, il est préférable de ne pas faire cette distinction, ce qui aura pour conséquence d'inclure des mots tels beau, dos et go dans une même cohorte.

On doit souligner que selon *Carlson* et *Adda*, cette approche n'est pas aussi efficace pour le Français qu'elle l'est pour d'autres langues tel l'anglais. C'est-à-dire que pour un même nombre de traits, le nombre de cohortes distinctes est plus petit et donc le nombre de mots par cohorte plus grand.

Le groupe de recherche du CSELT est l'un de ceux qui poursuit activement des recherches sur la parole en mots isolés. Ainsi, le dernier article paru [LAF95] porte sur une nouvelle méthode de pairage rapide potentiellement plus facile à adapter à la parole continue.

8.1.2 TANGORA

Le système TANGORA a été élaboré par un groupe le "Speech Recognition Group" d'IBM Thomas J. Watson Research Center [AVE87] ; [BAH88] ; [BAH89]. Son vocabulaire se compose d'environ 20 000 formes anglaises. Une forme est une instance phonétique distincte d'un mot. Cela veut dire qu'un mot ayant deux prononciations distinctes compte pour deux entrées. Tout comme dans le système du CSELT, la reconnaissance dans TANGORA comporte deux étapes: un pairage rapide et un pairage exact [AVE86].

La fonction du pairage rapide est ici encore de fournir au pairage exact un sous-ensemble des mots les plus prometteurs. La technique employée est cependant totalement différente. Il s'agit d'une méthode dans laquelle un vote ou valeur est attribué à chacun des mots du lexique par l'ensemble des vecteurs d'observation $y = y_1, y_2, ..., y_L$ d'un signal acoustique [BAH88]. Ce vote est calculé par une équation de la forme:

$$S_m = \sum_{i=1}^{L} v(y_i, m) + cte_m$$

Où L représente le nombre d'unités de temps de l'énoncé, une valeur d'initialisation pour chaque mot et la fonction de votation. Il est à noter que la fonction n'étant pas fonction du temps mais seulement des valeurs des paramètres des vecteurs , la valeur octroyée à deux mots formés des mêmes phonèmes mais pas nécessairement dans le même ordre doit être la même. Cela revient à dire que c'est l'identité des phonèmes d'un mot et non leur ordre qui est surtout importante dans l'identification de ce mot. C'est d'ailleurs à cette conclusion qu'on aboutit également pour l'anglais [MAR95].

Les mots ayant obtenu un vote supérieur à une valeur donnée sont qualifiés pour l'étape du pairage exacte. Typiquement on fixe cette valeur de façon à ce que le nombre de mots à retenir soit entre 20 et 25. Par la suite, on rajoute à cette liste de candidats au pairage exacte, un ensemble de mots « similaires » d'un point de vue acoustique aux mots initialement sélectionnés (un type de cohorte d'en moyenne 100 mots) [BAH90]. C'est cette liste allongée (typiquement d'environ 2000 mots) qui est considérée lors du pairage exacte consistant en l'algorithme « avant ».
Les articles récents publiés par IBM portent plutôt sur des systèmes à moyens et grands vocabulaires.

8.1.3 Système de l'INRS

Le système développé par l'INRS-Télécommunication (Institut national de recherche scientifique, Québec, Canada) pour l'anglais, est basé sur un vocabulaire de 75 000 entrées [GUP88]. Il se compose, non pas de deux, mais de trois étapes dont une première étape d'estimation du nombre de syllabes, une deuxième étape de production de scores heuristiques et une troisième étape consistant en un pairage exact (algorithme « avant ») effectué non plus de manière synchrone mais plutôt de manière asynchrone à l'aide d'un algorithme de type A* [NIL82].

C'est *Jelinek* [JEL69] et [JEL76] qui a appliqué le premier l'algorithme A*, une technique d'intelligence artificielle, en reconnaissance de la parole. Dans l'algorithme A*, les hypothèses les plus prometteuses sont conservées dans une pile. Le degré de promesse d'une hypothèse est représenté par une valeur numérique appelée score total. Cette valeur E est égale à la somme du score courant de l'hypothèse et d'une valeur heuristique représentant un estimé du score associé au meilleur chemin possible entre l'état courant et un état final. Si pour toutes les hypothèses partielles et pour toutes les solutions sol on a $E(sol) \geq S(sol)$ ou si pour toutes les solutions sol, on a $E(sol) = S(sol)$ où $S(sol)$ est le score de la solution, c'est-à-dire que l'heuristique surestime toujours le potentiel d'un chemin, l'algorithme A* est admissible et donnera toujours la meilleure solution.

Le résultat de la deuxième étape du système de l'INRS ne consiste donc pas en une liste de mots possibles comme c'était le cas pour les systèmes précédents mais plutôt en des scores heuristiques admissibles devant servir de guide à la fouille exacte. Ces scores sont obtenus en parcourant un graphe syllabique au moyen de l'algorithme de Viterbi. Ce graphe ne contenant pas de boucle, il en existe un pour les mots d'une syllabe, un pour ceux de deux syllabes, etc. Chaque graphe est constitué par l'ensemble des syllabes occupant cette position dans un mot du lexique. Par exemple, le graphe pour les mots de trois syllabes est composé de trois sous-graphes, l'un incluant toutes les premières syllabes de tous les mots du lexique, un autre incluant toutes les deuxièmes syllabes de tous les mots du lexique et un dernier incluant toutes les troisièmes syllabes de tous les mots du lexique.

Le choix du graphe est fonction du nombre de syllabes de l'énoncé, nombre estimé, lors de la première étape, par l'application de l'algorithme de Viterbi sur un structure composée de HMM génériques capables de détecter les voyelles des consonnes. Les scores heuristiques obtenus au moyen des graphes syllabiques sont ensuite utilisés pour guider l'application de l'algorithme « avant » à une structure d'arbre représentant de façon compacte tous les mots du lexique.

Les travaux de l'INRS ont porté sur les systèmes de RAP en parole continue et de concert avec Bell Northern Research (BNR) sur le développement de produits [LEN92].

8.1.4 PARSYFAL

PARSYFAL réalisé par les laboratoires de *Derounault* et *Merialdo* était, à l'époque, l'un des rares systèmes de RAP pour le Français. Avant de le présenter en détails, soulignons d'abord quelques particularités du français qui rendent sa reconnaissance plus complexe à traiter et, dans le cadre de cette partie du document, la dictée en mots isolés moins naturelle que pour l'Anglais. Il y a d'abord l'existence de nombreuses élisions, soit la suppression de la voyelle finale de certains mots devant un mot débutant par une voyelle ou un *h* muet, et de liaisons, soit la prononciation d'une consonne normalement muette à la fin d'un mot devant un mot débutant par une voyelle ou un *h* muet. Ces deux phénomènes tendent, en effet, à lier dans une même prononciation deux mots qu'un système de RAP en mots isolés voudrait séparer. Ensuite, le fait que, pour obtenir une même couverture de texte, un lexique français doive contenir un nombre appréciable d'entrées de plus qu'un lexique anglais. Ceci est dû, entre autre, au fait qu'en Français, la

plupart des mots, en particulier les verbes, ont un nombre d'inflexions phonétiques beaucoup plus grand qu'en anglais. Par exemple, à la phonie anglaise /t a k/ (talk) correspond les phonies françaises /p a r l/ (parle, parles), /p a r l é/ (parlez, parler), /p a r l on/ (parlons), l'ajout de la marque anglaise du future /w i l/ (will), force l'ajout des phonies /p a r l e r é/ (parlerai, parlerez), /p a r l e r a/ (parlera), /p a r l e r on/ (parlerons, parleront) et ainsi de suite.

Pour contourner ces problèmes, la dictée dans PARSYFAL se fait syllabe par syllabe. Ce faisant, le nombre d'entrées effectivement distinctes, n'est plus de 200 000 mots mais bien de 6400 syllabes liaisons incluses. En effet, avec cette approche, le problème des liaisons et élisions est réglé car les syllabes supplémentaires dues à la présence de ces dernières seront déjà incluses dans l'ensembles des syllabes probablement présentes à l'intérieur des mots. Ainsi, la syllabe *pa* que la liaison dans trop amis qui nous forcerait à rajouter est déjà présente à cause de son occurrence dans le mot patate. Cela n'aurait pas été le cas, si c'est le mot *pami* qui aurait dû être rajouté.

Malheureusement, cette approche fait que PARSYFAL n'est qu'à moitié un système de reconnaissance à grand vocabulaire. Il serait plutôt un système à moyen vocabulaire (avec en prime des « mots » très courts), qui de plus, oblige les énoncés à être présentées sous la forme de syllabes isolées, ce qui semble être encore moins naturelle que la parole isolée.

Devant l'adaptation réussie pour le Français de « vrais » systèmes de RAP à grand vocabulaire, la recherche sur la parole en syllabes isolées a été abandonnée et de ce fait, les algorithmes et structures qui y étaient dédiés.

8.1.5 Dragon Dictate

Dragon Dictate est actuellement le plus connu des systèmes de dictée à grand vocabulaire commercial [BAH91]. C'est un produit de Dragon Systems Inc. compagnie basée aux Etats-Unis. Pour des raisons assez évidentes, son fonctionnement n'est pas divulgué en détails. Il est basé sur un algorithme de reconnaissance à deux passes, la première un pairage rapide fournissant une liste de mots, à la seconde un pairage exact [BAK74]. Caractéristique dû au fait que l'algorithme est utilisé au sein d'une application de dictée, le pairage exact ne produit pas une seule solution mais plutôt une liste ordonnée des meilleures solutions parmi lesquelles l'utilisateur peut, en cas d'erreur, choisir une alternative.

Il faut souligner cependant que Dragon Dictate n'est pas un système indépendant du locuteur. C'est un système de type adaptatif, c'est-à-dire qu'un ré-entraînement partiel des HMM a lieu de temps en temps à partir d'énoncés tirés d'un locuteur particulier. C'est là un compromis effectué dans le but d'obtenir des taux de reconnaissance acceptable en un temps acceptable. En effet, le temps est un facteur très important pour une application commerciale. Or, les variations pour un même énoncé et un même locuteur étant en général moindre qu'au travers plusieurs locuteurs, le pouvoir discriminant des HMM obtenus par ré-entraînement sera plus grand. Ce faisant, le pairage sera plus exact et certaines méthodes d'accélération de la fouille, comme l'emploi de plus petites cohortes ou de plus petits faisceaux d'émondage pourront être utilisés sans conséquence au niveau du taux de reconnaissance.

Depuis le début des années 90, c'est la parole continue qui est le sujet de publications chez Dragon.

9 CONCLUSION

Les systèmes actuels de reconnaissance automatique de la parole sont construits à partir d'une modélisation probabiliste avec des modèles de Markov cachés [WOO95]. L'approche probabiliste a montré son efficacité, et de nombreux travaux continuent à améliorer les performances des systèmes existants [HUA93] ; [ZHA94]. Cependant, ces systèmes doivent être perfectionnés sur de nombreux points. Ainsi, localisation et identification phonétique sont très étroitement associées durant la phase de décodage par des modèles de Markov classiques, et la segmentation devient un résultat annexe du décodage; pourtant, la modélisation de la durée phonétique reste insuffisante. De plus, la grande variabilité inter-locuteurs diminue l'efficacité de modèles phonétiques indépendants du locuteur, qui deviennent faiblement discriminants. Enfin, la représentation du signal choisie est généralement appliquée de manière identique à des zones stationnaires comme le centre des voyelles et à des zones fortement transitoires comme les explosions. Les analyses du signal utilisées font généralement l'hypothèse d'une stationnarité à court terme du signal de parole qui n'est pas pertinente. Nous pensons qu'il est nécessaire de ré-introduire dans ces systèmes des traitements spécifiques, adaptés aux multiples aspects de la parole.

Le principe des techniques de Quantification Vectorielle les plus utilisées dans le traitement de la parole a été décrit dans ce chapitre. Néanmoins, plusieurs recherches ont montré que la QV fournit une décision dure non probabilisée qui ne transmet pas assez d’informations sur les observations. Ce qui nous a poussé à développer dans le cadre de ce travail, des nouveaux algorithmes de partitionnement dont les avantages et le principe seront décrits dans le chapitre suivant.

PARTIE II

Conception & Réalisation

Chapitre 4

Nouveaux algorithmes de partitionnement

Le chapitre4 est organisé comme suit : Après avoir rappelé synthétiquement ce qu'est une classification, et sur quoi elle travaille, nous présentons quelques stratégies traitant les données hétérogènes (intervalle, numérique, symbolique, ensembliste). Parmi l'ensemble de solutions proposées, nous nous intéressons particulièrement, aux méthodes impliquant une classification non supervisée par partition et nous en retenons deux méthodes comme base pour notre travail. Nous évoquons ensuite le contexte de l'étude ainsi que le principe et les avantages des algorithmes choisis pour la classification. Dans la dernière section, afin d'illustrer la performance des méthodes proposées, nous présentons des exemples d'application.

Chapitre 5

Modèle Hybride HMM – MLP

Le chapitre 5 présente l'étude et le test d'un système de reconnaissance de la parole arabe à base des modèles hybrides HMM-MLP. Nous rapportons éventuellement ses avantages et les résultats obtenus avec ces modèles.

1 INTRODUCTION

Ces dernières années, il y a eu d'événement fondamental fonçant le connexionisme sur une nouvelle voie. On constate une volonté de dépasser les limitations actuelles du connexionisme. Ainsi, de nouveaux types de systèmes voient le jour, inspirés de la neurobiologie, de la psychologie ou mixant des techniques connexionnistes avec d'autres symboliques [LAZ00] ; [LAZd02] ou stochastiques [LAZc02] : Modèles de Markov cachés (HMM), ces modèles sont couramment appelés "modèles hybrides". Dans ce sens, l'ère du Perceptron Multi-Couches (MLP) devra évoluer pour dépasser la "simple" classification de forme.

Les modèles hybrides HMM/MLP ont déjà été utilisé avec succès dans divers systèmes qui proposent de remplacer l'évaluation des probabilités d'observation soit discrètes soit avec des gaussiennes dans les formalismes classiques des HMM, par un réseau de neurones de type MLP, aussi bien pour le manuscrit [GOR98], l'audio visuel [ROB99] ou pour la reconnaissance de la parole [BOU97] ; [BER99] ; [HAGa00] ; [HAGb00] ; [LAZa02] ; [LAZb02] ; [LAZc02] ; [MOR00] ; [MORb01] ; [YAN97].

Le connexionisme a naturellement été appliqué à la reconnaissance de la parole au vu des bonnes capacités en classification de formes. Cependant, les modèles connexionnistes n'ont pour l'instant pas surpassé les HMM. Ceci est en grande partie due à leur relative inadéquation au problème du traitement séquentiel de l'information. C'est pourquoi une grande tendance dans le domaine du neuromimétisme consiste à étendre les systèmes classiques ou à développer de nouveaux modèles pour prendre en compte les variabilités inhérentes à la parole. A partir des modèles statiques tel le perceptron, les structures ont été adaptées pour tenir compte de l'aspect temporel sans pour autant changer complètement l'architecture des réseaux. Des modèles hybrides HMM/MLP ont été conçus ces dernières années pour l'Anglais britannique et américain [RII97] et pour le Français [AVE87] afin d'additionner les qualités de chacun des modèles fusionnés mais sans réellement homogénéiser l'architecture.

C'est la raison pour laquelle, nous étions intéressés d'utiliser les modèles hybrides HMM/MLP dans le cadre de la Reconnaissance Automatique de la Parole (RAP) arabe isolée indépendante de locuteur [LAZa02] ; [LAZb02] ; [LAZc02], néanmoins que le traitement de la parole arabe est encore à ses débuts au quel nous espérons apporter une contribution à travers cette thèse. Pour augmenter la performance du système hybride proposé, nous proposons d'intégrer dans le système une nouvelle technique de segmentation reposant sur la logique floue. L'algorithme est nommé Fuzzy C-Means (FCM) repose sur l'utilisation d'une dissimilarité pour mesurer par des variables non floues les vecteurs acoustiques de la base considérée, tandis l'appartenance de ces derniers aux classes est floue. Nous rapportons dans ce chapitre, les résultats obtenus sur une base personnelle. La comparaison de la performance du système hybride FCM/HMM/MLP [LAZa03] ; [LAZb03] ; [LAZe03] ; [LAZf03] avec celle d'autres reconnaisseurs utilisant les HMM discrets donne déjà des résultats très prometteurs.

2 RESEAUX DE NEURONES ET MODELES HYBRIDES

2.1 Généralités

Ces 20 dernières années, les réseaux de neurones ont pris une part de plus en plus importante dans le monde de la recherche notamment depuis que Rumelhart en 1986 a montré les multiples possibilités des réseaux de neurones à couches multiples. Ils sont en particulier utilisés comme estimateur statistique. Il existe de nombreux types de réseaux de neurones utilisés en reconnaissance de la parole:

- Les réseaux de neurones multi-couches perceptrons (MLP : Multi-Layer Perceptron). De nombreuses publications ont montré l'utilité de ce type de réseau en reconnaissance automatique de la parole [HAGa00] ; [HAGb00] ; [MOR00] ; [MORb01]. Nous n'utiliserons que ce type de réseau dans les expériences décrites dans ce travail.

- Les réseaux récurrents (Recurrent Neural Network [ROB94]). Robinson et d'autres ont montré dans de nombreuses publications que ce type de réseau pouvait être utilisé efficacement en reconnaissance de la parole. En outre, un système de démonstration ABBOT basé sur les modèles hybrides utilise ce type de réseau de neurones pour estimer les probabilités d'émission et effectuer une reconnaissance de la parole continue en anglais pour un vocabulaire de 20000 mots.

- Les réseaux de neurones à délais temporels (TDNN : Time Delay Neural Network [LAN90]). Ce type de réseau introduit un délai temporel visant d'une part à apprendre la structure temporelle des événements acoustiques et les relations entre ces événements et d'autre part à rendre ce réseau insensible à de possibles erreurs d'alignement (ne reposant donc pas sur une segmentation précise de l'entrée). Les résultats publiés avec ce type de réseau ne portent que sur de petites t“aches telles que la reconnaissance de phonèmes particuliers pour quelques locuteurs.

- Les réseaux prédictifs (Predictive Neural Network) [JOD94]. Ces réseaux sont très peu utilisés en reconnaissance automatique de la parole, pourtant ils ne sont pas à écarter complètement.

Ces réseaux ont été appliqués avec succès à une multitude de problèmes de classification. Les réseaux de neurones les plus utilisés en reconnaissance de la parole sont les perceptrons multi-couches [HAGa00] ; [LAZa02] ; [LAZb02] ; [LAZc02] ; [LAZd02] ; [MORb01]. Une seule couche cachée est généralement utilisée[1] . L'addition de cette couche cachée permet au réseau de modéliser des fonctions de décision complexes et non linéaires entre n'importe quel espace d'entrée et de sortie. Nous allons décrire dans ce chapitre comment ils peuvent être utilisés conjointement avec les HMM en reconnaissance de la parole.

1. En effet, il a été démontré qu'un réseau à plusieurs couches cachées est équivalent à un réseau à une couche cachée de taille plus importante.

2.2 Présentation

La figure 5.1 représente un réseau de neurones perceptrons à une couche cachée. Nous utiliserons ce type de réseau pour développer la théorie associée à ces nouveaux modèles (gardant en mémoire le fait que les autres réseaux cités au début du paragraphe peuvent aussi être utilisés).

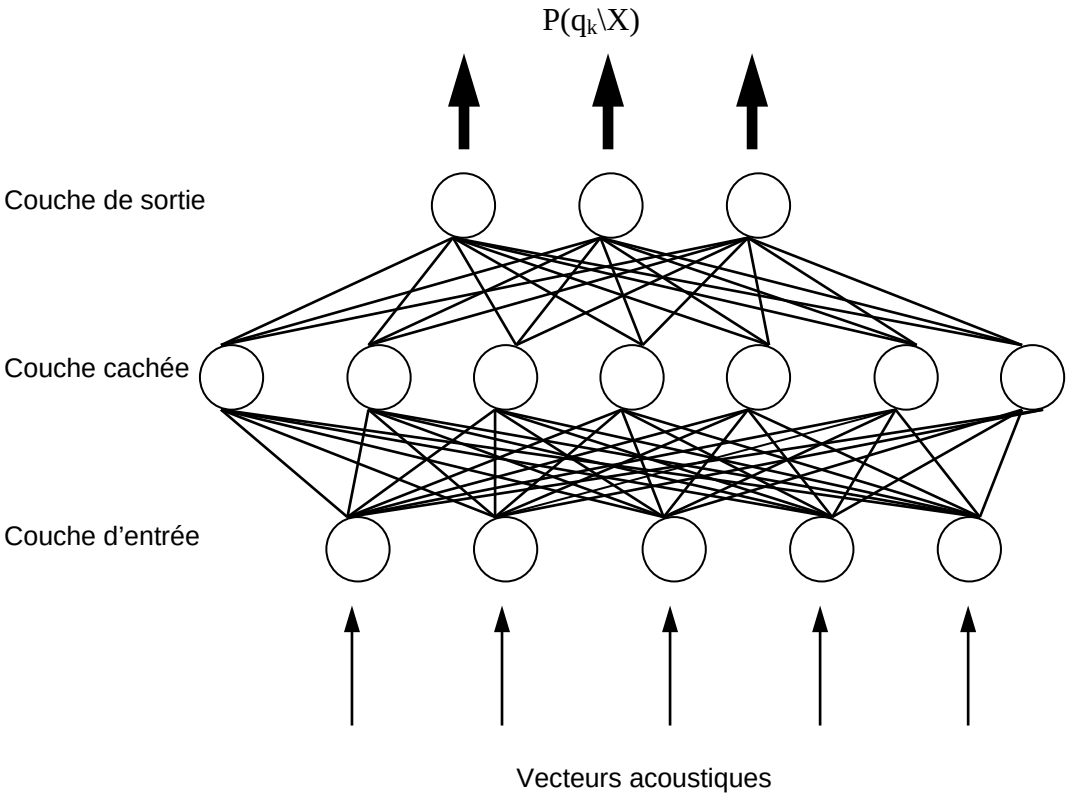


FIG 5.1 *MLP à une couche cachée*

Ils sont généralement constitués d'unités élémentaires interconnectées par des liens avec des poids variables. Le réseau présenté en figure 5.1 est constitué d'unités élémentaires appelées neurones. Le neurone calcule une somme pondérée des composantes du vecteur d'entrée et ajoute un biais à cette somme. Ensuite une fonction non-linéaire est appliquée au résultat. L'architecture du neurone peut donc être résumée simplement sur la figure 5.2. Plusieurs types de fonctions non-linéaires peuvent être définies :

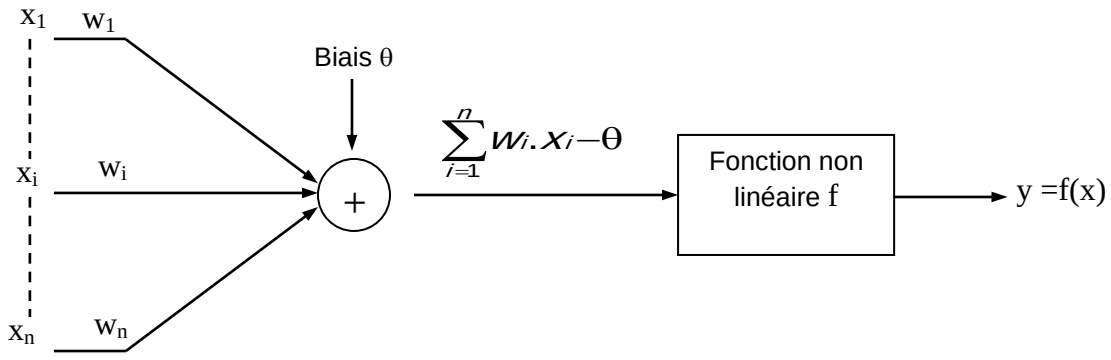


FIG 5.2 *Le neurone*

- **Fonction softmax**

$$f(z_i)=\frac{e^{z_k}}{\sum_{k=1}^{n} e^{z_k}} \tag{5.1}$$

- **Fonction signe**

$$\begin{cases} f(z) = +1 & \text{si } z>0 \\ -1 & \text{si } z<0 \end{cases} \tag{5.2}$$

- **Fonction sigmoïde**

$$f(z)=\frac{1}{1+e^{z}} \tag{5.3}$$

La fonction non linéaire la plus utilisée est la sigmoïde. Les différents termes de la somme pondérée des nœuds du réseau constituent les paramètres (ou poids) de celui-ci. Ils sont estimés lors d'un entraînement dit supervisé. Un entraînement est dit *supervisé* lorsque les entrées sont présentées au réseau les unes après les autres avec la sortie imposée correspondant à chaque entrée. Le problème de l'entraînement d'un réseau revient à trouver un ensemble de paramètres (poids) W qui minimise une fonction d'erreur E. Il existe plusieurs critères d'erreur qui peuvent être utilisés. Notons T l'ensemble des données d'entraînement, les sorties réelles y et désirées τ. Les critères les plus connus sont :

- **Le critère des moindres carrés**

$$E=\sum_{t\in T}\frac{1}{2}\cdot\sum_{l=1}^{L}\left(y_l^t-\tau_l^t\right)^2 \tag{5.4}$$

- **Le critère entropique**

$$E = \sum_{t \in T} \sum_{l=1}^{L} \tau_l^t \ln(y_l^t) \tag{5.5}$$

Les techniques du type descente de gradient sont les plus adaptées pour estimer ces paramètres (les poids du réseau W). Elles sont basées sur le calcul du gradient de l'erreur par rapport aux paramètres du réseau ($\delta E/ \delta w_{ij}$) et l'ajustement de ces paramètres jusqu'à ce qu'un minimum de l'erreur soit atteint.

La méthode de "back-propagation" (rétro-propagation) [JOD94], minimisant un des critères d'erreur (erreur quadratique ou entropie) entre les sorties réelles et désirées pour chacun des vecteurs d'entrée, est utilisée pour modifier la valeur des poids. La fonction d'erreur (pour un réseau à plus d'une couche cachée) est une fonction non linéaire des poids avec plusieurs minima possibles. Le gradient de l'erreur est employé pour ajuster les poids de l'élément de sortie et est propagé pour modifier les poids des couches cachées (d'où le nom de rétro-propagation de la couche de sortie vers la couche cachée). L'ajustement des paramètres se fait généralement selon la formule :

$$W^{T+1} = W^T - \eta \frac{\delta E}{\delta W_{ij}} \Big| W^T \tag{5.6}$$

où η est appelé taux d'apprentissage et doit être assez faible pour garantir la convergence du processus et assez grand pour éviter une convergence trop lente. Cet algorithme est le plus utilisé de nos jours pour l'entraînement des MLPs.

Le but de l'hybridation avec un Réseau de Neurones (RN) et en particulier un MLP (dans notre cas) est de mieux évaluer les probabilités d'observation des états, non plus par une modélisation de l'espace des vecteurs acoustiques mais par un apprentissage discriminant au niveau local de la fenêtre d'analyse. Un unique RN répond pour chaque observation les probabilités a posteriori de tous les états possibles des HMM hybrides (cf. FIG 5.3).

Le principe du développement de la théorie associée à ces nouveaux modèles est le suivant : Les vecteurs acoustiques présentés à l'entrée du MLP sont classés en k classes correspondant aux états stationnaires des HMM.

Lorsque le RN est entraîné à minimiser un critère des moindres carrés ou d'entropie. Dans notre application, nous avons utilisé la fonction coût la plus utilisée : Le critère des moindres carrés (cf. Equation 5.4), les valeurs optimales obtenues en sorties peuvent être interprétées en terme de probabilités a posteriori des classes de sortie conditionnées par les entrées. Ainsi, il est possible de démontrer que la sortie correspondante à l'état q_k est une estimation de la probabilité locale $P(q_k/x_n)$ si le vecteur acoustique x_n est présenté à l'entrée du MLP. Rappelons que les réseaux récurrents (RNN – Recurrent Neural Network [ROB94]) ou à délai temporel (TDNN – Time Delay Neural Network [LAN90]) peuvent être utilisés pour estimer cette probabilité. Or, en appliquant la loi de Bayes, il vient :

$$P(q_k/x_n) = \frac{P(x_n / q_k).P(q_k)}{P(x_n)} \quad (5.7)$$

En divisant la probabilité locale fournie par le RN $P(q_k/x_n)$ par la probabilité a priori $P(q_k)$ facilement calculable à partir de l'alignement Viterbi, il est possible d'obtenir la vraisemblance $P(x_n/q_k)/P(x_n)$. Lors de la phase de reconnaissance, la probabilité $P(x_n)$ étant constante, nous possédons donc une estimation de la probabilité d'observation des vecteurs acoustiques dans les états : $P(x_n/q_k)$ à la base de l'équation des HMM notamment des récurrences α et β de l'algorithme Baum-Welch et des récurrences Viterbi permettant le calcul de la probabilité de séquence des vecteurs acoustiques conditionnés par les états du modèle considéré P(X/M). Il n'est donc plus nécessaire d'effectuer, comme pour les HMM classiques, d'hypothèses restrictives sur la distribution des densités de probabilités. Les transitions sont toujours apprises par les HMM qui évaluent les vraisemblances des classes mots. Enfin, par l'application de la règle de Bayes, les probabilités a posteriori des classes mots $P(M_j/X)$ sont calculées, ce qui nous permet de calculer le score du modèle acoustique :

$$k = \arg \max_{\forall j} P(M_j / X) \quad (5.8)$$

Nos travaux ont été basés sur l'application d'un MLP avec une seule couche cachée, une fonction sigmoïde et une couche de sortie dont le nombre de neurones vaut le nombre d'états des HMM utilisés. Dans cette première application à vocabulaire réduit, un HMM gauche-droit est utilisé pour modéliser chaque mot. Dans une application de vocabulaire plus large non connu à l'avance, nous pensons d'utiliser des HMM pour modéliser les phonèmes arabes, qui seront concaténés pour former des mots.

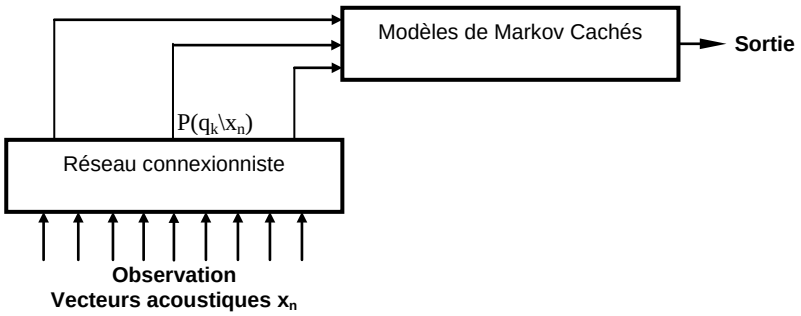


FIG 5.3 *Structure du modèle hybride HMM/MLP*

2.3 Initialisation

Pour ne pas avoir à annoter une base de fenêtres d'analyse, ce qui réduirait l'intérêt des HMM, des HMM discrets particuliers ont été appris, permettant d'associer chacun des leurs états à des parties de mots à analyser. Ainsi, par programmation dynamique (algorithme de Viterbi) on peut aligner les observations sur les états et annoter une base d'apprentissage pour une première génération du MLP.

2.4 Apprentissage & Reconnaissance

L'apprentissage global du modèle HMM/MLP se fait itérativement selon l'algorithme EM (Expectation-Maximization) [LAZh03]. Alternativement, le MLP et les HMM sont évalués et estimés sur les bases de données pour maximiser leurs critères d'apprentissage.

Un ANN est utilisé pour décoder les bases de mots, représentés par des séquences de vecteurs acoustiques, et calculer les probabilités d'observation des états. Les probabilités de transition des HMM sont réévaluées par l'algorithme de Baum-Welch pour maximiser la vraisemblance (critère MLE) de chaque modèle sur sa classe mot étant donné les probabilités d'observation fournies par le ANN.

Ce jeu des HMM permet de décoder à nouveau les bases de mots et de créer une nouvelle annotation pour réapprendre le RN, selon l'une des procédures suivantes : L'application de l'algorithme de Viterbi, nous permet d'obtenir la séquence d'états la plus probable, ce qui permet d'attribuer une classe à chaque observation. Les fonctions forward-backward permettent d'évaluer les probabilités des états pour chaque observation et de leurs attribuer des distributions de probabilités cibles pour le RN. Son apprentissage par l'algorithme de rétro- propagation minimise la distance entre ses sorties et les annotations.

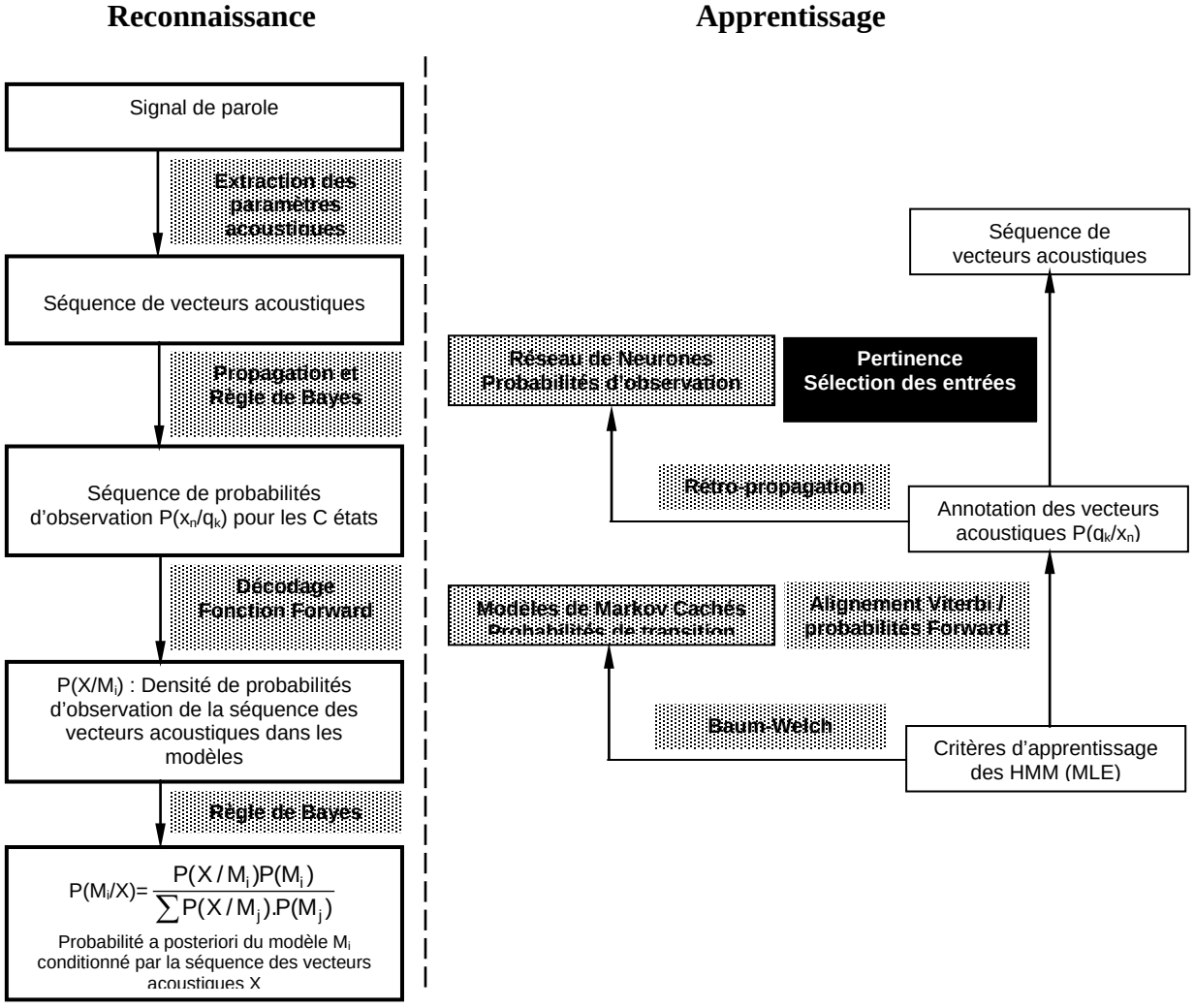


FIG 5.4 *Processus d'apprentissage et de reconnaissance du système hybride HMM/MLP*

Le MLP est utilisé comme classificateurs statistiques. Il permet de classifier des vecteurs acoustiques en différentes classes, chaque classe étant associée à un état stationnaire de l'ensemble Q. Le vecteur d'observation x_n est introduit aux entrées du MLP. Si l'ensemble Q contient K états stationnaires, le réseau présentera K nœuds de sortie. Etant donné x_n à l'entrée du MLP, on peut montrer que la sortie k de ce réseau est une estimation de la probabilité locale $P(q_k \backslash x_n)$. En utilisant la loi de Bayes :

$$P(q_k \mid x_n) = \frac{P(x_n \mid q_k) P(q_k)}{P(x_n)} \tag{5.9}$$

Il suffit de diviser cette probabilité locale par la probabilité a priori $P(q_k)$ pour obtenir un rapport de vraisemblance ("scaled likelihood") $P(x_n \mid q_k) / P(x_n)$. Comme pendant la reconnaissance, $P(x_n)$ est constant et ne modifie en rien la classification, on se ramène au formalisme des HMM présentés précédemment. Ce formalisme permet alors de modéliser le caractère séquentiel du signal de parole. Remarquons que l'architecture du réseau permet aisément d'introduire plusieurs vecteurs acoustiques consécutifs. Il suffit pour cela d'augmenter le nombre d'entrées du MLP. Cela a simplement pour conséquence d'augmenter la dimension des vecteurs d'entrées des perceptrons de la couche cachée.

Durant l'entraînement, les vecteurs d'observation x_n de l'ensemble d'entraînement sont consécutivement présentés aux entrées du MLP. L'entraînement est dit supervisé car on présente également au MLP les sorties désirées de celui-ci. La sortie associée à l'état stationnaire correspondant au vecteur d'entrée est forcée à 1 alors que les autres sorties sont à 0. L'algorithme opère alors par rétro-propagation de l'erreur d'estimation du vecteur de sortie du MLP et utilise la méthode itérative du gradient pour estimer les poids du réseau.

Nous avons montré que la probabilité globale $P(M \backslash X)$ peut s'exprimer en fonction des sorties d'un réseau de neurones estimant $P(q_k \backslash x_n)$. Par conséquent, en plus des avantages caractéristiques des réseaux de neurones, un système HMM/ANN bénéficie aussi de tous les avantages liés aux HMM, à savoir leur capacité à traiter les données séquentielles et leur possibilité d'entraîner l'ensemble des paramètres Θ (les paramètres du réseau de neurones dans le cas HMM/ANN) sans nécessiter la segmentation explicite de la base d'entraînement en termes de classes de Ω. Comme dans le cas HMM, il est donc possible d'estimer les paramètres Θ par un algorithme de type Baum-Welch ou Viterbi. Nous rapportons ici le mode d'entraînement à l'aide des fonctions avant-arrière décrit dans [BOI99].

Entraînement "avant-arrière" (Baum-Welch)

De façon équivalente aux approches HMM, l'entraînement des modèles a pour but d'estimer les paramètres

$$\Theta^* = \arg\max_{\Theta} \prod_{j=1}^{J} P(M_j \mid X_j, \Theta) \tag{5.10}$$

Dans le cas de l'entraînement "avant-arrière" (Baum-Welch) de systèmes hybrides HMM/ANN, on veut donc estimer l'ensemble des paramètres Θ maximisant P(M\X), c'est à dire :

$$\arg\max_{\Theta} \sum_{l_1,\ldots,l_N} \left[\prod_{n=1}^{N} P\left(q_{l_n}^n \mid X_{n-c}^{n+c}\right) \frac{P\left(q_{l_n}^n \mid M\right)}{P\left(q_{l_n}^n\right)} \right] P(M) \qquad (5.11)$$

Selon le même principe que l'entraînement des HMM, il est toujours possible de définir une fonction auxiliaire dont la maximisation est équivalente à la maximisation de (5.10). Nous pouvons alors utiliser une variante de l'algorithme EM pour entraîner un système hybride HMM/ANN :

- *Initialisation*
  - Choisir un réseau de neurones initial (ensemble de paramètres) et une distribution de probabilités a priori $P(q_k)$ initiale, ou
  - estimer les paramètres initiaux $\Theta^{(0)}$ du réseau ANN et les probabilités a priori à partir d'une segmentation initiale. Dans notre travail et pour ne pas avoir à annoter une base de fenêtres d'analyse, ce qui réduirait l'intérêt des HMM, des HMM discrets ont été appris, permettant d'associer chacun des leurs états à des parties de mots à analyser. Ainsi par programmation dynamique (algorithme de Viterbi) on peut aligner les observations sur les états et annoter une base d'apprentissage pour une première génération du ANN choisi (MLP dans notre cas).
- *Etape d'estimation* (E) : étant donné les paramètres $\Theta^{(t)}$ d'un réseau de neurones à l'itération *t* et les estimateurs de probabilités a priori $P(q_k)$, calculer
  - les nouvelles sorties désirées $P\left(q_k^n \mid X, M, \Theta^{(t)}\right)$ associées à chaque vecteur d'entraînement x_n, grâce aux récurrences α (Fonction Forward) et β (Fonction Backward) calculées à partir des sorties du réseau de neurones et des estimateurs courants de probabilités a priori $P(q_k)$;
  - les nouveaux estimateurs des probabilités a priori $P(q_k)$;
- *Etape de maximisation* (M) : nouvel entraînement (EBP) du réseau de neurones avec $P\left(q_k^n \mid X, M, \Theta^{(t)}\right)$(28) comme sortie désirée. On peut montrer que cet entraînement ANN conduit à un nouvel ensemble de paramètres $\Theta^{(t+1)}$ maximisant la fonction auxiliaire et garantissant donc que :

$$\prod_{j=1}^{J} P\left(M_j \mid X_j, \Theta^{(t+1)}\right) \geq \prod_{j=1}^{J} P\left(M_j \mid X_j, \Theta^{(t)}\right)$$

- *Itérer* : la convergence de ce processus itératif peut être démontrer en prouvant que (1) les nouvelles sorties désirées maximisent $P(M\backslash X)$ pour un ensemble de paramètres donnés et (2) que l'algorithme de gradient EBP, en convergeant vers ces sorties désirées, vont dans la direction d'un accroissement de $P(M\backslash X)$.

Reconnaissance

La reconnaissance de la parole par système hybride HMM/ANN se pratique selon le même principe que la reconnaissance HMM. Comme illustré dans la figure 5.5, le réseau de neurone est simplement utilisé comme estimateur de probabilités locales pour les HMM. Après division par les estimateurs de probabilités a priori, le réseau de neurones (MLP dans notre cas) fournit donc les vraisemblances normalisées $p(x_n \backslash q_k)/p(x_n)$ qui sont utilisées, soit dans un algorithme Viterbi, soit dans une récurrence "avant", afin d'estimer $P(M_j \backslash X)$ pour tous les modèles M_j possibles et d'assigner la séquence X au modèle M_k conduisant au maximum de probabilités a posteriori.

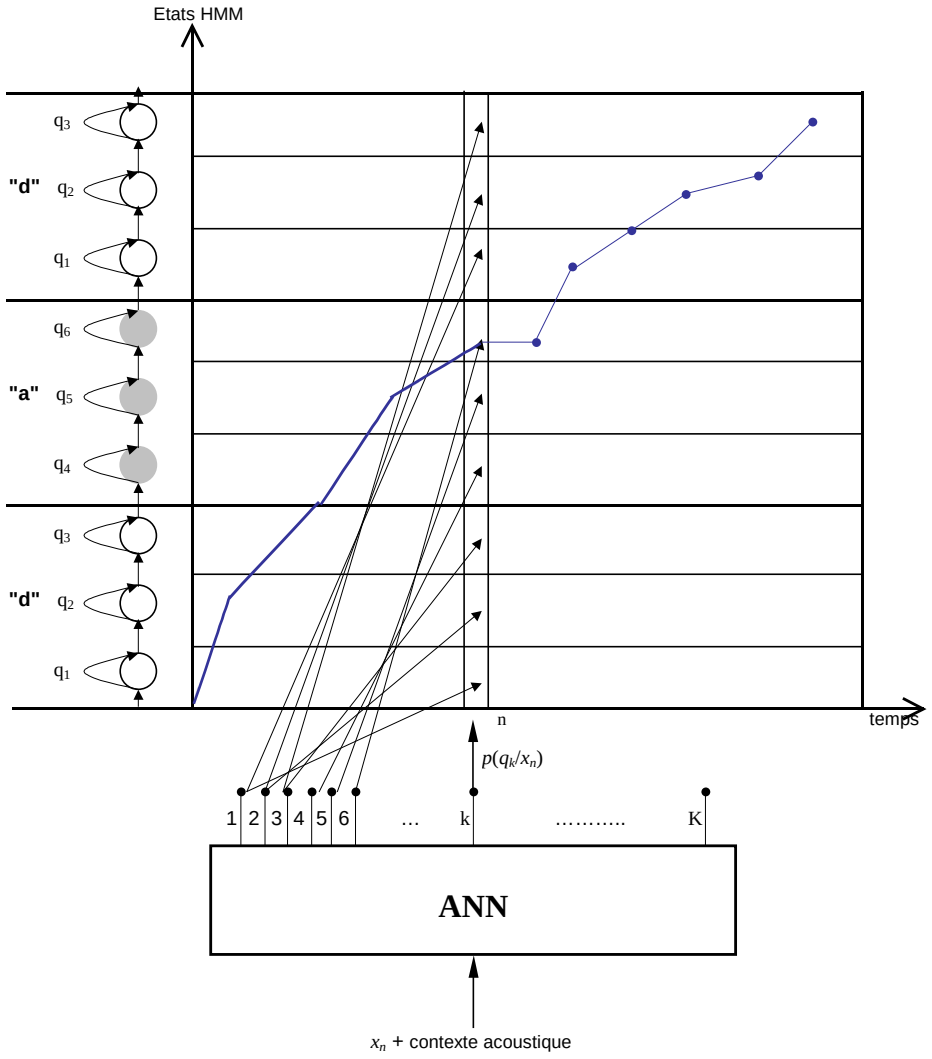


FIG 5.5 *Schéma de fonctionnement d'un système hybride HMM/ANN : pour chaque vecteur x_n (dans son contexte) présenté à l'entrée du réseau de neurones, celui génère l'ensemble des probabilités $P(q_k \backslash x_n)$ qui après division par les probabilités a priori $P(q_k)$ estimées sur l'ensemble d'entraînement, donnent les vraisemblances locales requises pour les différents états HMM.*

2.5 Lissage des probabilités a posteriori

Dans le cadre de ce travail, nous avons essayé d'intégrer une technique de lissage afin d'augmenter la capacité du système en se basant sur les idées décrites dans [BOI94]. Le principe est de combiner les probabilités a posteriori obtenues à la sortie du MLP avec ceux d'autres estimateurs (gaussiennes ou discrètes) du fait que l'entraînement supervisé du MLP selon le critère des moindres carrés estime uniquement mieux les probabilités ayant des valeurs grandes. Pour le travail considéré, nous avons multiplié les probabilités a posteriori fournies par le RN par les probabilités données à l'aide d'un estimateur discret. Ainsi, la probabilité d'observation du vecteur acoustique x_n dans l'état q_k réellement utilisée par le système hybride HMM/MLP est donnée par :

$$P(x_n / q_k) = \frac{P_{MLP}(q_k / x_n)}{P(q_k)} . P_d(x_n / q_k) \tag{5.12}$$

Où $P_{MLP}(.)$ et $P_d(.)$ représentent respectivement les probabilités fournies par le MLP et l'estimateur discret. Bien que les motivations théoriques de lissage des probabilités a posteriori ne soient toujours pas claires, ceci a uniformément mené au moins pour la tâche considérée ici à une amélioration significative au niveau du mot d'environ 2%.

3 APPORT DE L'HYBRIDATION HMM/MLP

Parmi les nombreux avantages de l'hybridation HMM/MLP, nous pouvons mentionner [BOI94] ; [LAZa02] ; [LAZb02] ; [LAZc02]; [LAZ05]; [ROB05] :

- Entraînement discriminant local au niveau de la fenêtre d'analyse entre les états HMM représentés par les sorties du MLP. L'entraînement à l'aide de l'algorithme Viterbi est basé sur la maximisation du critère de maximum de vraisemblance, ce qui le rend non discriminant.

- Aucune prétention au sujet des paramètres d'entrée. Ceci a deux avantages :

  1. Le MLP peut estimer n'importe quel genre de fonction de densité de probabilité et par conséquent, il peut également extraire la corrélation possible entre les composantes des vecteurs acoustiques.

  2. Lorsque plusieurs trames acoustiques sont présentées à l'entrée du MLP, le temps de corrélation entre les fenêtres successives peut aussi être modélisé.

- Les réseaux connexionnistes sont des structures fortement parallèles et régulières qui les rend particulièrement favorables du point de vu hardware et architecture de haute performance.

4 COMPARAISON DES DIFFERENTS MODELES

L'objectif de ce paragraphe est de faire une étude comparative entre le système hybride proposé et celui utilisant les HMM discrets pour un processus de reconnaissance de mots isolés prononcés en Arabe.

Dans toutes les expériences décrites dans ce rapport, la même topologie des HMM a été employée pour les trois types de modèles définis dans la suite de cette section.

4.1 Corpus utilisé

Trois bases de données ont été utilisées dans ce travail :

1) La première base (BD1) est lue par 30 locuteurs, chaque locuteur doit prononcer respectivement son nom & prénom, le nom des villes de naissance et résidence. Chaque son devrait être prononcé 10 fois. Nous avons choisit le vocabulaire d'une façon artificielle afin d'éviter les répétitions dans les noms des locuteurs et des villes, aussi bien pour les prénoms des locuteurs. Ce qui nous a permet d'avoir un vocabulaire de 1200 mots.

2) La deuxième base de données (BD2) est lue par les même locuteurs utilisés dans la première expérience. Ce vocabulaire contient 13 mots de commande (ex. sauvegarder\ sauvegarder sous\ sauvegarder tout\ précédent\ suivant, etc.) de sorte que chaque locuteur prononce chaque mot de commande 10 fois, ce qui donne un vocabulaire de 3900 sons.

3) La troisième base de données (BD3) est lue aussi par les même locuteurs utilisés dans la première expérience. Ce vocabulaire contient 10 chiffres arabes (*0*-*9*), de sorte que chaque chiffre locuteur prononce 10 fois chaque chiffre, ce qui donne un vocabulaire de 3000 sons.

Pour les données de test ont été énoncées par 8 locuteurs (4 hommes et 4 femmes) qui n'ont pas participé à l'entraînement du système et qui prononcent la séquence "nom – prénom – ville de naissance – ville de résidence" 5 fois concernant le premier corpus, 5 fois les mots de commande sélectionnés au hasard (le nombre des mots de commande prononcés par chaque locuteur est entre 5 à 10 mots), et prononcent aussi 5 fois les chiffres arabes.

Pour des fins de reconnaissance automatique de mots isolés prononcés en Arabe. L'objectif de nos expériences est de faire en premier une étude comparative entre (1) le système HMM discrets (2) le système hybride HMM/MLP utilisant des distributions discrètes (application de l'algorithme c-moyennes pour la segmentation acoustique) (3) le système hybride HMM/MLP utilisant des distributions discrètes floues (application de l'algorithme c-moyennes floues pour la segmentation acoustique) (4) le système hybride HMM/MLP utilisant les AG pour la segmentation acoustique.

4.2 Paramètres Acoustiques

L'ensemble de nos travaux est basé sur l'utilisation des vecteurs acoustiques de type log RASTA-PLP (RelActive SpecTrAl processing – Perceptral Linear Predictive) [HER94] ainsi que les MFCC (Mel-scale Frequency Cepstral Coefficients) [DAV80]. Ces paramètres sont calculés toutes les 10 ms avec un chevauchement de 25 ms. Chaque trame est représentée par 12 composantes plus l'énergie (MFCC\ log RASTA-PLP + E). Les valeurs des 13 coefficients sont normalisées par leur écart-type mesuré sur l'ensemble des trames d'apprentissage. Ainsi le vecteur acoustique fournit à l'entrée de notre MLP est composé de 26 paramètres (les paramètres cepstraux, leurs dérivées premières ainsi que les dérivées première et seconde de l'énergie). De plus, 9 trames de contexte acoustique sont utilisées correspondant à la configuration connue pour donner les meilleurs résultats, dans le sens que chaque vecteur acoustique courant est précédé par 4 vecteurs acoustiques dans le contexte gauche et suivi par 4 vecteurs acoustiques dans le contexte droit [BOI94].

4.3 Reconnaissance par le modèle 1 – HMM discret

Les vecteurs acoustiques ont été quantifiés en 4 dictionnaires selon le principe de l'algorithme de k-means comme suit :

- 128 prototypes pour les coefficients log RASTA-PLP/MFCC
- 128 prototypes pour les Δ (log RASTA-PLP/MFCC)
- 32 prototypes pour la dérivée première de l'énergie ΔE
- 32 prototypes pour la dérivée seconde de l'énergie $\Delta\Delta$E

Des modèles de mots à 10 états ont été utilisés pour modéliser chacun des unités élémentaires (mots). Notons seulement que le choix de 10 états par modèle a été fait d'une façon empirique.

4.4 Reconnaissance par le modèle 2 – Modèle hybride HMM/MLP avec des entrées fournies par k-means

Un MLP possédant en entrée 2880 neurones correspondant à 9 trames de contexte. Le vecteur binaire fournit à l'entrée du réseau composé seulement de 36 bits à "1" (obtenu en appliquant les concepts de la quantification vectorielle "k-means"). Une couche cachée de taille variable, une couche de sortie composée d'autant de neurones qu'il y a d'états HMM. Le nombre de neurones de la couche cachée a été choisit de manière à satisfaire la règle heuristique suivante [JOD94] :

Nombre de neurones cachés = (nombre de neurones d'entrée*nombre de neurones de sortie)$^{1/2}$

Ainsi un MLP à une seule couche cachée comprenant 2880 neurones à l'entrée, 288 neurones pour la couche cachée et 10 neurones de sortie a été entraîné.

4.5 Reconnaissance par le modèle 3 – Modèle hybride HMM/MLP avec des entrées fournies par FCM

Pour ce dernier cas, nous avons essayé de comparer la performance du modèle hybride précédent avec celle du modèle hybride HMM/MLP utilisant en entrée du réseau un vecteur acoustique composé de valeurs réelles qui ont été obtenues en appliquant l'algorithme FCM. Nous avons présenté chaque paramètre cepstrale (log RASTA-PLP/MFCC, Δlog RASTA-PLP/ΔMFCC, ΔE, ΔΔE) par un vecteur réel dont les composantes définissent les degrés d'appartenance du paramètre aux différentes classes des "code-book". La topologie du MLP est similaire au modèle 2, néanmoins la couche d'entrée est composée d'un vecteur réel avec 2880 composantes réelles correspondant aux différents degrés d'appartenance des vecteurs acoustiques aux classes des "code-book".

4.6 Modèle 4 - Modèle hybride HMM/MLP avec des entrées fournies par les AG

Pour ce dernier cas, les vecteurs acoustiques quantifiés qui sont présentés à l'entrée du MLP, sont fournis par le biais de l'application des AG. Les paramètres de l'AG utilisé sont récapitulés dans le tableau suivant :

TAB 5.1 *Paramètres de l' AG*

Nombre de générations maximal	100
Taille population	08
Probabilité de croisement	0.7
Probabilité de mutation	0.2
Pourcentage de remplacement de la Ω	0.5

4.6 Résultats et discussion

Ces quatre types de modèles ont été comparés dans le cadre d'une reconnaissance de mots arabes isolés. Le tableau 2 et la figure 6 résume les différents résultats obtenus des modèles utilisant seulement les paramètres acoustiques fournis par une analyse log RASRA-PLP. Tous les essais d'entraînement et de test des trois modèles ont été effectués sur la BD1 en premier puis sur la BD2 et enfin sur la BD3.

TAB 5.2 *Taux de reconnaissance pour les quatres types de modèles. Coefficients log RASTA-PLP*

Corpus utilisé	***Modèle 1***	***Modèle 2***	***Modèle 3***	***Modèle 4***
BD 1	79.3%	81.9%	84%	82.7%
BD 2	78.7%	83.9%	84.7%	85.2%
BD 3	83.4%	84.2%	85.8%	85%

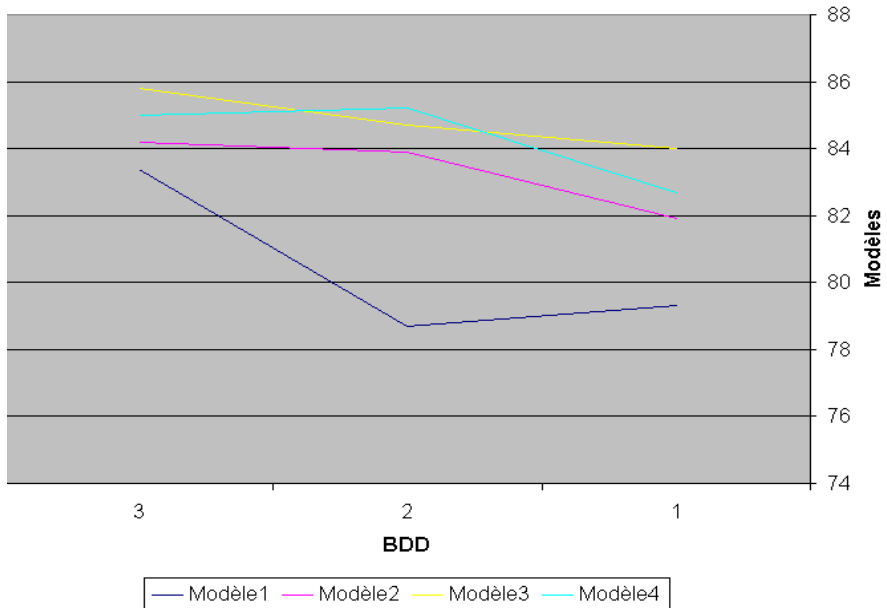


FIG.5.6 *Taux de reconnaissance pour les quatres types de modèles.*

Les taux de reconnaissance obtenus ainsi que l'étude des erreurs commises lors du processus de reconnaissance montrent que :

1. L'approche du modèle hybride HMM/MLP discret (modèle 2, 3 et 4) est toujours plus performante que celle des HMM discrets.

2. Le modèle hybride HMM/MLP utilisant en entrée du MLP, un vecteur dont les composantes sont obtenues en appliquant l'algorithme FCM et des AG, a donné les meilleurs résultats pour les trois corpus utilisés.

5 CONCLUSION

Malgré que les HMM bénéficient d'algorithmes d'entraînement et de décodage performants. Néanmoins, les hypothèses nécessaires à la mise en oeuvre de ces algorithmes peuvent pénaliser les performances de ces modèles. Les principales hypothèses les plus contraignantes sont :

- Entraînement non discriminant (maximisation de la vraisemblance au lieu des probabilités a posteriori).
- Forme des densités de probabilité fixée (multi-gaussiennes ou discrète).
- Les composantes des vecteurs acoustiques sont supposées non corrélées.
- La séquence des états est un processus de Markov du premier ordre.
- Pas de contexte acoustique pris en compte[2]. Aucune corrélation entre les vecteurs acoustiques n'est directement modélisable.
- Le formalisme est rigide, l'intégration d'autres sources de connaissance (syntaxe, sémantique ...) est difficile.

Notons tout de même que la plupart des systèmes de reconnaissance proposés sur le marché actuellement sont basés sur ce type de technique[3].

Certaines de ces hypothèses peuvent être supprimées (ou adoucies) en utilisant conjointement les modèles de Markov et un réseau de neurones.

Les premières expériences réalisées au sein de ce travail ont permis de fixer les performances de base pour les modèles hybrides HMM/MLP. Des taux de reconnaissance meilleurs ont été observés pour des locuteurs n'ayant pas participé à l'entraînement pour une reconnaissance de mots arabes isolés. Du point de vue efficacité des modèles, il semble évident (cf.§4) que les modèles hybrides sont au moins aussi performants voir même supérieurs que les HMM discrets. De nombreuses publications ont montré l'avantage de combiner les réseaux de neurones et les modèles de Markov cachés en reconnaissance de la parole (isolée ou continue) indépendamment du locuteur pour de petits ou de grands vocabulaires [BER99] ; [HAGa00] ; [HAGb00] ; [MOR00] ; [MORb01]. Les avantages dans l'utilisation cumulée des HMM et des MLP se reflètent bien dans les résultats obtenus.

---

2. Une solution à ce problème proposée par *Furui* [FUR86] puis par *L88ee* [LEE] consiste à utiliser les dérivées des vecteurs acoustiques. Une amélioration sensible du taux de reconnaissance est observée, mais ce n'est pas la solution à ce problème.

3. Watson d'AT&T, Voice type d'IBM et Easy Speaking de Dragon Dictate …

De plus, nous avons proposé de nouvelles méthodes pour la segmentation des vecteurs acoustiques: l'algorithme fuzzy c-means et l'algorithme génétique; ces deux dernières nous ont permis d'obtenir une meilleure performance du système hybride HMM/MLP proposé que celui utilisant l'algorithme classique k-means.

Plusieurs résultats récents en reconnaissance automatique de la parole (obtenus sur différentes bases de données allant des petits lexiques aux très grands lexiques) ont montré que les systèmes hybrides HMM/ANN conduisent généralement à des performances de reconnaissance équivalentes ou meilleures que celles des systèmes HMM utilisés dans les mêmes conditions. Néanmoins l'un des principaux défauts liés à ces modèles hybrides réside dans le fait que le nombre de paramètres est en quelque sorte borné. En effet, aucune amélioration n'est généralement observée (comme habituellement pour les HMM continus) lorsque le nombre des données d'entraînement et / ou de paramètres est fortement augmenté.

Nous proposons dans le sixième chapitre, une nouvelle méthode visant à explorer ce problème. Cette méthode est basée sur des expériences qui ont déjà montré qu'il est possible d'améliorer sensiblement les performances des systèmes en combinant plusieurs modèles.

Chapitre 6

Méthode de fusion de données

Pour pallier la limitation des systèmes hybrides HMM/MLP, nous proposons dans ce chapitre une nouvelle méthode visant à explorer le problème d'augmentation du nombre des données d'apprentissage. Nous rapportons les résultats des expériences qui ont déjà montré qu'il est possible d'améliorer sensiblement les performances des systèmes hybrides en combinant plusieurs modèles.

Chapitre 7

Conclusion générale

Les modèles hybrides sont de plus en plus utilisés en reconnaissance de la parole, ils sont couramment employés de nos jours pour nommer des systèmes qui tentent d'allier les avantages de plusieurs méthodes différentes pour augmenter le taux de reconnaissance et arriver à des systèmes avec des performances comparables à celles des humains.

1 EXTRACTION DES PARAMETRES ACOUSTIQUE

Tout d'abord, pour extraire les paramètres acoustiques qui seront fournis au système de reconnaissance, nous avons utilisé principalement les coefficients log RASTA-PLP (RelActive SpecTrAl processing-Perceptral Linear Predictive) dans les expériences décrites dans ce document. La justification de l'utilisation de ces coefficients est leur indépendance vis-à-vis du canal de transmission, et donc leur possible utilisation dans un système de démonstration sans dégradations significatives des taux de reconnaissance par l'utilisation d'un microphone standard différent de ceux utilisés lors de l'enregistrement des bases de données ayant servi à l'entraînement du système. Nous rapportons à titre de comparaison quelques résultats obtenus avec des coefficients MFCC (Mel-scale Frequency Cepstral Coefficients).

2 MODELE D'ENTRAÎNEMENT ET DE RECONNAISSANCE

Malgré que les HMMs bénéficient d'algorithmes d'entraînement et de décodage performants néanmoins, les hypothèses nécessaires à la mise en œuvre de ces algorithmes peuvent pénaliser les performances de ces modèles parmi lesquelles le fait que les vecteurs acoustiques sont supposés non corrélés ou encore l'hypothèse sur la distribution des densités de probabilités de chaque état HMM (distribution discrètes ou multi-Gaussiennes). Afin de supprimer certaines de ces hypothèses contraignantes, nous avons utilisé les HMMs conjointement avec un MLP, le but de l'hybridation avec un MLP dans le système proposé est de mieux évaluer les probabilités d'observation des états, non plus par une modélisation de l'espace des vecteurs acoustiques mais par un apprentissage discriminant au niveau local de la fenêtre d'analyse. Un unique MLP répond pour chaque observation les probabilités a posteriori de tous les états possibles des HMMs hybrides. Puis par règles de Bayes, les probabilités d'observation (scaled likelihood) des états sont calculées en normalisant par les probabilités a priori des classes. Les transitions sont toujours apprises par les HMMs qui évaluent les vraisemblances des classes mots. Enfin par règle de Bayes, les probabilités a posteriori des classes mots sont calculées. Les avantages principaux de cette hybridation est que les MLPs ne nécessitent pas d'hypothèses sur la forme des distributions statistiques associées à chaque état des HMMs. Du fait de l'entraînement discriminant des ANNs (ce qui est une de leurs propriétés majeures), on aboutit à des HMMs avec discrimination locale (au niveau de la fenêtre d'analyse). D'autre part, l'utilisation de l'information temporelle est plus aisée avec ce type de système : Il est facile de fournir plusieurs vecteurs acoustiques à l'entrée du MLP. Une information contextuelle est donc prise en compte dans les probabilités estimées et la corrélation entre des fenêtres successives n'est pas négligée. Pour diverses raisons, cela n'est pas possible avec les HMMs classiques. Nous avons rapporté à titre de comparaison le résultat obtenu du système utilisant uniquement les HMMs discrets. Une réduction significative des taux d'erreur par rapport aux HMMs discrets a été observée avec le modèle hybride HMM-MLP.

3 METHODES DE SEGMENTATION

Enfin, afin d'augmenter la performance du système hybride de reconnaissance proposé, nous avons développé dans le cadre de ce travail :
D'un part, un nouveau algorithme de partitionnement flou, de type FCM pour la segmentation de la parole arabe, du fait que la quantification vectorielle (l'algorithme k-means dans notre application) fournit une décision dure non probabilisée qui ne transmet pas assez d'informations sur les observations. De plus, les frontières entre les unités élémentaires du signal, ne sont pas acoustiquement définies convenablement. L'intérêt de l'algorithme FCM réside dans l'utilisation d'une dissimilarité pour mesurer par des variables non floues les vecteurs acoustiques. Tandis, l'appartenance de ces vecteurs acoustiques aux classes est floue. FCM utilise des centres de gravité comme représentants de classes. Ensuite, le critère optimisé par cet algorithme repose sur l'écart entre un vecteur acoustique et son centre de gravité qui fait intervenir la dissimilarité.

D'une autre part, concernant le deuxième algorithme proposé, nous nous sommes intéressés plus particulièrement, aux méthodes impliquant une classification supervisée par partition et nous avons retenu une comme base pour notre travail. Cette solution consiste à faire le choix d'une mesure que nous utilisons dans notre application. Cet algorithme cherche une "bonne" partition relativement à un critère qui mesure la qualité d'une partition. Nous sommes donc ramenés à un problème d'optimisation. Les propriétés de cet algorithme ne garantissent pas la convergence vers un optimum global, c'est pourquoi nous nous sommes intéressés à une heuristique de type Algorithmes Génétiques (AG), moins susceptibles d'être piégés par les minima locaux et désormais largement employés dans les problèmes d'optimisation.. Si sur un plan théorique, aucun résultat général ne prouve que cette méthode conduise à une solution optimale, en pratique la convergence globale est souvent constatée.

Par la suite, chaque paramètre cepstrale est présenté par un vecteur réel dont les composantes définissent les degrés d'appartenance du paramètre aux différentes classes. Ce vecteur réel est fournit à l'entrée du MLP pour calculer les probabilités d'observation du vecteur acoustique dans les différents états d'HMM. Nous avons rapporté à titre de comparaison les résultats obtenus avec le système hybride de base. Nous avons observé encore une réduction significative des taux d'erreur par rapport au système hybride HMM-MLP utilisant à l'entrée du MLP un vecteur binaire dont les composantes sont fournies en appliquant le principe de l'algorithme classique k-means. Des expériences préliminaires au niveau du mot, utilisant des vocabulaires de tailles 1200 et 3900 mots ont été rapportées.

4 METHODE DE FUSION DE DONNEES

Nous avons constaté que ces modèles hybrides HMM\MLP souffrent de nombreux défauts parmi lesquelles le fait que le nombre de paramètres est en quelque sorte borné. En effet, aucune amélioration n'est généralement observée (comme habituellement pour les HMMs continus) lorsque le nombre des données d'entraînement et/ou de paramètres est fortement augmenté.

Pour pallier cette limitation des systèmes hybrides HMM/MLP, nous avons proposé dans cette thèse, une nouvelle méthode visant à explorer ce problème. Cette méthode est basée sur des expériences qui ont déjà montré qu'il est possible d'améliorer sensiblement les performances des systèmes hybrides en combinant plusieurs modèles. A la base, l'hypothèse est que, si les modèles sont entraînés sur différentes parties du fichier d'entraînement, ils vont sélectionner des propriétés différentes des données, permettant ainsi une amélioration des résultats lorsque les sorties sont combinées. Lors de la reconnaissance, les modèles sont tous les deux utilisés et la sortie correspondant au meilleur score est sélectionnée. Ces modèles sont combinés selon plusieurs critères pour fournir le mot le plus probable.

5 CONCLUSION FINALE

Nous avons présenté dans ce document l'étude et le test de performance d'un modèle hybride qui combine efficacement la technologie la plus utilisée en reconnaissance de la parole : Les modèles de Markov cachés (HMM – Hidden Markov Models) et des réseaux de neurones artificiels (ANN – Aritificial Neural Networks) particulièrement les perceptrons multi-couches (MLP – Multi-Layer Perceptrons) pour l'entraînement et la reconnaissance de la parole arabe isolée indépendante de locuteurs pour un vocabulaire de 1200 et 3900 sons. Nous avons décrit dans le même cadre le principe de deux nouveaux algorithmes (FCM – Fuzzy C-Means) qui repose sur les concepts de la logique floue et l'algorithme non supervisé basé sur le principe des algorithmes génétiques pour la segmentation de la parole arabe. Le système hybride proposé a été testé et comparé à un reconnaisseur utilisant les HMMs discrets pour une tâche de reconnaissance améliore fortement les taux de reconnaissance et c'est à notre connaissance, la première fois que ce genre de modèles hybrides sont utilisés pour effectuer une reconnaissance de la parole arabe.

Nous avons défini dans cette thèse, une méthode permettant de diviser en plusieurs parties l'ensemble d'entraînement et d'entraîner plusieurs MLP sur chacune de ces parties. Nous espérons ainsi tirer profit de l'entraînement des réseaux sur des données filtrées par la procédure de fusion mettant en exergue des propriétés différentes du signal. Différents types de combinaisons des systèmes ont été testés :

- La combinaison linéaire.
- La combinaison linéaire dans le domaine logarithmique.
- La combinaison par le critère entropique.
- La combinaison par un MLP.

Une réduction significative du taux d'erreur a pu être observée en utilisant la méthode de fusion décrite dans ce document par rapport au système hybride de base (40 % pour 60 mots, 13 % pour 150 mots et 9% pour 700 mots) pour des distributions discrètes floues, et 37 % pour 60 mots, 12 % pour 150 mots et 9.5% pour 700 mots, pour des distributions obtenues par les AG. Cette procédure nous a permis de tirer au mieux parti des nombreuses données d'entraînement dont nous disposions. Cette amélioration, obtenue dans le cadre d'une reconnaissance de mots isolés arabe, devrait aussi être constatée pour un système de reconnaissance de la parole continue. Dans cette optique, la même procédure décrite pourra être appliquée et le même système utilisé.

Il semble que la méthode de combinaison des sorties des MLP la plus efficace (du moins pour l'expérience décrite ici), Il semble que la méthode de combinaison des sorties des MLP la plus efficace (du moins pour l'expérience décrite ici), dans les deux cas, consiste à combiner les sorties des trois réseaux de neurones par le biais d'un MLP classique.

Un seul petit inconvénient à la méthode : il est nécessaire de faire tourner en parallèle trois réseaux. Les temps de reconnaissance sont donc plus importants que pour le système de base. Ils restent cependant plus qu'acceptables. De plus il faut pouvoir disposer d'un nombre important de données d'entraînement, le peu de données que nous possédons risque d'être un facteur limitatif de l'amélioration que nous pourrions observer.

6 PERSPECTIVES

Malgré que le modèle hybride et la méthode de fusion proposés étaient plus performant que les HMM classiques qui sont à la base de la presque totalité des outils de reconnaissance de la parole présents sur le marché. Cependant, les taux de reconnaissance obtenus sont au moins bons que nous ne l'espérions. Nous prévoyons donc d'explorer les voies suivantes pour améliorer notre système :

- Pour améliorer la performance du système proposé, il serait alors important d'utiliser d'autres techniques d'extraction de paramètres et comparer le taux de reconnaissance du système avec celui utilisant l'analyse acoustique log RASTA-PLP. Nous pensons utiliser les techniques LDA (Analyse Discriminante Linéaire) et CMS (Cepstral Mean Substraction) du fait que ces représentations sont considérées actuellement parmi les plus performantes en RAP.

- Malgré que les techniques FCM et AG ont amélioré la performance du système hybride proposé néanmoins, ces algorithmes souffrent de quelques défauts soulevés aussi par les méthodes classiques de classification (k-means, nuées dynamiques, etc.) qui sont la nécessité de connaître a priori, le nombre de classes, la sensibilité au choix de la configuration initiale ainsi que la convergence vers des minima locaux. Nous avons essayé dans un travail antérieur [LAZc03] ; [LAZd03] ; [LAZg03] d'appliquer une nouvelle méthode de segmentation de la parole et qui a permis de pallier les principaux défauts soulevés, une comparaison de performance a déjà été réalisée avec celle du FCM et qui a donné des résultats très prometteurs. Nous penserons intégrer cette méthode dans le système de RAP proposé.

- Il paraît intéressant aussi d'utiliser des HMM continus avec une distribution multi-gaussiennes et comparer la performance du système avec celle des HMM discrets.

- D'autre part, pour un vocabulaire étendu, il est intéressant d'utiliser des modèles de phonèmes au lieu de mots, du fait que le nombre de phonèmes qui permet la construction de n'importe quel mot présent dans la plupart des langages est faible, ce qui facilite l'entraînement avec des bases relativement petites.

Références

[ANA95] A. Anastasakos, R. Schwartz, and H. Shu, ``Duration modeling in large vocabulary speech recognition," *in Proceedings of the International Conference on Acoustics, Speech and Signal Processing., vol. 1, pp. 628--631, IEEE, may 1995*.

[ATA91] B. Atal, “effectiveness of linear prediction characteristics of a speech wave for automatic speaker identification and verification," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, pp. 305--308, 1991*.

[AUB93] X. Aubert, Haeb-Umbach, and H. .Ney, ``Improvement in connected digit recognition using linear discriminant analysis and mixture densities," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 648--651, 1993.*

[AVE86] A. Averbuch, L. Bahl, R. Bakis, P. Brown, A. Cole, G. Daggett, S. Das, K. Davies, S. De Gennaro, P. de Souza, E. Epstein, D. Fraleigh, F. Jelinek, S. Katz, B. Lewis, R. Mercer, A. Nadas, D. Nahamoo, M. Picheny, G. Shich- man, P. Spinelli, ``An IBM-PC Based Large-vocabulary Isolated-utterance Speech Recognizer", *ICASSP 86, pp. 53-56, 1986*.

[AVE87] A. Averbuch et al. (collectif de 21 auteurs), "Experiments with the Tangora 20,000 word speech recognizer", *Proc. ICASSP, pp. 701-704, Dallas, 1987*.

[BEL57] R.E. Bellman, Dynamic Programming, Princeton Univ. Press, 1957.

[BAU72] L. E. Baum, “An inequaly and associated maximization technique in statistical estimation of probabilistic functions of markov processes.," *Inequalities 3, pp. 1--8, 1972*.

[BAK74] J.K. Baker, ``The Dragon Automatic Speech Recognition System", *Speech Communication Seminar, Stockholm, pp. 1-12, 1974.*

[BAK75] J. K. Baker, ``The dragon system -- an overwiew.", *IEEE Transactions on Acoustics Speech and Signal Processing, vol. 1, pp. 24--29, February 1975. ASSP-23.*

[BAK76] R. Bakis, ``Continuous speech word recognition via centisecond acoustic states," *In 91st Meeting of the Acoustical Society of America, April 1976*.

[BRI82] J.S. Bridle, M.D. Brown, & R.M. Chamberlain, "An algorithm for connected word recognition", *Proc. ICASSP, pp. 899-902, Paris, 1982*.

[BAH83] L. R. Bahl, F. Jelinek, and R. L. Mercer, `à maximum likelihood approach to continuous speech recognition," *IEEE Trans. on Patt. Anal. and Machine Intell., vol. PAMI-5, pp. 179--190, 1983*.

[BAH87] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, ``Speech recognition with continuous-parameter hidden markov models," *Computer Speech and Language, vol. 2, pp. 219--234, 1987.*

[BEZ81] J. C. Bezdek. "Pattern Recognition with Fuzzy Objective Function Algorithms". *Plenum Press*, New York, 1981.

[BEZ87] J. Bezdek, "Pattern Recognition With Fuzzy Objective Function Algorithms". *Second Edition, New-York: Plénum, 1987.*

[BEZ99] J.C. Bezdek. "Fuzzy models and algorithms for pattern recognition and image processing". *Handbooks of fuzzy sets series* ; FSHS 4. Boston: Kluwer Academic. xv, 776, 1999.

[BOI87] R.Boite, M.Kunt, « Traitement de la parole », *Lausanne, Presses, Polytechniques romandes, 1987.*

[BAH88] L.R. Bahl, R. Bakis, P.V. de Souza, R.L. Mercer, ``Obtaining Candidate Words by Polling in a Large Vocabulary Speech Recognition System", *ICASSP 88, pp. 489-492, 1988*.

[BAH89] L.R. Bahl, R. Bakis, J. Bellegarda, P.F. Brown, D. Burshtein, S.K. Das, P.V. de Souza, P.S. Gopalakrishnan, F. Jelinek, D. Kanevsky, R.L. Mercer, A.J. Nadas, D. Nahamoo, M.A. Picheny, ``Large Vocabulary Natural Language Continuous Speech Recognition", *ICASSP 89, pp. 465-467, 1989*.

[BAH90] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo, ``Constructing Groups of Acoustically Confusable Words", *ICASSP 90, pp. 85-88, 1990*.

[BEL90] J.R. Bellegarda & D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing, vol. no. pp. 2033-2045, 1990*.

[BOU90] H. Bourlard, N. Morgan., and C. Wellekens, ``Statistical inference in multilayer perceptrons and hidden markov models with applications in continuous speech recognition," *Neurocomputing Algorithms ,Architectures and Applications, pp. 217--226, 1990. F. FogelmanSoulie and J. Herault (eds), NATO ASI Series*.

[BAH91] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, M.A. Picheny, ``Decision Trees for Phonological Rules in Continuous Speech", *ICASSP-91, pp. 185-188, 1991*.

[BEN92] Y. Bengio, R. De Mori, G. Flammia & R. Kompe, "Phonetically motivated acoustic parameters for continuous speech recognition using artificial neural networks", *Speech Communication, vol. 11, pp. 261-271, 1992.*

[BOI94] J-M. Boîte, H. Bourlard, B. D'Hoore, *al*, "Task independent and dependent training: performance comparison of HMM and hybrid HMM/MLP approaches". *IEEE 1994, volI, pp.617-620, 1994*.

[BOU94] H. Bourlard and N. Morgan, "Connectionist Speech Recognition-A Hybrid Approach". *Kluwer Academic Publisher, 1994.*

[BAH95] L.R. Bahl, S. Balakrishnan-Aiyer, M. Franz, P.S. Gopalakrishnan, R. Gopinath, M. Novak, M. Padmanabhan, S. Roukos, ``The IBM Large Vocabulary Continuous Speech Recognition System for the ARPA NAB News Task", *95 ARPA SLT Workshop, pp. 121-126, 1995*.

[BOU97] H. Bourlard, S. Dupont, "Sub-band-based speech recognition". *In Proc, IEEE Internat, Conf, Acoust, Speech and Signal Process, Munich, pp.1251-1254, 1997.*

[BER99] F. Berthommier, H. Glotin. "A new snr-feature mapping for robust multistream speech recognition". *In Berkeley University of California, editor, Proc.Int.Congress on Phonetic Sciences (ICPhS), Vol1 of XIV, pp. 711-715, Sanfrancisco, 1999.*

[BAL99] L. O. Ball, B. Ozyurt, & J. C. Bezdek. "Clustering with a genetically optimized approach". *IEEE Transactions on Evolutionnary Computation*, 3(2) :103–112, 1999.

[CHA81] J-L. Chandon & S. Pinson, "Analyse typologique, Théories et applications". *Masson, 1981*.

[CEL89] G. Celeux, E. Diday, *al*, "Classification automatique des données". *Dunod, 1989*.

[CHO87] K. Choukri, "Quelques approches pour l'adaptation aux locuteurs en reconnaissance

automatique de la parole ", *Thèse de l'ENST, 1987*.

[CAL89] Calliope, "La parole et son traitement automatique". *Masson et CNET-ENST, Paris, 1989*.

[CHI91] K. Chidananda Gowda & E. Diday, "Symbolic clustering using a new dissimilarity measure". *Pattern Recognition Letters, 24(6) : pp. 567–588, 1991*.

[CHI92] B. Chigier & H. C. Leung, "The effects of signal representations, phonetic classification techniques, and the telephone network", *Proc. ICSLP, pp. 97-100, Banff, 1992*.

[DEM77] A. P. Demster, N. M. Laird, and D. B. Rubin, ``Maximum likelihood from incomplete data via the em algorithm.," *Journal Of the Royal Statistical Society, 1977. Series B 34.1-38*.

[DAV80] S-B. Davis, P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". *Proceedings of the International Conference on Acoustics, Speech and Signal processing, pp. 357-366, 1980*.

[DUP96] S. Dupont, H. Bourlard, and C. Ris, "Multi-stream speech recognition," *Tech Rep IDIAP-RR 96-07,. IDIAP, Martigny, 1996*.

[DER97] O. Deroo, C. Rii, *al*, "Hybrid HMM/ANN System for speaker independent continuous speech recognition in French". *Faculté Polytechnique de Mons – TCTS, Belgium, 1997*.

[DIA00] J. Diatta, I. Kojadonovic, & H. Ralambondrainy, "Une mesure de dissimilarité pour les données hétérogènes floues". *In Logique Floue et applications (LFA 2000), 2000*.

[ELB90] M. El-Bèze, « Choix d'unités appropriées et introduction de connaissances dans des modèles probabilistes pour la reconnaissance automatique de la parole », *Thèse de l'Université Paris 7, 1990*.

[ELS98] Y. El-Sonbaty & M-A. Ismail, "Fuzzy clustering for symbolic data". *IEEE Transactions on fuzzy systems, 6(2) : pp. 195–204, 1998*.

[FOR73] G. D. Forney, ``The viterbi algorithm". *Proceedings of the IEEE, vol. 61, pp. 268--278, March 1973*.

[FURa86] S. Furui, "Speaker independent isolated word recognizer using dynamic features of speech spectrum". *Proceedings of the International Conference on Acoustics, Speech and Signal Processing., vol. 34, pp. 52--59, Tokyo, Japan, April 1986.*

[FURb86] S. Furui, "Speaker independent isolated word recognition based on emphasized spectral dynamics". *Proceedings of the International Conference on Acoustics, Speech and Signal Processing., pp. 1991--1994, Tokyo, Japan, April 1986.*

[FON97] V. Fontaine, C. Ris, and J. Boite, ``Non linear discriminant analysis for improved speech recognition," *Proceeding of the European Conference On Speech Communication and Technology., vol. 4, pp. 2071-- 2075, 1997*.

[FIS88] L. Fissore, P. Laface, G. Micca, R. Pieraccini, ``Very Large Vocabulary Isolated Utterance Recognition: a Comparison Between One Pass and Two Pass Strategies", *ICASSP 88, pp. 203-206, 1988*.

[FIS89] L. Fissore, P. Laface, G. Micca, R. Pieraccini, ``Lexical Access to Large Vocabularies For Speech Recognition", *IEEE Trans. on ASSP, Vol. 37, No. 8, pp. 1197-1989, Août 1989*.

[GUP88] V.N. Gupta, M. Lennig, P. Mermelstein, ``Fast Search Strategy in a Large Vocabulary Word Recognizer", *Journal Acoust. Soc. Am., Vol. 6, pp. 2007- 2017, Décembre 1988.*

[GAG90] C. Gagnoulet & D. Jouvet, "Reconnaissance de la parole et modélisation statistique: expérience du CNET", *Traitement du signal, vol. 7, no. 4, pp. 267-274, 1990*.

[GAL90] E.Gallais, P.Alinat, G.Souvay, J.M.Pierrel, « Intégration d'un système de reconnaissance analytique de la parole dans une console sonar : Vers un dialogue naturel », *Traitement du signal, Vol 7, n°4, pp. 367-379, 1990*.

[GAU94] J.-L. Gauvain & C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains", *IEEE Trans. on Speech and Audio Processing, vol. 2, no. 2, pp. 291- 298, 1994.*

[GHI94] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition", *IEEE Trans. on Speech and Audio Processing, vol. 2, no. 1, pp. 115-132, 1994*.

[GOL94] D. E. Goldberg. "Algoritmes génétiques - Exploration, optimisation et apprentissage automatique". *Addison Wesley*, 1994.

[GLO95] H.Glotin, F.Berthommier, E.Tessier, H.Bourlard, "Interfacing of CASA and Multistream recognition", *Proceedings of EuroSpeech'95, pp. 135-138, 1995*.

[GOR98] N. Gorsky, V. Anisimov, *al*, "A new A2iA Bankcheek Recognition System". *Third European Workshop on Handwriting Analysis and Recognition. IEE 1998, pp? 1998.*

[HAN90] B. Hanson & T. Applebaum, "Robust speaker independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech", *Proc. ICASSP, pp. 857-860, Albuquerque, 1990.*

[HER90] H. Hermansky, ``Perceptual linear predictive analysis of speech," *Journal of The Acoustic Soc. Am., vol. 87, 1990*.

[HER91] H. Hermansky, A. Bayya, N. Morgan, and P. Khon, "Compensation for the effect of the communication channel in perceptual linear predictive (plp) analysis of speech". *Proceeding of the European Conference On Speech Communication and Technology., pp. 1367--1370, Genova, Italy, 1991.*

[HAT91] J.-P. Haton J.-M. G. J. Pierrel, G. Perennou, J. Caelen et J.-L.. Gauvain, " Reconnaissance automatique de la parole ", *Dunod, Paris, 1991*.

[HAE92] R. Haeb-Umbach, D. Geller, and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition". *Proceedings of the International Conference on Acoustics, Speech and Signal Processing., vol. 1, pp. 13--16, 1992*.

[HUA93] X. Huang, F. Alleva, M.-Y. Hwang, R. Rosenfeld, "An Overview of the SPHINX-II Speech Recognition System". *93 ARPA HLT Workshop, pp. 81- 87, 1993*.

[HER94] H. Hermansky, N. Morgan. "RASTA Processing of speech". *IEEE Trans.On Speech and Audio Processing, vol.2, no.4, pp. 578-589, 1994*.

[HUA98] Z. Huang. "Extensions to the k-means algorithm for clustering large data sets with categorical values". *In Data Mining and Knowledge Discovery, pp. 283–304, 1998.*

[HÖP99] F. Hôppner, Klawonn, *al*. "Fuzzy Cluster Analysis". *Methods for Classification, Data*

Analysis and Image Recognition, John Wiley & Sons, Ltd, 1999.

[HAGa00] A. Hagen, A. Morris . "Comparison of HMM experts with MLP experts in the full combination multi-band approach to robust asr". *Int.Conf.on Spoken Language Processing, Beijing 2000*.

[HAGb00] A. Hagen, A. Morris. "From multi-band full combination to multi-stream full combination processing in robust asr". *ISCA Tutorial Research Workshop ASR2000, Paris, France, 2000*.

[HECa00] M. Heckmann, F.Berthommier, V.Charbonneau, K.Kroschel, "A Multi-Stage methodology to setup an ANN/HMM Audio-Visual speech recognition system". *In Proc. ICSLP2000, Beijing, China, 2000*.

[HECb00] M. Heckmann, F.Berthommier, K.Kroschel, "A hybrid ANN/HMM Audi-Visual speech recognition". *In Proc. ICSLP, Beijing, China, 2000*.

[HER00] H.Hermanssky, D.P.W.Ellis, S.Sharma, "Tandem connectionist feature extraction for conventionnal HMM systems". *International Computer Science Institute, Berkeley, California, USA, 2000*.

[JEL69] F. Jelinek, "A Fast Sequential Decoding Algorithm Using a Stack". *IBM J. Research and Development, Vol. 13, pp. 675-685, Novembre 1969*.

[JEL76] F. Jelinek, "Continuous speech recognition by statistical methods". *Proceedings of the IEEE, vol. 64, no. 4, pp. 532--536, 1976*.

[JUA85] B. H. Juang and L. R. Rabiner. "Mixture autoregressive hidden markov models for speech signals". *Proceedings of the International Conference on Acoustics, Speech and Signal Processing., vol. 33, no. 6, pp. 1404-- 1413, 1985*.

[JUN90] J.C.Junqua. "Utilisation d'un modèle d'audition et de connaissances phonétiques en reconnaissance automatique de la parole". *Traitement du signal, Vol 7, n°4, pp. 275-284, 1990*.

[JUN93] J.-C. Junqua, H. Wakita & H. Hermansky. "Evaluation and optimization of perceptually-based ASR front-end". *IEEE Trans. on Speech and Audio Processing, vol. 1, no. 1, pp. 39-48, 1993*.

[JOD94] J-F. Jodouin. "Les réseaux de neurones : Principes & définitions". *Edition Hermès, Paris, France, 1994*.

[KLA77] D.Klatt. "Review of the ARPA speech understanding project ". *J.Acoust.Soc.Am, Vol JASA-62, n°6, pp. 1345-1366, 1977*.

[KUM96] N. Kumar and A. G. Andreou. "On generalizations of linear discriminant analysis". *Tech. Rep. JHU/ECE-96-07, Electrical and Computer Engineering, John Hopkings University, 1996*.

[LAN77] A.Landercy, R.Renard. " Eléments de phonétique". *Bruxelles, Didier, Centre International de Phonétique Appliquée de Mons, 1977*.

[LAG85] E. Lagger, A.Waibel. "A Coarse Phonetic Knowledge Source for Template Independent Large Vocabulary Word Recognition". *ICASSP 85, pp. 2.7.1- 2.7.4, 1985*.

[LAF87] P. Laface, G. Micca, R. Pieraccini. "Experimental Results on a Large Lexicon Access Task". *ICASSP 87, pp. 20.4.1-20.4.4, 1987*.

[LAF88] P. Laface. "Recognition of Words in Very Large Vocabulary". *Recent Advances in Speech Understanding and Dialog Systems, NATO ASI Series F Volume 46, pp. 235-254, Edited by H. Niemann, M. Lany and G. Sagerer, 1988*.

[LEE88] K.-F. Lee "Large-vocabulary speaker-independant continuous speech recognition: The SPHINX system". *Ph.D. Thesis, Carnegie Mellon University, 1988*.

[LEEa89] K.F. Lee, H.-W. Hon, M.-Y. Hwang, S. Mahajan, R. Reddy. "The SPHINX Speech Recognition System". *ICASSP 89, pp. S9.3, 1989*.

[LEEb89] K.-F. Lee. "Automatic Speech Recognition: Development of the SPHINX System". *Kluwer Academic Publisher, forward by D. Raj Reddy, 1989.*

[LAN90] K-J. Lang, A-H. Waibel. "A time-delay neural network architecture for isolated word recognition". *Neural Networks, vol.3, pp. 23-43, 1990*.

[LEE90] K.-F. Lee. "Context-dependant phonetic hidden Markov models for speaker-independent continuous speech recognition". *IEEE Trans. Acoust., Speech, Signal Processing, vol. no. pp. 1990*.

[LEE92] C.H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini & A.E. Rosenberg. "Improved acoustic modeling for large vocabulary continuous speech recognition". *Computer Speech and Language, vol. 6, no. 3, pp. 197-213, 1992*.

[LEN92] M. Lennig, D. Sharp, P. Kenny, V. Gupta, K. Precoda. "Flexible Vocabulary Recognition of Speech". *ICSLP 92, pp. 93-96, 1992*.

[LJO94] A. Ljolje. "The importance of cepstral parameter correlations in speech recognition". *Computer Speech and Language, vol.I, pp.? 1994.*

[LAF95] P. Laface, C. Vair, L. Fissore. "A Fast Segmental Viterbi Algorithm for Large Vocabulary Recognition". *ICASSP 95, pp. 560-563, 1995*.

[LAZ00] L.Lazli. « Système Expert Connexionniste : Application pour la Reconnaissance des Mots Arabes Isolés ». *Mémoire d'Ingénieur d'état en Informatique*, département d'Informatique, Université d'Annaba, Algérie, Juin 2000.

[LAZa02] L. Lazli, M. Sellami. "Proposition d'une Architecture d'un Système Hybride HMM-PMC pour la Reconnaissance de la Parole Arabe". *Proceedings of t he Seventh Magrebian Conference on Computer Sciences, 7th MCSEAI'02, Mai, vol I, pp. 101-109, Annaba, Algérie, 2002*.

[LAZb02] L. Lazli, M. Sellami. "Modèle Neuro-Markovien pour la Reconnaissance de la Parole Arabe Indépendante de Locuteur". *SNAS'02, Séminaire National sur l'Automatique et traitement de Signaux,* pp. 13 (Résumé), Oct. 27-28, Annaba, Algérie, 2002.

[LAZc02] L. Lazli, M. Sellami. "Reconnaissance de la parole arabe par système hybride HMM/MLP". *CGE'02, Conférence sur le Génie Electrique*, pp. 93 (Résumé), Décembre.17-18, EMP. Ecole Militaire Polytechnique, Alger, Algérie, 2002.

[LAZd02] L. Lazli, H. Bahi, M. Sellami. "Modèle Neuro-symbolique pour la Reconnaissance de la Parole Arabe". *CGE'02, Conférence sur le Génie Electrique*, Décembre.17-18, EMP. Ecole Militaire Polytechnique, Alger, Algérie, 2002.

[LAZa03] L. Lazli, M. Sellami. "Connectionist Probability Estimators in HMM Arabic Speech Recognition using Fuzzy Logic". *MLDM 2003: the 3rd international conference on Machine Learning & Data Mining in pattern recognition, LNAI 2734, Springer-verlag,*

pp.379-388, 5-7 Juillet, Leipzig, Allemagne, 2003.

[LAZb03] L. Lazli, M. Sellami. "Hybrid HMM-MLP system based on fuzzy logic for Arabic speech recognition". *PRIS 2003: the third international workshop on Pattern Recognition in Information Systems with ICEIS 2003: the 5th International Conference on Entreprise Information Systems, Springer-verlag,* pp. 150-155, 22-23 Avril, Angers, France, 2003.

[LAZc03] L. Lazli, M. Sellami. "Speech Segmentation using a New Unsupervised Approach". *ISPS 2003: the International Symposium on Programming & Systems,* pp. 295-305, Mai 5-7, Alger, Algérie, 2003.

[LAZd03] L. Lazli, M. Sellami. "Speech clustering using a new algorithm". *AICCSA 2003 (ACS/IEEE): international conference on Applications & Computer Systems,* pp. 131 (Abstract), Juillet 14-18, Tunis, Tunisie, 2003.

[LAZe03] L. Lazli, M. Sellami. "Speaker independent isolated speech recognition for language using hybrid HMM-MLP-FCM system". *AICCSA 2003 (ACS/IEEE): international conference on Applications & Computer Systems,* pp. 108 (Abstract), Juillet 14-18, Tunis, Tunisie, 2003.

[LAZf03] L. Lazli, M. Sellami. « Modèle hybride HMM-MLP basé flou : Appliqué à la reconnaissance de la parole ». *SETIT2003/IEEE : conférence internationale sur les Sciences Electroniques, Technologies de l'Information & des Télécommunications,* pp. 104 (Résumé), 17-21 mars , Sousse, Tunisie, 2003.

[LAZg03] L. Lazli, M. Sellami. "A new method for unsupervised classification: Application for the speech clustering". *SETIT2003, Conférence internationale : Sciences Electroniques, Technologies de l'Information et des Télécommunications,* pp. 97 17-21 Mars, Sousse, Tunisie, 2003.

[LAZh03] L. Lazli, " Système de reconnaissance neuro-markovien pour la parole arabe isolée basé sur une segmentation floue". Mémoire de MAGISTER en Informatique, département d'Informatique, Université d'Annaba, Algérie, Juin 2003.

[LAZa04] L. Lazli, M-T. Laskri. "Nouvelle méthode d'entraînement des systèmes hybrides HMM/ANN appliquée pour la reconnaissance automatique de la parole". *CARI 2004: the 7ème Colloque Africain sur la Recherche en Informatique,* pp. 331-338, Novembre 22-25, Hammamet, Tunisie, 2004.

[LAZb04] L. Lazli, M.T. Laskri. "Application de la méthode de fusion de données pour la reconnaissance automatique de la parole indépendante de locuteur". *MCSEAI 2004: the 8th Magrebian Conference on Computer Sciences,* pp. 523-533, 9-12 Mai, Sousse, Tunisie, 2004.

[LAZ05] L. Lazli, M-T.Laskri . "Nouvelle méthode de fusion de donnée pour l'apprentissage des systèmes hybrides MMC/RNA". Revue ARIMA: Revue africaine de la recherche en Informatique et Mathématiques appliquées, Volume 3 - numéro spécial CARI, novembre 2005, pp. 125-170.
http://wwwdirection.inria.fr/international/arima/docs/02articles.html

[MYE81] C.S. Myers & L.R. Rabiner. "Connected digit recognition using a level building DTW algorithm", *IEEE Trans. Acoust., Speech, Signal Processing, vol. 1981*.

[MAR85] S.M. Marcus. "Associative Models and the Time Course of Speech". *Bibliotheca Phonetica, No.12, pp. 36-52, Karger, Basel 1985*.

[MOR00] A.Morris, A.Hagen, H.Glotin, H.Bourlard. " Multi – Stream adaptive evidence combination for noise robust ASR". *IDIAP Research* Report 99-26, Accepted for publication in Speech Communication, January 2000.

[MORa01] A.Morris. "Some applications of a priori knowledge in Multi-Stream HMM and HMM/ANN based ASR". IDIAP Research Report, 2001.

[MORb01] A.Morris, A.Hagen, H.Bourlard. "MAP combination of Multi-Stream HMM or HMM/ANN experts". *IDIAP Research Report 01-14, EuroSpeech2001, Special Event "Noise Robust Recignition",* Aalborg, Denmark, June 2001.

[NIL82] N. Nilsson. "Principles of Artificial Intelligence". *Tioga Publishing Company*, 1982.

[PIE69] J.Pierce. "Whither speech recognition". *J.Acoust.Soc.Am*, Vol JASA-46, n°6, pp. 1049-1051, 1969.

[PHA99] D-L. Pham, J-L. Prince. "An Adaptive Fuzzy C-means algorithm for Image Segmentation in the presence of Intensity Inhomogeneities". *Pattern Recognition Letters.* 20(1), pp. 57-68, 1999.

[RAB79] L.R. Rabiner, S.E. Levinson, A.E. Rosenberg & J.G. Wilpon. "Speaker-independent recognition of isolated words using clustering techniques". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 1979.

[RAB86] L. R. Rabiner and B. H. Juang. "An introduction to hidden markov models". *ASSP Magazine*, pp. 4-16, January 1986.

[RAB87] L. R. Rabiner, J. G. Wilpon, and B. H. Juang. "A performance evaluation of a connected digit recognizer". *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing., (Dallas, TX)*, pp. 101-104, Apr. 1987.

[RUS87] M. J. Russel and A. E. Cook. "Experimental evaluation of duration modelling techniques for automatic speech recognition". *Proceedings International Conference on Acoustics Speech and Signal Processing*, pp. 2376--2379, April 1987.

[RAB89] L. R. Rabiner and B. Juang. "A tutorial on hidden markov models and selected applications in speech recognition". *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-285, 1989.

[ROB90] T. Robinson, J. Holdsworth, R. Patterson & F. Fallside. "A comparison of preprocessors for the Cambridge recurrent error propagation network speech recognition system". *Proc. ICSLP*, pp. 1033- 1036, 1990.

[RUS90] M. J. Russel, K. M. Ponting, S. M. Peeling, S. R. Browning, J. S. Briddle, R. K. Moore, I. Galiano, and P. Howell. "The arm continuous speech recognition system". *Proceedings International Conference on Acoustics Speech and Signal Processing*, pp. 69--72, April 1990.

[RAB93] L. R. Rabiner and B.-H. Juang. "Fundamentals of Speech Recognition". *PTR Prentice Hall*, 1993.

[ROB94] A-J. Robinson. "An application of recurrent nets to phone probability estimation". *Proceedings of the IEEE Transactions on Neural Network*, vol.5, pp. 298-305, 1994.

[RAL95] H. Ralambondrainy. "A conceptual version of the k-means algorithm". *Pattern Recognition Letters, 16 :* pp. 1147–1157, 1995.

[RII97] S-K. Rii, A. Krogh. "Hidden Neural Networks: A framework for HMM-NN hybrids". *IEEE 1997, ICASSP-97,* Apr 21-24, Munich, Germany, 1997.

[ROB99] P-J. Robert-Ribes, J-L. Schwartz, A. Guerin-Dugue. "Comparing models for audiovisual fusion in noisy-vowel recognition task". *IEEE Trans. Speech Audio Processing 7*, pp. 629-642, 1999.

[ROB05] M. Robenson, M.R. Azimi-Sadjadi, Senior Member, IEEE, and J. Salazar."Multi-Aspect Target Discrimintion Using Hidden Markov Models and Neural Networks", *IEEE transaction on neural networks*, vol. 16, No. 2, pp. 447-459, March 2005.

[SCH77] M. R. Schroeder. "Recognition of complex acoustic signal". *Life Science Research* Report edited by T. H. Bullock, 1977.

[SAK78] H. Sakoe & S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, no. pp. 1978.

[SAK79] H. Sakoe. "Two-level DP-matching - A dynamic programming-based pattern matching algorithm for connected speech recognition". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. pp. 595, 1979.

[SOO88] F.K. Soong & A.E. Rosenberg. "On the use of instantaneous and transitional spectral information in speaker recognition". *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 6, pp. 1988.

[SIO95] O. Siohan. "On the robustness of linear discriminant analysis as a pre-processing step for noisy speech recognition". *Proceedings International Conference on Acoustics Speech and Signal Processing*, pp. 125--128, 1995.

[TAB97] Y.Tabet. " Analyse et synthèse du signal de la parole arabe ". *Mémoire de Magister, Département d'Electronique*, Université de Annaba, Algérie, Juillet 1997.

[TIM01] H. Timm. "Fuzzy Cluster Analysis of Classified Data". *IFSA/Nafips 2001, Vancouver, 2001*.

[VIT67] A.J. Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". *IEEE Trans. Inform. Theory*, vol. 13, no. 2, pp. 260-269, 1967.

[VIN68] T.K. Vintsjuk. "Recognition of words of oral speech by dynamic programming". *Kibernetika*, vol. 81, no. 8, 1968.

[VEL70] V.M. Velichko & N.G. Zagoruyko. "Automatic recognition of 200 words" *Int. J. Man-Machine Studies*, vol. 2, pp. 223-234, 1970.

[VER89] G.J. Vernooij, G. Bloothooft, Y Van Holsteijn. "A Simulation Study on the Usefulness of Broad Phonetic Classification in Automatic Speech Recognition". *ICASSP 89*, pp. 85-88, 1989.

[WIL93] J.G. Wilpon, C.-H. Lee & L.R. Rabiner "Connected digit recognition based on improved acoustic resolution". *Computer Speech and Language*, vol. pp. 1993.

[WOO95] P.C. Woodland, C.J. Leggetter, J.J. Odell, V. Valtchev & S.J. Young. "The 1994 HTK large vocabulary speech recognition system". *Proc. ICASSP*, pp. 73-76, Detroit, 1995.

[YOU92] S.J. Young. "The general use of tying in phoneme-based HMM speech recognizers".

Proc. ICASSP, pp. San Francisco, 1992.

[YAN97] Y. Yan, M. Fanty, R. Cole. "Speech Recognition Using Neural Networks with Forward-Backward Probability Generated Targets". *ICASSP*, vol.4, pp. 32-41, 1997.

[ZUA90] V. Zue, S. Seneff, and J. Glass "Speech database development : Timit and beyond". *Speech Communication*, pp. 351--356, 1990.

[ZHA93] Y. Zhao. "A speaker-independent continuous speech recognition system using continuous mixture gaussian density HMM of phoneme-sized units". *IEEE Trans. on Speech and Audio Processing*, vol. no. pp. 1993.

[ZHA94] Y. Zhao. "An acoustic-phonetic-based speaker adaptation technique for improving speaker independent continuous speech recognition". *IEEE Trans. on Speech and Audio Processing*, vol. no. pp. 394, 1994.