



École Doctorale  
d'Informatique,  
Télécommunications  
et Électronique de Paris

Thèse  
présentée pour obtenir le grade de docteur  
de l'École Nationale Supérieure des Télécommunications

Spécialité : **Signal et Images**

**Belgacem BEN MOSBAH**

Utilisation de la mémoire de parole pour la  
reconnaissance : Application pour des  
personnes handicapées

Soutenue le 19 Janvier 2005 devant le jury composé de

BENNANI Younes	Président
ZARADER Jean-luc	Rapporteurs
ENNAJI Abdel	
ABED MERIEM Karim	Examineurs
CHARBIT Maurice	
RICHARD Gael	

Ecole nationale supérieure des télécommunications

**A mes parents , mes beaux parents et ma femme**

**A mes enfants Samar et Iyad**

A mes frères et soeurs, a mes amis Hatem, Mourad, Mohamed, Hédi ..

# Résumé

Durant ces dernières années, et grâce à l'évolution de l'informatique, nous avons assisté à une évolution assez importante des systèmes de reconnaissance automatique de la parole. Les systèmes développés dépendent des applications, il y'a des systèmes de reconnaissance de mots isolés et des systèmes de reconnaissance de la parole continue. Les systèmes de reconnaissance de mots isolé sont comme principale application la commande vocale tandis que les systèmes de reconnaissance de la parole continue ont comme application principale la dictée vocale.

Pour les personnes handicapées l'absence des bases de données et la diversité des handicaps articulatoires sont des obstacles majeurs pour la construction de systèmes de reconnaissance de la parole fiables, ce qui explique la pauvreté du marché en systèmes de reconnaissance de la parole pour des personnes handicapées.

Le travail développé durant cette thèse consiste à adapter certains des systèmes de reconnaissance de la parole existants aux personnes qui ont des handicaps articulatoires.

Pour les systèmes de reconnaissance de mots isolés, nous avons utilisé une approche d'apprenissage dynamique qui permet aux systèmes de s'adapter aux utilisateurs au fur et à mesure de son utilisation. Cette approche permet à un utilisateur handicapé d'utiliser le système sans passer par une longue phase d'apprentissage qui est généralement lourde et pénible pour ces personnes. Elle permet aussi aux systèmes de reconnaissance d'utiliser une base d'apprentissage dans les mêmes conditions que les tests.

Pour les systèmes de reconnaissance de parole continue nous avons utilisé deux types d'approches :

- 1 - Une adaptation dynamique des modèles phonétiques des systèmes de reconnaissance de parole continue pour les personnes handicapées. Cette approche permet d'adapter le système de reconnaissance à l'utilisateur et peut être appliquée aux personnes normaux parlants.

- 2 - Utilisation d'une segmentation indépendante de la langue (ALISP) pour la reconnaissance. Cette approche consiste à utiliser la correspondance entre la segmentation ALISP et la phonétique et les modèles des segments ALISP pour construire un système de reconnaissance de la parole continue.

# Abstract

In the last few years, thanks to fast evolution of data processing, we readed a rather important evolution of the speech recognition system.

The developed systems are application depended. Two classes of systems exist those for the recognition of isolated words and those for continous speech recognition. The isolated words recognition systems have as application the vocal order while the continous speech recognition systems have, as principal application, the vocal dictation.

For disabled people, the absence of appropriate data bases and the diversity of the articulator handicaps are major obstacles for the construction of reliable speech recognition systems. This explains the poverty of the market in speech recognition systems for disabled people.

The work developped during this thesis consists in adapting the existing speech recognition systems to the people who have articulator handicaps.

For isolated words recognition systems, we used a dynamic approach of training which makes it possible for the system to progressively adapt to the users during his functioning. In particular this approach allows disabled users to exploit the system without passing through a long phase of training which is generally too heavy and painful for them. Also by this approach, one trains the recognition system under the same test conditions.

For the continous recognition systems, we proposed two approaches :

1 - A dynamic adaptation of the phonetic models of the continous recognition systems for the disabled people. This approach makes it possible to adapt the recognition system to the user ande hence can be used for normal speaking people.

2 - An approach that uses an independent language segmentation (ALISP) for the recognition. This approach consists of using the correspondance between ALISP segmentation and the phonetics and the models of segments ALISP to build the recognition system.

# Table des matières

<b>1</b>	<b>Introduction générale</b>	<b>15</b>
<b>2</b>	<b>État de l'art</b>	<b>17</b>
2.1	Introduction . . . . .	17
2.2	Historique . . . . .	17
2.3	État de l'art de la reconnaissance de la parole . . . . .	18
2.3.1	Les commandes vocales . . . . .	19
2.3.2	Les systèmes de compréhension . . . . .	19
2.3.3	Les systèmes de dictée vocale . . . . .	19
2.4	Conclusion . . . . .	21
<b>3</b>	<b>Généralités sur les systèmes de reconnaissance de la parole</b>	<b>23</b>
3.1	Introduction . . . . .	23
3.2	Les systèmes de reconnaissance de la parole . . . . .	24
3.2.1	Les différentes méthodes de reconnaissance de la parole . . . . .	24
3.2.2	Le décodage acoustico-phonétique . . . . .	24
3.2.3	Outils pour la reconnaissance de la parole . . . . .	26
3.2.4	Méthodes de reconnaissance de la parole . . . . .	27
3.3	Conclusion . . . . .	36
<b>4</b>	<b>Étude des handicaps articulatoires</b>	<b>37</b>
4.1	Introduction . . . . .	37
4.2	Production de la parole . . . . .	37
4.2.1	La production du son . . . . .	38
4.2.2	Consonnes et voyelles . . . . .	38
4.2.3	Point d'articulation et mode d'articulation . . . . .	40
4.2.4	Sourdes ou sonores . . . . .	41
4.2.5	Orale et nasale . . . . .	41
4.3	Étude de la voix des parkinsoniens . . . . .	41
4.3.1	Analyse fréquentielle et temporelle . . . . .	42
4.3.2	Commentaires . . . . .	48
4.4	Conclusion . . . . .	48

<b>5</b>	<b>Apprentissage dynamique</b>	<b>49</b>
5.1	Introduction . . . . .	49
5.2	Description de la méthode . . . . .	50
5.3	Paramétrisation du signal parole . . . . .	50
5.3.1	Fenêtrage et normalisation du signal . . . . .	51
5.3.2	Coefficients MFCC . . . . .	51
5.4	L'algorithme de comparaison dynamique . . . . .	52
5.5	Résultats expérimentaux et discussions . . . . .	53
5.5.1	Procédure expérimentale . . . . .	53
5.5.2	Résultats et discussion . . . . .	54
5.5.3	Taux de confiance . . . . .	54
5.6	Conclusion . . . . .	56
<b>6</b>	<b>Utilisation d'ALISP</b>	<b>57</b>
6.1	Introduction . . . . .	57
6.2	Principe de la segmentation ALISP . . . . .	57
6.2.1	Segmentation initiale . . . . .	58
6.2.2	Segmentation statistique . . . . .	59
6.3	Transcription phonétique d'une base de données . . . . .	60
6.3.1	Voix d'essai . . . . .	60
6.3.2	Mise en forme du signal acoustique . . . . .	60
6.4	Étude des correspondances entre une segmentation phonétique et une segmentation ALISP . . . . .	61
6.4.1	Caractéristiques de la base de données . . . . .	61
6.4.2	Correspondances ALISP-phonétique . . . . .	61
6.4.3	Étude de la correspondance ALISP-polyson . . . . .	62
6.5	Résultats expérimentaux et commentaires . . . . .	63
6.5.1	Correspondance utilisant les classes ALISP . . . . .	63
6.5.2	Correspondance utilisant les sous-classes . . . . .	65
6.6	Utilisation de la segmentation ALISP pour la reconnaissance de la parole . . . . .	66
6.6.1	Utilisation de la segmentation ALISP pour la reconnaissance des mots isolés . . . . .	66
6.6.2	Utilisation de la correspondance ALISP-polyson pour la re- connaissance de la parole continue . . . . .	68
6.7	Conclusion . . . . .	71
<b>7</b>	<b>Adaptation dynamique</b>	<b>73</b>
7.1	Modèles de Markov cachés . . . . .	73
7.1.1	Modélisation du signal parole . . . . .	73
7.1.2	Reconnaissance d'un modèle . . . . .	75
7.1.3	Recherche des états cachés . . . . .	77
7.1.4	Apprentissage d'un modèle . . . . .	79
7.2	Adaptation des HMM . . . . .	81



7.2.1	Adaptation de la moyenne (MLLR : Processus d'adaptation de régression linéaire de maximum de vraisemblance) . . . . .	81
7.2.2	Adaptation de la covariance . . . . .	84
7.3	Reconnaissance de la parole continue . . . . .	85
7.3.1	Modèles acoustiques . . . . .	85
7.3.2	Modèles de langage . . . . .	86
7.3.3	Apprentissage en parole continue . . . . .	87
7.3.4	Reconnaissance de la parole continue . . . . .	88
7.4	Système de commande vocale de l'environnement domestique pour les handicapés . . . . .	89
7.4.1	Description des bases de données de tests . . . . .	90
7.4.2	Description du système de reconnaissance de la parole . . . . .	90
7.4.3	Utilisation du modèle du monde . . . . .	91
7.4.4	Adaptation des modèles (adaptation des moyennes) du monde aux locuteurs handicapés . . . . .	92
7.4.5	Adaptation des modèles (adaptation des moyennes et des covariances) du monde aux locuteurs handicapés . . . . .	93
7.4.6	Utilisation des modèles adaptés aux locuteurs handicapés pour la reconnaissance de la parole des locuteurs normaux . . . . .	94
7.5	Système de reconnaissance de parole continue pour des parkinsoniens . . . . .	95
7.5.1	Système de référence . . . . .	95
7.5.2	Procédure expérimentale . . . . .	96
7.6	Conclusion . . . . .	105
<b>8</b>	<b>Conclusions et perspectives</b>	<b>107</b>
	<b>Annexes</b>	<b>108</b>
<b>A</b>	<b>correspondance sous-classes-ALISP</b>	<b>111</b>



# Liste des tableaux

2.1	Quelques produits du marché . . . . .	21
4.1	Quelques statistiques sur la fréquence fondamentale pour la personne 1	43
4.2	Quelques statistiques sur la fréquence fondamentale pour la personne 2	43
4.3	Mesure de taux de silences des fichiers enregistrés pour le niveau 0 . .	44
4.4	Mesure de vitesse d'élocution des fichiers enregistrés pour le niveau 0	44
4.5	Quelques statistiques sur la fréquence fondamentale pour la personne 1	45
4.6	Quelques statistiques sur la fréquence fondamentale pour la personne 2	45
4.7	Taux de silence pour des fichiers enregistrés pour le niveau 1 . . . . .	45
4.8	Vitesse d'élocution pour des fichiers enregistrés pour le niveau 1 . . .	46
4.9	Quelques statistiques sur la fréquence fondamentale pour une per- sonne de niveau 2 (exemple 1) . . . . .	46
4.10	Taux de silence des fichiers enregistrés pour le niveau 2 . . . . .	46
4.11	Vitesse d'élocution des fichiers enregistrés pour le niveau 2 . . . . .	47
4.12	Quelques statistiques sur la fréquence fondamentale pour une per- sonne de niveau 3 . . . . .	47
4.13	Taux de silence des fichiers enregistrés pour le niveau 3 . . . . .	47
4.14	Vitesse moyenne d'élocution des fichiers enregistrés pour le niveau 3 .	48
4.15	Tableau récapitulatif . . . . .	48
5.1	Évolution de taux de reconnaissance au cours de l'apprentissage dy- namique . . . . .	54
5.2	Évolution de taux de confiance au cours de l'apprentissage dynamique pour la phrase D . . . . .	55
5.3	Évolution de taux de confiance au cours de l'apprentissage dynamique pour la phrase B . . . . .	56
6.1	Nombre de segments ALISP dans la base de données ELAN . . . . .	64
6.2	Quelques correspondances entre les phonèmes et les segments ALISP	64
6.3	Taux de reconnaissance pour le locuteur 1 . . . . .	67
6.4	Taux de reconnaissance pour le locuteur 2 . . . . .	68
7.1	T.rec avec les modèles du monde pour le locuteur 1 . . . . .	91
7.2	T.rec avec les modèles du monde pour le locuteur 2 . . . . .	91
7.3	T.rec avec les modèles adaptés (adaptation des moyennes) pour le locuteur 1 . . . . .	92
7.4	T.rec avec les modèles adaptés pour le locuteur 2 . . . . .	92

7.5	T.rec avec les modèles partiellement adaptés . . . . .	93
7.6	T.rec avec les modèles adaptés (adaptation des moyennes et des co- variances) pour le locuteur 1 . . . . .	93
7.7	T.rec avec les modèles adaptés pour le locuteur 2 . . . . .	94
7.8	T.rec avec les modèles partiellement adaptés . . . . .	94
7.9	T.rec avec les modèles adaptés au locuteurs handicapés pour le locu- teur normale 1 . . . . .	95
7.10	T.rec avec les modèles adaptés au locuteurs handicapés pour le locu- teur normale 2 . . . . .	95
7.11	Taux de reconnaissance avec les modèles du BREF pour la base de données BREF . . . . .	96
7.12	Nombre de locuteurs par niveau . . . . .	97
7.13	Taux de reconnaissance pour les locuteurs handicapés avec les mo- dèles de BREF . . . . .	98
7.14	Évolution de taux de reconnaissance pour le degré 0 . . . . .	99
7.15	Évolution de taux de reconnaissance pour le degré 1 . . . . .	99
7.16	Évolution de taux de reconnaissance pour le degré 2 . . . . .	100
7.17	Évolution de taux de reconnaissance pour le degré 3 . . . . .	100
7.18	Évolution de taux de reconnaissance pour le degré 0 . . . . .	101
7.19	Évolution de taux de reconnaissance pour le degré 1 . . . . .	102
7.20	Évolution de taux de reconnaissance pour le degré 2 . . . . .	102
7.21	Évolution de taux de reconnaissance pour le degré 3 . . . . .	102
7.22	Évolution de taux de reconnaissance pour le degré 0 . . . . .	103
7.23	Évolution de taux de reconnaissance pour le degré 1 . . . . .	103
7.24	Évolution de taux de reconnaissance pour le degré 2 . . . . .	104
7.25	Évolution de taux de reconnaissance pour le degré 3 . . . . .	104
A.1	Correspondances de quelques sous-classes de HB avec des polysons . .	111
A.2	Correspondances de quelques sous-classes de HC avec des polysons . .	112
A.3	Correspondances de quelques sous-classes de HD avec des polysons . .	112
A.4	Correspondances de quelques sous-classes de HG avec des polysons . .	113
A.5	Correspondances de quelques sous-classes de HH avec des polysons . .	113
A.6	Correspondances de quelques sous-classes de HI avec des polysons . .	114
A.7	Correspondances de quelques sous-classes de HJ avec des polysons . .	114
A.8	Correspondances de quelques sous-classes de HK avec des polysons . .	114
A.9	Correspondances de quelques sous-classes de HL avec des polysons . .	115
A.10	Correspondances de quelques sous-classes de HM avec des polysons . .	115

# Table des figures

3.1	<i>Chaîne de Markov à trois états gauche-droite . . . . .</i>	29
4.1	<i>appareil respiratoire . . . . .</i>	38
4.2	<i>le principe de production des consonnes et des voyelles . . . . .</i>	40
4.3	<i>La voix nasale. . . . .</i>	41
5.1	<i>Mise à jour dynamique du dictionnaire des références . . . . .</i>	50
5.2	<i>Calcul des coefficients MFCC . . . . .</i>	52
5.3	<i>contrainte de cheminement utilisée . . . . .</i>	53
6.1	<i>Principe de la segmentation ALISP . . . . .</i>	58
6.2	<i>La correspondance ALISP-phonèmes . . . . .</i>	62
6.3	<i>La correspondance ALISP-polyson . . . . .</i>	62
6.4	<i>Définition des sous-classes . . . . .</i>	66
6.5	<i>Détermination de la correspondance ALISP-polyson pour la reconnaissance . . . . .</i>	70
6.6	<i>Détermination hypothèses de mots du dictionnaire . . . . .</i>	70
6.7	<i>Détermination du résultat final de la reconnaissance . . . . .</i>	71



# Chapitre 1

## Introduction générale

Avec l'évolution de l'informatique, les systèmes de reconnaissance de la parole ont constamment évolué, en partant de système de reconnaissance monolocuteur de mots isolés de quelques mots jusqu'aux systèmes de parole continue indépendants du locuteur avec des dictionnaires qui contiennent des dizaines de milliers de mots [HL00], [Bar96].

Les systèmes actuels de parole continue indépendants du locuteur utilisent une phase d'adaptation et atteignent plus de 95% de taux de reconnaissance tels que Viavoice et Dragon [Lok99].

Mais malheureusement ces systèmes grand public ne tiennent pas compte des personnes qui ont des handicaps articulatoires. Nous avons fait une expérience à l'hôpital de Garche, une personne normale a mis 30 minutes pour adapter sa voix au système Dragon tandis qu'une personne handicapée même à l'issue d'une heure d'adaptation n'a pas pu utiliser le système.

Reste à ajouter que pour des personnes handicapées, l'enregistrement d'une base de données est pénible. A ces problèmes s'ajoute la variation des types du handicaps [NT00], [BM00]. Il s'avère alors que la construction d'un système de reconnaissance de parole pour les personnes handicapées, en utilisant les méthodes classiques, est très difficile voire impossible dans ces conditions.

Notre travail de recherche consiste à chercher des solutions pour ces personnes handicapées en partant de systèmes existants ou en utilisant des approches qui s'affranchissent des difficultés articulatoires de ces personnes.

L'approche que nous proposons consiste à utiliser une mémoire de parole (parole de test) enregistrée au cours de l'utilisation du système pour la reconnaissance de la parole.

Le chapitre 1 est consacré à l'état de l'art de la reconnaissance de la parole, en citant quelques systèmes à commande vocale, de compréhension de parole et de dictée vocale. En fait les systèmes actuels de reconnaissance de mots isolés atteignent des performances pouvant atteindre jusqu'à 100% pour des vocabulaires  $\leq 1000$  mots, les systèmes de reconnaissances de parole continue peuvent atteindre 95% de taux de reconnaissance [Jou88].

Dans le chapitre 2 nous analysons les grands axes de reconnaissance de la parole qui sont les commandes vocales, systèmes de compréhension et la dictée vocale puis

nous expliquons les différentes méthodes de reconnaissance de la parole à savoir la méthode globale et la méthode analytique [HCD91]. Nous analysons également les différents outils utilisés en reconnaissance de la parole en expliquant les deux approches les plus utilisées que sont la comparaison dynamique et les modèles de Markov cachés.

Le chapitre 3 est consacré à l'étude des handicaps articulatoires, nous commençons par une étude de la production de la parole chez un être humain en expliquant les différents types de classification des sons de parole, tel que les consonnes et les voyelles, orales et nasales etc..., puis nous analysons le signal parole de quelques sujets.

Dans le chapitre 4, nous exposons la première approche proposée pour la reconnaissance de la parole de personnes handicapées. Cette approche consiste à utiliser un apprentissage dynamique au cours de l'utilisation du système de reconnaissance. Ce système se base sur la méthode de comparaison dynamique. L'utilisateur doit enregistrer la base de référence une seule fois qui dispense la personne handicapée d'enregistrer cette base plusieurs fois et l'apprentissage se fait au cours de l'utilisation du système de reconnaissance. Quand l'utilisateur est satisfait du résultat de la reconnaissance, le mot test est ajouté automatiquement au dictionnaire des références. Cet apprentissage augmente les performances des systèmes au cours de leur utilisation. Après un certain nombre d'utilisations, les performances du système se stabilisent ce qui nous permet d'arrêter ce processus.

Notre deuxième approche est détaillée au chapitre 5. Cette approche consiste à utiliser la segmentation ALISP pour la reconnaissance de la parole de personnes handicapée. Un premier système de reconnaissance de mots isolés a été développé. Ce système utilise une méthode statistique avec les segments ALISP comme unités de reconnaissance et un dictionnaire de mots transcrits en segments ALISP. Le deuxième système est un système de reconnaissance de parole continue. Ce système de reconnaissance utilise un dictionnaire phonétique et un modèle de langage et utilise les segments ALISP comme unités de reconnaissance. Il nécessite une phase intermédiaire pour passer de la segmentation ALISP à un découpage phonétique en utilisant la correspondance ALISP-phonétique.

Dans le chapitre 6 nous exposons la méthode d'adaptation dynamique des modèles statistiques proposés pour adapter les systèmes de reconnaissance de parole aux personnes handicapées. Nous avons développé deux systèmes de reconnaissance de la parole, le premier est un système de reconnaissance de parole à petit vocabulaire et le deuxième est un système de reconnaissance de parole continue à grand vocabulaire.

Un dernier chapitre récapitule les résultats et les perspectives de nos travaux dans cette thèse.



# Chapitre 2

## État de l'art

### 2.1 Introduction

La reconnaissance automatique de la parole pose de nombreux problèmes aussi bien au niveau théorique que pratique. Sa complexité fait que des sous-problèmes ont pu être résolus. Ces solutions partielles correspondent à des contraintes plus ou moins fortes, et les systèmes existants supposent une coopération des utilisateurs [Jou88].

Pour classer les systèmes de reconnaissance automatique, on a généralement recours aux critères suivants [Spa99] :

- le mode d'élocution (mots isolés, mots connectés et parole continue)
- la taille du vocabulaire (petit, moyen ou grand vocabulaire)
- la dépendance vis à vis du locuteur (mono-locuteur, multilocuteurs ou indépendant du locuteur)
- l'environnement (normal ou bruité).

Pour les systèmes de reconnaissance de la parole handicapée l'état de l'art est vierge, jusqu'à aujourd'hui il n'y a pas un système commercialisé de reconnaissance de la parole handicapés. Dans ce chapitre nous allons mentionner quelques systèmes de reconnaissance de parole pour des personnes normaux parlant en s'appuyant sur les méthodes utilisées pour pouvoir s'inspirer de ces systèmes pour résoudre le problème de reconnaissance de la parole handicapée.

### 2.2 Historique

La reconnaissance de la parole est une discipline récente. Vers 1950 apparut le premier système de reconnaissance de chiffres, appareil entièrement câblé et très imparfait. Vers 1960, l'introduction des méthodes numériques et l'utilisation des ordinateurs changent la dimension des recherches. Néanmoins, les résultats demeurent modestes car la difficulté du problème avait été largement sous-estimée, en particulier en ce qui concerne la parole continue. Vers 1970, la nécessité de faire appel à des contraintes linguistiques dans le décodage automatique de la parole avait été jusque là considérée comme un problème d'ingénierie. La fin de la décennie 70 voit

se terminer la première génération des systèmes commercialisés de reconnaissance de mots. Les générations suivantes, mettant à profit les possibilités sans cesse croissantes de la micro-informatique, posséderont des performances supérieures (systèmes multilocuteurs , parole continue) [Pye94], [Bak75].

On peut résumer en quelques dates les grandes étapes de la reconnaissance de la parole [HCD91] :

- 1952 : reconnaissance des 10 chiffres, pour un monolocuteur , par un dispositif électronique câblé,
- 1960 : utilisation des méthodes numériques
- 1965 : reconnaissance de phonèmes en parole continue
- 1968 : reconnaissance de mots isolés par des systèmes implantés sur gros ordinateurs (jusqu'à 500 mots)
- 1969 : utilisation d'informations linguistiques
- 1971 : lancement du projet ARPA aux USA (15 millions de dollars) pour tester la faisabilité de la compréhension automatique de la parole continue avec des contraintes raisonnables
- 1972 : premier appareil commercialisé de reconnaissance de mots
- 1976 : fin du projet ARPA ; les systèmes opérationnels sont HARPY, HEARSAY I et II et HWIM
- 1978 : commercialisation d'un système de reconnaissance à microprocesseurs sur une carte de circuits imprimés
- 1981 : utilisation de circuits intégrés VLSI (Very Large Scale Integration) spécifiques du traitement de la parole
- 1981 : système de reconnaissance de mots sur un circuit VLSI
- 1983 : première mondiale de commande vocale à bord d'un avion de chasse en France
- 1985 : commercialisation des premiers systèmes de reconnaissance de plusieurs milliers de mots
- 1986 : lancement du projet japonais ATR de téléphone avec traduction automatique en temps réel
- 1988 : apparition des premières machines à dicter par mots isolés
- 1989 : recrudescence des modèles connexionnistes neuromimétiques
- 1990 : premières véritables applications de dialogue oral homme-machine
- 1994 : IBM lance son premier système de reconnaissance vocale sur PC
- 1997 : lancement de la dictée vocale en continu par IBM

## 2.3 État de l'art de la reconnaissance de la parole

Dans le domaine de la reconnaissance de la parole, on distingue trois grands types d'applications [Bar96] :

- les systèmes de commande vocale
- les systèmes de compréhension
- les machines à dicter

### 2.3.1 Les commandes vocales

Il existe aujourd'hui un nombre important de produits fiables sur le marché qui permettent de contrôler l'environnement au moyen d'une entrée vocale [HL00]. Les applications multiples et variées vont du jouet gadget à l'outil de travail sophistiqué. voilà quelques exemples :

- commandes de la voiture
- jeux vidéo
- noms de gares SNCF
- aides aux handicapées : ces applications sont soit du type aide aux personnes ayant un organe défaillant hors la voix (aveugle, problèmes moteurs etc..), soit du type outil de rééducation pour les malentendants.
- reconnaissance des chiffres : comme le projet cabine vocale (pour composer un numéro de téléphone il suffit de prononcer la suite des chiffres).

Pour ces applications la taille du vocabulaire est limitée (elle ne dépasse pas quelques centaines de mots).

Bien qu'une souplesse soit donnée à l'utilisateur quant au choix du vocabulaire, il est recommandé de choisir des mots contrastés pour réduire le risque d'ambiguïté. Ces systèmes sont d'autant plus performants (jusqu'à 99% de taux de reconnaissance) que les mots sont bien différenciables par leur longueur ou leur transcription phonétique. Certains de ces systèmes sont multilocuteurs, et ne nécessitent pas d'apprentissage préalable [Jou88].

### 2.3.2 Les systèmes de compréhension

Les systèmes de compréhension se caractérisent par un vocabulaire limité et par une sémantique fermée.

Les applications généralement choisies sont de type par exemple comme l'interrogation d'une base de données, les standards téléphoniques automatisés qui donnent des renseignements météo, ou même permettent de réserver des places. Ces systèmes sont connectés à des modules d'interprétation de message reconnu, dont le but est de réagir soit par l'émission d'une réponse vocale soit par une action mécanique sur l'environnement après prise de décision. De ce fait, la performance de ces systèmes doit être jugée sur la base du nombre de phrases reconnues.

Beaucoup de grands systèmes ont été conçus autour du projet ARPA lancé en 1977 aux USA. Aujourd'hui, plusieurs laboratoires travaillent sur le projet DARPA (CMU, MIT, BBN, SRI ..). Par exemple, le système SPHINX développé au CMU, suppose l'emploi d'un petit vocabulaire (les milles mots du corpus "Ressources Management Data Base"). Il a par contre le mérite de permettre aux locuteurs une parole continue et ne nécessite pas d'apprentissage préalable. Le taux de reconnaissance sur les mots est de l'ordre de 96%.

### 2.3.3 Les systèmes de dictée vocale

Les machines à dicter ont pour but de retranscrire un texte dicté par un locuteur devant un microphone aussi bien qu'une secrétaire, c'est à dire, en respectant au

mieux les règles d'usage et d'accord orthographique propres à la langue utilisée. La compréhension des phrases n'est nullement requise [Bar96].

On peut remarquer que ce domaine occupe un lieu important, à la frontière de l'oral et de l'écrit. De fait, les registres de langue traités ne sont pas ceux du langage parlé, mais plutôt ceux de l'écrit. En fonction de l'application envisagée, seront dictés des rapports, des articles de journaux, des lettres administratives ... Il en résulte une complexité moindre que s'il fallait retranscrire des dialogues à l'état brut. Dans le vif d'une conversation, les phrases grammaticales se mêlent aux phrases incomplètes, tandis que fourmillent hésitations, reprises, retours en arrière, ou autres répétitions.

Historiquement, l'équipe de recherche IBM dirigée par F. Jelinek, est la première à avoir montré qu'un système à grand vocabulaire (Tangora 5000 mots en 1995) pouvait tenir dans une petite boîte portable. Par la suite, l'ensemble des grands systèmes développés se sont inspirés du système Tangora. Avec un taux de réussite supérieure à 95% pour un vocabulaire de 20 000 mots, Tangora tend à devenir un système multi-lingue existant pour l'Anglais, l'Italien, le Français et l'Allemand.

Le système Dragon fonctionne en mots isolés avec un vocabulaire de base de 16 000 mots extensible à 30 000 mots. Un de ses points forts est sa capacité d'adaptation [Bak75].

Aux USA, le second grand système vendu est la machine Kurzweil (1 000 à 10 000 mots). Le voice terminal de Kurzweil (KVT) ne s'est pas vraiment détaché de la commande vocale. Ses concepteurs affirment qu'il offre la possibilité de dicter en mots isolés un texte. Cependant les spécialistes ne lui confrère pas le statut de système de dictée [Lee88].

En France, le système de dictée développé au LIMSI (5 000 à 10 000 mots) autour du circuit  $\mu$ PCD a abouti au produit DATAVOX (5 000 mots) commercialisé par la société VECSYS. Le taux de reconnaissance publié est de 95% pour un locuteur masculin. Le système Halmet est une maquette développée parallèlement, il peut traiter un vocabulaire de 7 000 mots comme mots isolés [Bar96].

Le système développé à l'INRS (Bell Northern) par M. Lenning fonctionne en anglais avec une capacité de 86 000 mots. On trouve aussi des systèmes dédiés aux langues asiatiques comme celui réalisé pour le mandarin, qui peut traiter 60 000 mots.

Dans le tableau suivant nous donnons quelques produits qui existent sur le marché :

Société	Type de produit	Support	Approche	Reconnaissance	Application	Vocabulaire
DRAGON	Logiciel	Windows	HMM	Monolocuteur Continue	commande et dictée	5000 à 60000
SIEMENS	DSP		R.N	Multilocuteur	tél'ephonie	
IBM	Logiciel	Windows	HMM	Multilocuteur mots isolés	dictée	35000 à 65000
IBM	Logiciel	Windows	HMM	Multilocuteur continu	dictée et commande	65000 à 350000
DeDris Xi-linx			Globale	Monolocuteur mots isolés		
Spectro-chip	Carte	PC		Identification et vérification	Protection informatique	35000 à 65000
VOX	DSP	Carte	HMM	mots isolés	France télécom	
Nuance	Logiciel	Unix	HMM		Téléphonie commande et dictée	15000

TAB. 2.1 – Quelques produits du marché

Tous ces produits de marché sont destinés aux personnes normaux parlants. Les personnes qui ont un handicap articulaire n'ont pas un produit propre à eu

## 2.4 Conclusion

Dans ce chapitre nous avons décrit certains systèmes actuels de reconnaissance de la parole. Ces derniers donnent des résultats assez importants. En effet les systèmes de commande vocale donnent un taux de reconnaissance supérieur à 95 % pour des applications avec des petits vocabulaires. Alors que les systèmes de parole continue donnent des taux de reconnaissance qui peuvent atteindre 95%, mais ces systèmes nécessitent des phases important d'adaptation. Les systèmes de reconnaissance de la parole utilisent généralement la DTW pour les systèmes de reconnaissance de mots isolés et les HMM pour les systèmes de parole continue.

Malheureusement actuellement, il n'existe pas sur le marché un système de reconnaissance dédié aux personnes handicapées.

Dans le prochain chapitre nous allons donner quelques généralités sur les systèmes de reconnaissance de la parole.



## Chapitre 3

# Généralités sur les systèmes de reconnaissance de la parole

### 3.1 Introduction

Il y a quelques années, la recherche en Reconnaissance Automatique de la Parole (RAP) était considérée par le grand public comme un aimable passe-temps durant lequel on ne se préoccupait que des problèmes sans fondement réel. Aujourd'hui l'honnête homme est tout de même troublé par les performances des systèmes de reconnaissance actuels. En effet, il a quelques difficultés à réaliser correctement quelque chose d'aussi trivial que de reconnaître sa voix, à l'heure où l'informatique triomphe dans les calculs sur des milliards de données avec des résultats spectaculaires. S'il reste en effet un domaine où la réalité a du mal à dépasser la fiction, celui-ci risque fort d'être la Reconnaissance Automatique de la Parole [Jou88].

L'enregistrement du signal lui-même ne suffit pas à identifier les prononciations étant donné l'extrême variabilité du signal due aux locuteurs et au milieu environnant. La solution idéale réside dans une représentation adéquate du signal et l'élaboration d'un système réel de reconnaissance indépendant du vocabulaire et du locuteur. A l'heure actuelle, les modules classiques de reconnaissance de la parole les plus efficaces utilisent des approches statistiques et plus particulièrement des Modèles de Markov Cachés (MMC). Il en existe une grande diversité car leur utilisation est très dépendante de l'application à laquelle ils sont destinés. Les recherches autour de ces modèles sont nombreuses, chacune apportant sa contribution. Mais, le système idéal n'existe pas et les études en reconnaissance automatique de parole donnent l'impression de piétiner. Il semble que pour surpasser les performances actuelles, il fallait faire preuve d'idées originales et d'audace. Certains chercheurs le montrent en voulant introduire de nouveaux paramètres (prosodie), de nouveaux modèles (modèles articulatoires, modèles dynamiques, fusion de données). La possession d'un outil suffisamment souple permettant d'accéder à ces différents niveaux d'abstraction, est un besoin réel en recherche fondamentale sur la reconnaissance automatique de la parole [Den92].

## 3.2 Les systèmes de reconnaissance de la parole

### 3.2.1 Les différentes méthodes de reconnaissance de la parole

Nous trouvons deux méthodes de reconnaissance de parole [Jou88] :

1. La méthode globale : cette méthode considère le plus souvent le mot ou le phonème comme unité de reconnaissance minimale, c'est-à-dire indécomposable. Dans cette méthode nous comparons globalement le message d'entrée (mot, phrase) aux différentes références stockées dans un dictionnaire en utilisant des algorithmes de programmation dynamique ou des modèles de Markov cachés (HMM = Hidden Markov Model). L'avantage de cette méthode est d'éviter l'explicitation des connaissances relatives aux transitions qui apparaissent entre les phonèmes.

La généralisation de la méthode à des unités enchaînées présente un certain intérêt. En effet les unités phonétiques sont représentées par des modèles et les connaissances phonétiques, lexicales et syntaxiques sont compilées dans un seul réseau, ce qui rend le système de reconnaissance très homogène, des niveaux acoustiques jusqu'aux niveaux linguistiques. La reconnaissance consiste alors à trouver le meilleur chemin dans le réseau global pour reconnaître une phrase prononcée [Bar96].

Ce type de méthode est utilisé dans les systèmes suivants :

- reconnaissance de mots isolés.
- reconnaissance d'unités enchaînées.
- reconnaissance de parole dictée avec pauses entre les mots.

2. La méthode analytique : cette méthode fait intervenir un modèle phonétique de langage. Il y a plusieurs unités minimales pour la reconnaissance qui peuvent être choisies (syllabe, demi-syllabe, diphone, phonème, phone homogène, etc.). Le choix parmi ces unités dépend des performances des méthodes de segmentation utilisées. La reconnaissance par cette méthode, passe par la segmentation du signal de la parole en unités de décision puis par l'identification de ces unités en utilisant des méthodes de reconnaissance des formes (classification statistique, réseaux de neurones, etc.) ou des méthodes d'intelligence artificielle (systèmes experts par exemple) [Bar96].

Cette méthode est beaucoup mieux adaptée aux systèmes à grand vocabulaire et pour la parole continue. Les problèmes qui peuvent apparaître dans ce type de système sont dus en particulier aux erreurs de segmentation (délétions, insertions, substitutions, recouvrements) et d'étiquetage phonétique. C'est pourquoi le DAP (Décodage Acoustico-Phonétique) est fondamental dans une telle approche.

### 3.2.2 Le décodage acoustico-phonétique

La reconnaissance et la production de séquences sont des problèmes quotidiens pour l'être humain ; c'est que l'ajout d'une représentation mentale d'objets connus est une tâche relativement simple et inconsciente.

Elle passe par diverses étapes de traitement qui ont deux tâches principales :

- percevoir les objets
- les nommer.



Ces deux tâches sont difficilement séparables mais résument bien les enjeux de la reconnaissance des formes et de l'intelligence artificielle. Un certain nombre de traitements perceptifs et/ou moteurs sont câblés et permettent de caractériser les formes sensibles. Ces formes alors dégrossies, séparées, peuvent être communiquées au traitement cognitif qui s'attachera à les classer et à les nommer vis-à-vis d'un code appris.

### Nature du son de parole

Le signal de parole peut être considéré comme la concrétisation d'une suite de symboles abstraits (ou phonèmes) organisés selon un code linguistique et liés entre eux par des relations structurantes. La portée de ces relations sur l'axe syntagmatique définit les contextes internes, immédiats et lointains. Le contexte interne résulte de facteurs linguistiques, articulatoires etc... et agit directement sur la réalisation individuelle du phonème. Le contexte immédiat se manifeste par la déformation des phonèmes adjacents sous l'effet de la coarticulation (qui provoque des migrations de traits sur l'axe syntagmatique) et des amalgames lexicaux. Le contexte lointain opère par le biais de la prosodie de groupe en agissant sur l'émergence des phonèmes (via les syllabes accentuées) dans le groupe (accent lexical, accent de groupe, accent de phrase, etc...). Ainsi le signal parole résulte de la contribution de tous ces facteurs qui opèrent à différents niveaux de structuration [Jou88].

### Du signal parole à l'observation acoustique

Les premiers modules de traitement dans un système de reconnaissance de la parole sont les suivants :

- Acquisition et numérisation du signal : pour être utilisable par un ordinateur après son acquisition, un signal doit tout d'abord être numérisé. Cette opération tend à transformer un phénomène temporel analogique, le signal sonore dans notre cas, en une suite d'éléments discrets (les échantillons).

Le choix de la fréquence d'échantillonnage est aussi déterminant pour la définition de la bande passante représentée dans le signal numérisé. Un signal échantillonné à 16000 Hertz contient une bande de fréquences allant de 0 à 8000 Hertz.

- Extraction des paramètres du signal : Pour un signal de parole on détermine généralement les caractéristiques suivantes :

- \* l'énergie du signal : Elle est souvent évaluée sur plusieurs trames de signal successives pour pouvoir mettre en évidence des variations. Sur une fenêtre de longueur N allant de I à I + N la formule de calcul de l'énergie du signal S(k) est la suivante :

$$E(S)_I^{I+N} = \sum_{k=I}^{I+N} |S(k)|^2$$

- \* les coefficients cepstraux : les coefficients cepstraux représentent le signal parole dans le domaine fréquentiel. Ces coefficients vont servir directement dans le processus de reconnaissance de la parole.

### 3.2.3 Outils pour la reconnaissance de la parole

#### Paramétrisation de la parole

L'information acoustique pertinente du signal de parole se situe principalement dans la bande passante de fréquence allant de 50 Hz à 8000 Hz, la fréquence d'échantillonnage devrait donc au moins être égale à 16 kHz, selon le théorème de Shannon ; mais elle peut varier en fonction du domaine d'application ou des besoins ou contraintes matérielles.

La trame acoustique est un ensemble de coefficients ou paramètres, calculés sur un bloc d'échantillons. Dans la plupart des applications, ce bloc d'analyse est de taille fixe, il correspond à un temps de parole de 20 à 40 ms. La suite de vecteurs d'analyse est obtenue en déplaçant ce bloc de 10 à 20 ms ; il y a recouvrement de blocs, ce qui apparente cette analyse à une analyse de type fenêtre glissante.

De nombreuses paramétrisations ont été utilisées en parole ; les méthodes issues du traitement du signal sont classiquement répertoriées en deux catégories :

- les transformées non paramétriques usuelles telles que la transformée de Fourier
- les méthodes paramétriques qui s'appuient sur un modèle simplifié de production de la parole et qui exploitent le couplage "source/conduit" (codage prédictif linéaire).

#### 1 - Les méthodes non paramétriques

Ce type de paramétrisation fait appel aux techniques classiques utilisées en traitement de signal : les transformées temps-fréquence et temps-échelle. Malgré quelques tentatives récentes d'exploitation des transformées de type ondelettes, la transformée la plus utilisée en parole reste la transformée de Fourier discrète (FFT (Fast Fourier Transform)).

Cependant, d'une part, la période fondamentale fait apparaître de nombreuses harmoniques sur le spectre d'amplitude ainsi obtenu, et d'autre part, l'information reste redondante. Pour tenir compte de la perception humaine, le spectre est ramené à une échelle non linéaire Bark ou Mel, donnée par la formule suivante :

$$M = \frac{1000 \times \log_{10}(1 + \frac{F}{1000})}{\log_{10}(2)}$$

$F$  en Hz et  $M$  en Mel

Afin de réduire l'information, une suite de fenêtres (triangulaires, rectangulaires...) est appliquée dans le domaine spectral selon l'une des échelles précédemment décrites. Les coefficients obtenus sont alors synonymes d'énergie dans des bandes de fréquence.

#### 2 - Les méthodes paramétriques

Ces méthodes tiennent compte du processus de phonation et s'appuient sur un modèle linéaire simplifié de production de la parole. Le signal vocal est considéré comme la sortie d'un filtre excité par une source. Le filtre modélise le conduit nasal, le conduit vocal et le rayonnement aux lèvres, tandis que la source correspond à un signal périodique qui est un son voisé ou un bruit aléatoire qui est un son non voisé. L'analyse LPC (Linear Prediction Coding) simplifie ce modèle de production

en supposant que le filtre ne comporte que des pôles. Les paramètres sont alors les coefficients du filtre, ils décrivent la fonction de transfert du conduit vocal.

### 3.2.4 Méthodes de reconnaissance de la parole

Les approches traitées sont des approches statistiques qui ont donné naissance à des systèmes dits auto-organiseurs fondés soit sur [HL00] :

- une comparaison directe par calcul de distances avec des références acoustiques (comparaison dynamique).
- un calcul probabiliste à l'aide de modèles stochastiques; les plus utilisés actuellement dérivent des modèles de Markov cachés (Hidden Markov Models HMM)

#### La comparaison dynamique

Compte tenu de la forte variabilité inter-locuteur et même intra-locuteur, les prononciations d'un même mot se réalisent acoustiquement de manière fort différente; entre autres, on observe une distorsion temporelle qui implique que les échelles temporelles des deux occurrences du même mot ne coïncident pas. On ne peut pas donc comparer point à point les formes acoustiques, et il est nécessaire de procéder à un alignement dit temporel [Bel57].

Cette comparaison s'effectue par programmation dynamique. Elle est fondée sur les travaux de R. Bellman en 1957 pour la recherche de la trajectoire optimale. Pour formuler un système de pondération simple, celle-ci a été reprise par Sakoe et Shiba en 1970 .

Étant donné deux images acoustiques  $T = (T_i, 1 \leq i \leq I)$  et  $R = (R_j, 1 \leq j \leq J)$  de longueurs respectives  $I$  et  $J$  et la connaissance des distances entre deux vecteurs  $T_i$  et  $R_j$ ,  $d(i,j)$ , la distance entre  $T$  et  $R$  est constituée de l'accumulation des distances entre les événements  $T_{ik}$  et  $T_{jk}$  que l'on peut pondérer, le long d'un chemin  $C = ((i_k, j_k))$  de longueur  $K$ . La comparaison entre  $T$  et  $R$  revient à faire la recherche du chemin optimal  $C$  qui minimise cette somme cumulée en respectant les contraintes minimales suivantes :

- faire coïncider les extrémités : le chemin doit commencer en  $(T_1, R_1)$  et finir en  $(T_I, R_J)$ ,
- obtenir une croissance temporelle stricte  $C_{k-1} < C_k$ .
- respecter la continuité du chemin.

L'application de ces contraintes permet de respecter la dimension temporelle des signaux : on ne peut progresser indéfiniment dans les trames d'un signal tout en restant dans la même trame de l'autre signal . On ne peut pas non plus régresser dans le parcours des signaux. On est obligé de parvenir en fin d'analyse à la fin des représentations des deux images acoustiques comparées.

En plus de ces contraintes minimales s'ajoutent d'autres contraintes locales dites de cheminement liées aux connaissances a priori sur le débit oral. Le coût d'un chemin est obtenu par la formule récurrente de Sakoe et Shiba ici appliquée à une contrainte donnée.

1) Initialisation :

$$D(1, 1) = d(1, 1) \quad \text{et} \quad D(1, j) = D(j, 1) = \inf$$

2) Récurrence :

$$D(i, j) = \min \left\{ \begin{array}{l} D(i-1, j-1) + 2 \times d(i, j) \\ D(i-1, j-2) + d(i, j-1) + d(i, j) \\ D(i-2, j-1) + d(i-1, j) + d(i, j) \end{array} \right\}$$

3) Coût du chemin optimal D

$$D = \frac{D(I, J)}{I + J}$$

Dans un système de reconnaissance de mots isolés, fondé sur cette méthode, un ensemble de références est constitué lors de la phase d'apprentissage ; pour chaque mot de vocabulaire, est acquise une (ou plusieurs) prononciation, chacune d'elles est paramétrée et donne naissance à une référence ou suite de vecteurs acoustiques R. Lors de la phase de reconnaissance, pour chaque prononciation à reconnaître T, on calcule toutes les distances entre l'observation T et les références R par comparaison dynamique.

Le mot reconnu est celui qui correspond à la référence pour laquelle la distance à T est de coût minimal.

La comparaison dynamique était une des méthodes les plus utilisées dans les cartes vocales. Le système MOZART développé au LIMSI et commercialisé par la société VECSYS donne de très bons résultats.

Cette carte permet :

- la reconnaissance de mots isolés avec des pauses inférieures à 200 ms.
- la reconnaissance de mots connectés avec un premier apprentissage en mots isolés puis un second en mots connectés.
- la détection de mots clé (word-spotting) avec un premier apprentissage en mots isolés puis un second en prononçant une phrase contenant les mots clés dont on garde les références.

La tendance actuelle est d'utiliser des Modèles de Markov Cachés qui produisent aussi de bons résultats pour le même type d'application : la référence est constituée par un ou plusieurs MMC modélisant globalement le mot.

### Les modèles de Markov cachés

Un modèle de Markov caché (MMC) ou Hidden Markov Model en anglais (HMM) est un graphe probabilisé dans lequel chaque noeud est censé produire un ou plusieurs segments stables ou transitoires du signal. A chaque état ou noeud est associée une distribution de probabilité d'émettre un vecteur spectral [Rab89a].

1- Modèle théorique

Un modèle de Markov cachée est un double processus stochastique  $(X_t, Y_t) \quad 1 \leq t \leq T$ ). La chaîne interne  $X_t$  non observable, et la chaîne externe  $Y_t$  observable, s'allient pour générer le processus stochastique. La chaîne interne est supposée, pour

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

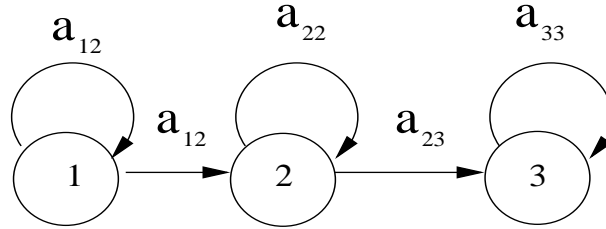


FIG. 3.1 – Chaîne de Markov à trois états gauche-droite

Le processus  $(X_t) \quad 0 \leq t \leq T$  est une chaîne de Markov d'ordre 1 s'il doit vérifier :

$$P(X_{t+1} = q_j / X_t = q_i, \dots, X_0 = q_0) = P(X_{t+1} = q_j / X_t = q_i) = a_{ij}$$

pour tout  $t \geq 0$ .

Le processus  $(Y_t) \quad 0 \leq t \leq T$ , processus observable, vérifie :

$$P(Y_t = y_t | X_t = q_i, \dots, X_1 = q_1, Y_{t-1} = y_{t-1}, \dots, Y_1 = y_1) = P(Y_t = y_t / X_t = q_i) = b_i(y_t)$$

Les observations sont supposées indépendantes les unes des autres conditionnellement à la suite d'états. Chaque réalisation de  $Y_t$  ne dépend que de l'état courant caché.

$$P(Y_{1:t} | X_{1:t}) = \prod_{i=1}^t P(Y_i | X_i)$$

Les observations  $Y_t$  peuvent être de nature :

- discrètes :  $b_i$  est une distribution de probabilité discrète : une loi discrète est généralement représentée par les fréquences d'apparition des observations discrètes.
- continues :  $b_i$  est une fonction de densité de probabilité : les densités traditionnelles utilisées sont des densités gaussiennes, entièrement définies par le vecteur moyenne et la matrice de covariance, ou des densités de type mélange de gaussiennes (sommes pondérées de densités gaussiennes).

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

– son ensemble fini d'états  $Q = (q_1, \dots, q_N)$

– son ensemble de probabilités de transitions entre les états

$$A = (a_{ij}) \quad 1 \leq i \leq N \quad 1 \leq j \leq N$$

– son ensemble de lois (ou densités) de probabilités associées à un état

$$B = (b_i(y_t)) \quad 1 \leq i \leq N$$

– son ensemble de probabilités initiales

$\pi = (\pi_i) \quad 1 \leq i \leq N$ , ou  $\pi_i$  désigne la probabilité d'entrer dans le modèle par l'état initial  $q_i$ . Cette probabilité est généralement égale à  $(\frac{1}{N})$ .

Un modèle de Markov caché est ainsi décrit par le jeu de paramètres  $\lambda = (\pi, A, B)$ .

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

soient :

$Y = y_1, \dots, y_T$  la suite d'observations

$Q_i = q_{i1}, \dots, q_{iT}$  la suite d'états de longueur T le long du chemin i

$$\begin{aligned} P(Y_1 = y_1, \dots, Y_T = y_T | \lambda) &= \pi_i \times \sum P(X_1 = q_{i1}, \dots, X_T = q_{iT}, Y_1 = y_1, \dots, Y_T = y_T | \lambda) \\ &= \pi_i \times P(Y_1 = y_1, \dots, Y_T = y_T | X_1 = q_{i1}, \dots, X_T = q_{iT}, \lambda) \\ &\quad \times P(X_1 = q_{i1}, \dots, X_T = q_{iT}) \end{aligned}$$

## 2- Utilisation des MMC en RAP

Il existe différentes manières d'utiliser les MMC en reconnaissance automatique de la parole, selon l'application visée :

- dans le cadre d'une reconnaissance de mots isolés en nombre limité (moins de 1000 mots), chaque mot  $m_v$  du vocabulaire V est modélisé par un  $MMC_v$ . La phase de reconnaissance consiste pour une suite d'observations acoustiques données, à calculer la vraisemblance de ces observations par rapport à chacun des modèles et à considérer comme mot reconnu le mot correspondant à la vraisemblance maximum. Cette approche n'est qu'une version améliorée de l'approche de type programmation dynamique.

Cette approche n'est réalisable qu'avec un vocabulaire limité dans la mesure où l'algorithme de reconnaissance considère tous les chemins modélisant chaque mot  $m_v$ , ce qui entraîne un important coût de calcul.

- lorsque la taille du vocabulaire augmente et/ou lorsqu'il s'agit de reconnaissance de mots connectés, un modèle de Markov global est construit à partir de modèles élémentaires.

- \* en reconnaissance de mots isolés en nombre inférieur à 1000, chaque mot  $m_v$  du vocabulaire V est modélisé par un  $MMC_v$ . Tous les  $MMC_v$  sont liés entre eux par une entrée et une fin commune dans un réseau global. L'entrée correspond à un MMC représentant un silence précédant le mot, tandis que la fin correspond à un MMC modélisant un silence suivant le mot.

Chaque chemin du réseau global correspond à la prononciation d'un mot. En phase de reconnaissance, le mot reconnu correspond alors au chemin le plus probable étant donné la suite d'observations.

- \* en reconnaissance de mots connectés, pour des vocabulaires de taille importante, l'utilisation de modèles globaux pour tous les mots du dictionnaire soulève quelques problèmes :

- le stockage de tous les mots du vocabulaire devient très important
- une grande quantité de parole est nécessaire pour réaliser l'estimation statistique de tous les paramètres (probabilité de transitions, lois d'émissions d'observations).

C'est pourquoi on préfère représenter phonétiquement les mots à partir d'unités phonétiques. Un MMC élémentaire correspond alors à une unité phonétique.

Actuellement le choix de l'unité phonétique reste un problème ouvert. On peut choisir des unités de l'ordre du phonème, mais on peut alors difficilement traduire les variations dues aux contextes.

Une unité plus simple d'un point de vue modélisation est le fenone ; unité introduite dans les systèmes développés chez IBM, elle correspond à un MMC élémentaire au sens "machine", réduit à un seul état émetteur et à une boucle, et représentant un état acoustique élémentaire.

A partir de ces modèles est construit par concaténation un modèle de Markov caché pour chaque mot du vocabulaire, en tenant compte des différentes variantes de prononciation ; le modèle global est obtenu en reliant les modèles de mots selon la syntaxe de l'application. Ce type de construction permet de prendre en compte lors des différentes substitutions et concaténations, des connaissances de type phonétique et phonologique.

Au niveau phonétique, il est ainsi possible de rendre compte de phénomènes spécifiques tels que la décomposition d'un son en phases élémentaires.

Au niveau phonologique sont introduites des règles inter-mot et intra-mot afin de prendre en compte les phénomènes de coarticulation et d'assimilation.

Tout chemin du graphe ainsi obtenu correspond à une phrase du langage à reconnaître et la phase de reconnaissance consiste alors, comme précédemment, à rechercher au sein du graphe le chemin le plus probable pour la suite d'observations donnée.

Le problème posé est en fait celui de la modélisation à plusieurs niveaux par des modèles MMC : les modèles de phrases sont construits à partir de modèles de mots qui eux-mêmes sont réalisés avec des modèles acoustiques. Ce procédé peut être réalisé de manière récursive. Lorsque l'on aborde la reconnaissance de parole continue (très grand vocabulaire, syntaxe non contrainte), on ne peut compiler toutes les phrases possibles à partir des mots du vocabulaire. Une solution consiste à combiner les modèles MMC de mots en les connectant à tous les mots successeurs potentiels, selon un modèle dit de langage.

Pour prendre en compte plusieurs locuteurs, la modélisation du signal de parole doit être améliorée par une augmentation du nombre de paramètres des MMC. En général, l'accroissement du nombre de paramètres s'effectue selon deux grandes méthodes : l'utilisation de multi-modèles et l'emploi de densités multi-gaussiennes.

Dans le cas des multi-modèles, chaque unité à reconnaître est modélisée par plusieurs modèles. En phase de reconnaissance, l'algorithme de Viterbi cherche le meilleur chemin dans le réseau global, chacune des branches menant aux modèles étant équiprobables. L'apprentissage peut être :

- standard : les observations sont utilisées sans information a priori en utilisant l'ensemble du réseau (Baum Welch ou Viterbi).
- à partir d'une classification a priori des locuteurs : on affecte à chaque branche les observations correspondants à un sous ensemble de locuteurs choisis (l'approche classique consiste à définir deux sous ensembles correspondant aux locuteurs masculins et féminins)
- à partir d'une classification automatique des locuteurs : les branches correspondent à des classes de locuteurs. Les observations sont alors affectées aux branches

à partir d'un critère (critère de vraisemblance ou de taux d'erreurs).

Dans le cas des multi-gaussiennes, on augmente le nombre initial de lois d'émissions de probabilité jusqu'à approcher une distribution donnée.

En phase de reconnaissance l'algorithme de Viterbi cherche le chemin optimal en "mélangeant" toutes les lois d'émissions. L'apprentissage est itératif ; avant toute nouvelle itération, chaque gaussienne donne naissance à deux nouvelles gaussiennes en perturbant légèrement les paramètres de la gaussienne considérée. Le nombre théorique de gaussiennes est multiplié par deux à chaque étape.

Pour un multi-modèles, lors de la phase de reconnaissance, une fois que l'algorithme de Viterbi a choisi un début de chemin, le reste du chemin est fortement "guidé", contrairement au modèle multi-gaussiennes où le chemin optimal emprunte des lois d'émissions indépendamment de leur mode de séparation lors de l'apprentissage.

En conclusion, la réalisation d'un système de reconnaissance à base de modèles de Markov cachés, s'effectue en trois phases :

- Décrire un réseau dont la topologie reflète les phrases, mots du vocabulaire ou unités élémentaires à traiter.
- Réaliser un apprentissage des paramètres du (ou des) réseau(x)  $\lambda = (\pi, A, B)$ .
- Effectuer la reconnaissance proprement dite par calcul de vraisemblance.

### 3- Phase de reconnaissance

La phase de reconnaissance selon l'application visée, consiste à calculer pour une suite d'observations acoustiques, soit sa vraisemblance par rapport à un modèle  $\lambda$ , soit la probabilité du chemin optimal l'ayant générée. Deux algorithmes ont été développés pour résoudre ces deux problèmes, l'algorithme de Baum Welch pour le calcul de vraisemblance et l'algorithme de Viterbi pour le calcul du chemin optimal.

\* L'algorithme de Baum-Welch repose sur le calcul de deux fonctions :

si  $Y = y_1, \dots, y_T$  désigne une suite d'observations,

- la fonction forward  $\alpha(t, q_j)$  représente la probabilité d'observer les  $t$  premières observations et d'être à l'instant  $t$  dans l'état  $q_j$ ,

- la fonction backward  $\beta(t, q_j)$  représente la probabilité d'observer les  $(T-t)$  dernières observations sachant que l'on est à l'instant  $t$  dans l'état  $q_j$ .

Ces deux fonctions se calculent par récurrence sur le temps :

$$\alpha(t, q_j) = P(X_t = q_j, Y_{1:t}) = \sum_i \alpha(t-1, q_i) \times a_{ij} \times b_j(y_t)$$

$$\alpha(1, q_j) = \pi_j \times b_j(y_1).$$

$$\beta(t, q_j) = \sum_i a_{ji} \times b_i(y_{t+1}) \times \beta(t+1, q_i).$$

$$\beta(T, q_j) = 1$$

On en déduit la vraisemblance de la suite d'observations par rapport au modèle :

$$P(Y_1 = y_1, \dots, Y_T = y_T | \lambda) = \sum_j \alpha(T, q_j)$$

ou

$$P(y_1 = y_1, \dots, Y_T = y_T | \lambda) = \sum_j \pi_j \times b_j(y_1) \beta(1, q_j).$$

\* L'algorithme de Viterbi recherche le meilleur chemin au sens probabiliste, ayant généré la suite d'observations  $Y = y_1, \dots, y_T$  :

Nous nous situons dans le cas général où les émissions des lois de probabilités sont sur les états et où il n'y a pas de transition vide. Si  $Q_i = q_{i1}, \dots, q_{iT}$  désigne une suite d'états de longueur  $T$  le long d'un chemin  $i$ , le chemin recherché vérifie



$$P(X_1 = q_{i1}, \dots, X_T = q_{iT}, y_1 = y_1, \dots, Y_T = y_t | \lambda) = \max P(X_1 = q_{i1}, \dots, X_T = q_{iT}, y_1 = y_1, \dots, Y_T = y_t | \lambda)$$

Soit à  $C(t, q_j)$  la probabilité d'émission de la sous suite d'observations  $y_1 \dots y_t$ , le long du chemin le plus probable tel que  $X_t = q_j$ ,

$$C(t, q_j) = \max C(t-1, q_i) \times a_{ij} \times b_j(y_t)$$

Le chemin optimal est obtenu à l'aide de la fonction auxiliaire

$$D(t, q_j) = \argMax D(t-1, q_i) \times a_{ij}$$

par récurrence arrière.

#### 4 - Apprentissage d'un MMC

La puissance de l'approche markovienne réside dans l'automatisation de l'apprentissage des paramètres  $\lambda$ , qui se réalise à l'aide des algorithmes de Baum-Welch ou de Viterbi.

\* APPRENTISSAGE par la procédure de BAUM WELCH :

Si  $W = (w_1, \dots, w_n, \dots, w_N)$  désigne un ensemble de prononciations de phrases acceptées, la phase d'apprentissage consiste à rechercher le modèle  $\lambda$  par rapport auquel la vraisemblance de  $W$  est maximale :

$$\hat{\lambda} = \arg \max_n P(w_n | \lambda)$$

La maximisation directe de la fonction de vraisemblance est impossible et est remplacée par un algorithme itératif qui, à chaque étape, maximise une fonction auxiliaire définie comme suit :

$$Q(\hat{\lambda}, \lambda) = P(i, w_n | \lambda) \log P(i, w_n | \hat{\lambda})$$

où  $i$  est un chemin quelconque du modèle défini par  $\lambda$ .

Moyennant quelques hypothèses sur la nature des lois d'observations, à chaque itération, partant d'un modèle  $\lambda$ , le nouveau jeu de paramètres  $\hat{\lambda}$  qui maximise la fonction  $Q$  par rapport à  $\lambda$ , est meilleur dans le sens où :

$$P(w_n | \hat{\lambda}) \geq P(w_n | \lambda)$$

La maximisation de la fonction  $Q$  est explicite et les formules de ré-estimation mettent à profit les fonctions forward et backward de Baum.

\* APPRENTISSAGE par la procédure de VITERBI :

Si  $W = (w_1, \dots, w_n, \dots, w_N)$  désigne de nouveau un ensemble d'apprentissage, le modèle  $\lambda$  est défini comme suit :

$$\hat{\lambda} = \argMax P(X_n, w_n | \lambda)$$

où  $X_n$  représente le chemin optimal générant la suite d'observations  $w_n$  par rapport à  $\lambda$ .

La procédure de ré-estimation est semblable à celle développée dans l'algorithme de Baum Welch. Elle consiste à maximiser la fonction auxiliaire :

$$Q(\lambda, \hat{\lambda}) = \sum P(x, w | \hat{\lambda}) \times \log[P(x, w | \lambda)].$$

où  $i$  est un chemin quelconque du modèle défini par  $\lambda$ ,  $X_n$  le chemin optimal générant  $w_n$  selon le modèle  $\lambda$ .

Les formules de ré-estimation s'apparentent alors à de simples comptages d'événements, en utilisant l'algorithme de Viterbi pour spécifier pour chaque prononciation le chemin optimal.

L'arrêt de ces deux procédures s'effectue lorsque les variations du critère utilisé sont inférieures à un seuil donné ou lorsque le nombre d'itérations est supérieur à un seuil également.

Un inconvénient de la RAP par MMC réside dans le choix du corpus d'apprentissage : il doit être le plus complet possible afin de pouvoir estimer correctement toutes les lois de transition et toutes les lois d'émissions d'observations. Dès que la syntaxe de l'application devient trop complexe, voire naturelle, et dès que la taille du vocabulaire dépasse certaines valeurs (1000 mots), la phase d'apprentissage devient longue et fastidieuse ; le volume des observations de l'ensemble d'apprentissage peut être prohibitif dans certaines applications.

### De l'observation acoustique à la forme lexicale finale

Dans les systèmes de reconnaissance de la parole utilisant les MMC nous avons besoin d'un dictionnaire et d'un modèle de langage pour construire un système de reconnaissance de la parole [Bar96]

1. Dictionnaire phonétique et modèles d'unités plus longues : Dans le cadre de la reconnaissance de la parole continue, même si le système acoustique est basé sur des phonèmes, il faut obtenir, pour chaque entrée du dictionnaire phonétique, un modèle qui lui est propre. Ces modèles sont obtenus par concaténation de HMMs de phonèmes. Par exemple le dictionnaire est constitué des mots et chaque mot est une concaténation d'une suite de phonèmes.

Exemple de dictionnaire :

Ouvrir oo vv rr ii rr

Fermer ff ee rr mm ei

porte pp oo rr tt

télé tt ei ll ei

Le dictionnaire définit la suite de phonèmes possible, le système de reconnaissance cherche donc dans un espace bien défini d'avance.

Bien entendu, dans l'éventualité de l'utilisation d'allophones, cette concaténation tient compte des contextes des phonèmes. Il est possible d'obtenir des modèles de phrases en concaténant à leur tour des modèles de mots. Cela est particulièrement intéressant dans les procédures d'alignement de phonèmes sur des signaux dont on possède la transcription phonétique.

2. Algorithmes de recherche : de nos jours, il existe beaucoup d'algorithmes de reconnaissance basés sur des HMMs. Il existe presque autant de variantes que de systèmes de reconnaissance. Avec les progrès intervenus au cours de la dernière décennie, les capacités des systèmes de reconnaissance ont considérablement augmenté. Elles sont passées de quelques mots reconnus en mode isolé pour un seul locuteur à des systèmes multilocuteurs avec plusieurs milliers de mots, en parole continue voire spontanée. Tout cela amène de nouveaux problèmes. Il est nécessaire de limiter l'espace de recherche, qui croît de manière exponentielle, pour obtenir le bon résultat avec un temps de traitement correct. Les algorithmes de reconnaissance incluent donc très souvent des stratégies permettant de choisir un nombre limité d'hypothèses à chaque instant, et ainsi de n'explorer que l'espace suffisant pour trouver la meilleure solution.

- Algorithme a étoile : L'algorithme a étoile a été adapté de nombreuses fois pour être intégré dans des systèmes de reconnaissance de la parole.

Considérons un graphe  $G$  défini par deux ensembles,  $S$  contenant l'ensemble des sommets et  $A$  celui des arcs, un arc étant un couple  $(a_n, a_m)$  de sommets reliés entre eux. Dans notre application, le graphe représente les différentes possibilités de progression sur le chemin acoustique. La particularité de l'algorithme  $A^*$  est d'utiliser une fonction heuristique pour guider la recherche qui dépend non seulement du chemin déjà parcouru mais aussi d'une estimation du chemin qui reste à parcourir. Cette fonction peut s'exprimer avec des probabilités exprimées sous forme logarithmique comme dans l'équation suivante.

$$f(n) = g(n) + h(n)$$

Dans cette équation [HL00], nous trouvons  $g(n)$  qui représente le score du chemin pour arriver à l'état courant  $n$  et  $h(n)$ , une estimation du score pour atteindre le noeud final. Cette deuxième fonction est souvent appelée sonde, puisqu'elle permet de guider l'algorithme pour qu'il choisisse le chemin le plus prometteur. Il est aisé de comprendre que le problème de l'utilisation du  $A^*$  se résume à la définition d'une fonction  $h$  exprimant correctement le poids du chemin jusqu'au noeud final. Plusieurs approches ont été proposées comme notamment, l'utilisation pour  $g$  et  $h$  des fonctions de l'algorithme forward-backward employées lors de la phase d'apprentissage du module acoustique. à chaque pas de l'algorithme, pour tous les chemins en cours, le chemin le plus prometteur au sens de  $f$  est étendu, l'hypothèse de reconnaissance étant le chemin qui atteint le noeud final.

L'avantage principal de l'algorithme  $A^*$  est de pouvoir fournir en une seule passe les  $n$  meilleurs chemins au sein du graphe, il suffit pour cela de garder une pile des  $n$  chemins les plus prometteurs et de les étendre à chaque fois, et ainsi de limiter l'espace de recherche tout en fournissant plusieurs résultats.

- Algorithme à base de modélisation arborescente : si l'on examine un dictionnaire phonétique, on se rend vite compte qu'il existe beaucoup de préfixes communs à tous les mots du dictionnaire. Une modélisation arborescente devient alors très efficace pour représenter les suites de phonèmes des mots. Elle permet un gain de place non négligeable en mémoire, mais introduit une incertitude au niveau des algorithmes de reconnaissance. En effet, dans une représentation standard des mots, il est facile de connaître instantanément le mot en cours lors de la phase de reconnaissance, alors que cela devient impossible dans une représentation arborescente tant que l'on n'a pas atteint une feuille de l'arbre.

Ce mode de stockage est aussi très intéressant, dans le cas de variantes phonétiques de mêmes mots. Certaines approches ont aussi utilisé la mise en commun des suffixes, ce qui aboutit à la construction d'un graphe. En utilisant ces méthodes, dans un système reconnaissant 65000 mots basés sur des phonèmes simples, il ne subsiste, au début de chaque nouveau mot, que l'ensemble des phonèmes connus à évaluer, puisque tous les débuts de mots ont été factorisés. Même dans le cas d'un système utilisant des allophones, cela entraîne un gain non négligeable en temps de reconnaissance puisque le maximum de HMMs à considérer est le nombre d'allophones présents dans le système. Pourtant, la non-connaissance du mot en cours introduit un problème dans l'intégration du modèle de langage. Ainsi, lorsque l'on arrive à la fin d'un mot (feuille de l'arbre), il faut choisir le meilleur chemin, soit le meilleur prédécesseur de ce mot, en utilisant le modèle de langage. Dans certains

cas, cette technique peut conduire à des erreurs, l'information du modèle de langage étant incluse trop tard, en fixant la meilleure position pour le début du mot courant en se basant uniquement sur les scores acoustiques de l'hypothèse en cours. Des travaux précédents ont montré que cela pouvait conduire à des problèmes de segmentation et de décision qui conduisent à un accroissement des erreurs pouvant aller jusqu'à 12%. Pour pallier ce problème, ils ont proposé de garder, pour tous les prédécesseurs possibles une copie de l'arbre. Pourtant, si l'on ne limite pas les hypothèses, ces copies peuvent rapidement prendre beaucoup de mémoire. On peut employer alors une méthode d'élagage d'arbre (pruning en anglais) nommée Beam search. Plutôt que d'utiliser un seuil fixe pour réduire l'espace de recherche en éliminant les hypothèses peu probables, cette technique utilise comme référence le score de la meilleure hypothèse. Un seuil empirique est alors fixé pour définir quels sont les chemins, par rapport au plus prometteur, qui seront gardés. Ainsi, il est possible de limiter le nombre de copies des arbres que l'on possède en mémoire. Cette technique possède aussi l'énorme avantage, par rapport à un seuil fixe, de ne jamais conduire à des non réponses car on garde toujours au moins la meilleure hypothèse.

- Algorithme de résolution de treillis de mots Les algorithmes résolvant des treillis de mots sont très souvent utilisés comme seconde passe de reconnaissance. Il s'agit d'utiliser les résultats d'une précédente phase de reconnaissance afin de construire un graphe de mots réévalué avec des méthodes plus coûteuses, en temps de calcul ou en mémoire, pour trouver une solution optimale, comme par exemple un modèle acoustique présentant plus de modèles d'allophones, un modèle de langage différent, etc. La construction de ce graphe est basée sur la pile des résultats de la phase de reconnaissance précédente. Ainsi, tous les mots finissants au début d'un autre mot sont ses prédécesseurs dans le graphe. Il est possible aussi d'inclure, sur chaque noeud, la probabilité de l'hypothèse ayant produit le chemin menant à ce mot.

### 3.3 Conclusion

L'approche la plus utilisée en reconnaissance de la parole est l'approche statistique, surtout pour les systèmes de reconnaissance de la parole continue. Cette approche a l'inconvénient de nécessité d'utiliser une grande base de données pour l'apprentissage. La méthode de comparaison dynamique est généralement utilisée pour les systèmes de reconnaissance de mots isolés.

Dans le prochain chapitre nous allons analyser la parole pour quelques personnes handicapée (Parkinsoniens)

## Chapitre 4

# Étude des handicaps articulatoires

### 4.1 Introduction

Le signal parole présente plusieurs propriétés qui lui sont spécifiques.

En premier lieu, la parole est un phénomène directionnel, elle est étalée dans le temps, et possède par définition un début, un milieu et une fin. De ce fait, l'auditeur reçoit les informations relatives au mot cible bout par bout seulement [Mas00].

En deuxième lieu, la parole est continue. Le signal de parole ne comporte pas d'espace, ou périodes de silences signalant à l'auditeur où se situent les frontières entre phonèmes ou entre mots. Le caractère continu de la parole soulève un problème de mise en correspondance entre un signal d'entrée continu et des représentations lexicales discrètes.

Pour une personne handicapée, une difficulté de production de quelques sons de parole peut agir sur les performances des systèmes de reconnaissance de la parole. En fait le handicap articulatoire peut provenir de deux sources, soit d'un handicap mental, soit d'un handicap physique de l'appareil phonatoire.

Dans ce chapitre nous allons étudier la production de la parole chez un être humain, puis nous analysons les difficultés de production de la parole chez quelques sujets.

### 4.2 Production de la parole

La parole est le résultat acoustique d'une série de mouvement des appareils respiratoires et articulatoires. Donc la production de la parole est le passage de l'air dans l'appareil phonatoire (un ensemble de cavités).

La production de la parole chez une personne handicapée articulatoire manifeste des troubles de production. Nous allons donner un bref aperçu sur la production de la parole, puis nous allons étudier ces perturbations pour quelques pratiques que nous avons rencontré.

### 4.2.1 La production du son

La majorité des sons du langage sont le fait du passage d'une colonne d'air venant des poumons, qui traverse un ou plusieurs résonateurs de l'appareil phonatoire [Mas00].

Les résonateurs principaux sont :

- le pharynx ;
- la cavité buccale ;
- la cavité labiale ;
- les fosses nasales.

La présence ou l'absence d'obstacles sur le parcours de la colonne d'air modifie la nature du son produit. C'est, entre autres, en classant ces obstacles éventuels que la phonétique articulatoire dégage les différentes classes de sons.

Pour un petit nombre de réalisations, l'air ne provient pas des poumons, mais de l'extérieur, par inspiration. Une articulation peut aussi être engendrée par une variation de pression entre l'air interne et l'air externe à la cavité buccale, voire même par une variation de pression purement interne (c'est le cas des clics par exemple). Le schéma de la figure 4.1 représente les éléments essentielles de l'appareil Phonatoire.

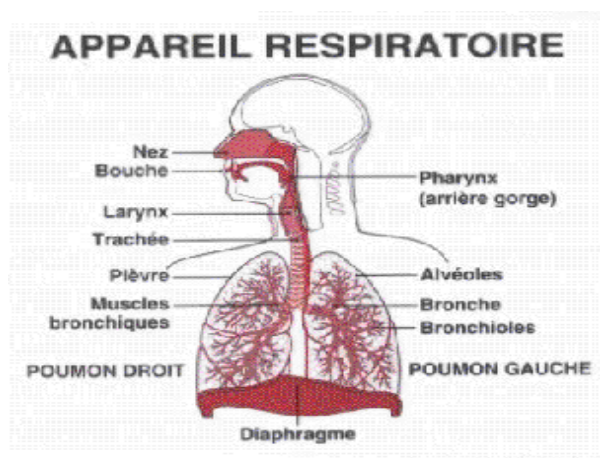


FIG. 4.1 – *appareil respiratoire*

### 4.2.2 Consonnes et voyelles

La distinction entre voyelles et consonnes s'effectue de la manière suivante :

- Si le passage de l'air se fait librement à partir de la glotte, on a affaire à une voyelle ;

En français, on compte 11 voyelles orales à savoir :

- [I] (pis)
- [EI] (épée)
- [AI] (épais)
- [A] (pas)

[AA] (pâte)

[O] (porc)

[OO] (peau)

[OU] (pou)

[U] (pur)

[EU] (peu)

[OE] (peur)

Il y a en outre 4 voyelles nasales :

[IN] (brin)

[AN] (branche)

[ON] (bon)

[EN] (brun)

- Si le passage de l'air à partir de la glotte est obstrué, complètement ou partiellement, en un ou plusieurs endroits, on a affaire à une consonne. Donc la production des consonnes est plus complexe que celle des voyelles. On distingue en français 4 groupes de consonnes :

\* Les fricatives sont caractérisées par un écoulement turbulent de l'air à travers une construction étroite formée dans le conduit vocal. Les fricatives non voisées sont [F], [S] et [CH]. les fricatives voisées sont [V], [Z] et [J].

\* Les plosives sont caractérisées par une brève occlusion suivie d'une détente du conduit vocal. Les plosives non voisées sont [P], [T] et [K]. les plosives voisées sont [B], [D] et [G].

\* Les nasales sont produites à l'aide de conduit nasal telle que [M], [N] et [GN].

La consonne latérale [L]

La consonne vibrante [R]

Le passage des consonnes aux voyelles ne se fait pas de manière abrupte, mais sur un continuum. On distinguera ainsi des articulations intermédiaires, comme les vocoïdes (par exemple les semi-voyelles) ou les spirantes. Le schéma suivant explique le principe de production des consonnes et des voyelles.

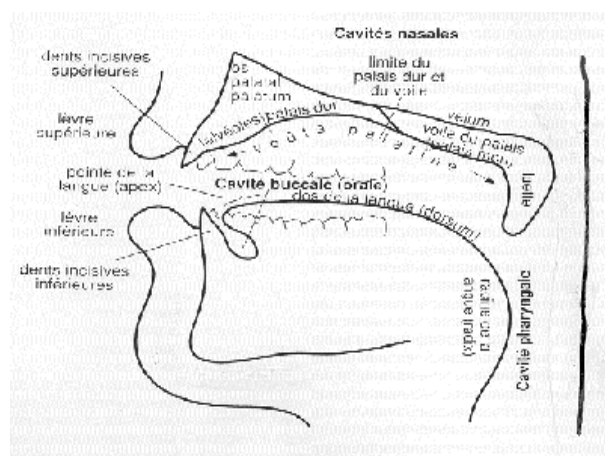


FIG. 4.2 – le principe de production des consonnes et des voyelles

### 4.2.3 Point d'articulation et mode d'articulation

La distinction entre mode d'articulation et point d'articulation est particulièrement importante pour le classement des consonnes.

Le mode d'articulation est défini par un certain nombre de facteurs qui modifient la nature du courant d'air expiré :

- libre passage, ou mise en vibration, de l'air au niveau de la glotte (sourde ou sonore) ;
- libre passage, ou non, en un point quelconque (le point d'articulation) des cavités supra-glottiques (voyelle ou consonne) ;
- passage par une voie unique ou deux voies différentes (orale ou nasale) ;
- passage, dans le conduit buccal, par une voie médiane ou latérale (la plupart des articulations opposées aux latérales).

Le point d'articulation est l'endroit où se trouve, dans la cavité buccale, un obstacle au passage de l'air. De manière générale, on peut dire que le point d'articulation est l'endroit où vient se placer la langue pour obstruer le passage du canal d'air.

Le point d'articulation peut se situer aux endroits suivants :

- les lèvres (articulations labiales ou bilabiales) ;
- les dents (articulations dentales) ;
- les lèvres et les dents (articulations labiodentales) ;
- les alvéoles (c'est-à-dire les gencives internes des incisives supérieures, articulations alvéolaires) ;
- le palais (vu sa grande surface, on peut distinguer des articulations pré-palatales, médio-palatales et post-palatales) ;
- le voile du palais (palais mou, articulations vélaires) ;
- la luette (articulations dites uvulaires) ;
- le pharynx (articulations pharyngales) ;
- la glotte (articulations glottales).



#### 4.2.4 Sourdes ou sonores

Une réalisation est dite sourde lorsque les cordes vocales ne vibrent pas ; si celles-ci entrent en vibration, la réalisation sera dite sonore. Les cordes vocales sont des replis musculaires situés au niveau de la glotte.

La vibration des cordes vocales est le résultat d'une obstruction de la glotte : celles-ci vibrent sous la pression de l'air interne qui force un passage entre elles.

#### 4.2.5 Orale et nasale

Au carrefour du pharynx, le passage de l'air peut s'effectuer dans une ou deux directions, selon la position du voile du palais :

- si le voile du palais est relevé, l'accès aux fosses nasales est bloqué, et l'air ne peut traverser que la cavité buccale ;
- si le voile du palais est abaissé, une partie de l'air traversera les fosses nasales (l'autre partie poursuivant son chemin à travers la cavité buccale).

Les réalisations du premier type sont dites orales, celles du second type nasales le schéma suivant explique les sons nasales .

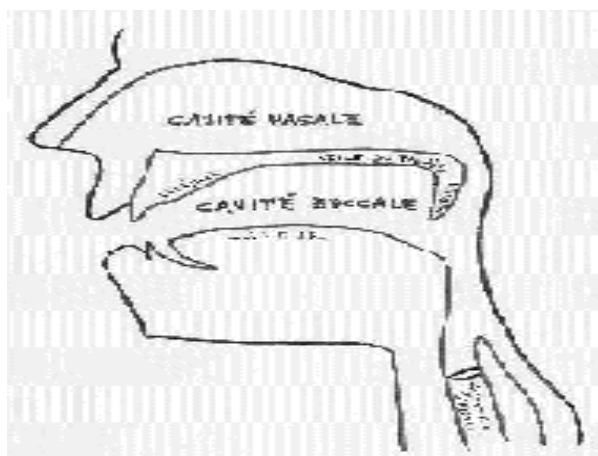


FIG. 4.3 – La voix nasale.

### 4.3 Étude de la voix des parkinsoniens

La base de données parkinsonien qui a été enregistrée par Mr TESTON comporte l'enregistrement de 200 personnes handicapées [GV00]. Chaque personne a enregistré le texte suivant extrait de “LA CHÈVRE DE MONSIEUR SEGUIN” qui est :

*M. Séguin n'avait jamais eu de bonheur avec ces chèvres. Il les perdait toutes de la même façon : un beau matin, elles cassaient leur corde, s'en allaient dans la montagne, et là haut le loup les mangeait. Ni les caresses de leurs maître, ni la peur du loup, rien ne les retenait. c'était paraît-il, des chèvres indépendantes, voulant à tout prix le grand air et la liberté. Le brave M. Séguin, qui ne comprenait rien au*

*caractère de ses bêtes, était consterné. Il disait :*

*C'est fini ; les chèvres s'ennuient chez moi, je ne garderai pas une. Cependant, il ne découragea pas, et, après avoir perdu six chèvres de la même manière, il en acheta une septième ; seulement, cette fois, il eut soin de la prendre toute jeune, pour qu'elle s'habitât à demeurer chez lui.*

*Ah ! Gringoire, qu'elle était jolie la petite chèvre de M. Séguin ! qu'elle était jolie avec ses yeux doux, sa barbiche de sous-officier, ses sabots noirs et luisants, ses cornes zébrées et ses long poils blancs qui lui faisaient une houppelande !.*

Pour cette étude nous allons nous baser sur une classification faite par des cliniciens allant de 0 à 3, 0 : la voix est presque normale, 1 : la voix indique qu'il y a des petites déformations au niveau de la production de la parole, 2 : la personne montre plus de handicaps au niveau de la production de la parole, 3 : la personne a eu de vrais handicaps articulatoires. Cette classification tient compte de plusieurs aspects cliniques et traitement du signal de la parole. Nous allons traiter l'aspect traitement du signal fréquentiel de la parole de ces personnes handicapées pour déterminer les effets de ces handicaps sur le traitement fréquentiel du signal.

Pour chaque niveau nous allons traiter la fréquence fondamentale sur un long fichier de parole, puis nous allons traiter la vitesse d'élocution, ensuite nous allons analyser le taux de silence dans la parole.

### 4.3.1 Analyse fréquentielle et temporelle

#### calcul de fréquentielle et temporelle

Pour le calcul de la fréquence fondamentale nous utilisons la méthode AMDF (Averaged Magnitude Difference Function Algorithm) qui est basé sur le calcul de la différence de l'amplitude du signal à différents instants.

On définit la distance  $D(k)$  avec ;

$$D(k) = \sum_{n=1}^{N-k} |x(n) - x(n+k)| \quad \text{avec} \quad k = 1, \dots, K$$

avec

$x(n)$  : le signal de parole à l'instant  $n$

$N$  : longueur de fenêtre

$K$  : coefficient glissant de la fenêtre.

La distance entre deux minimum successifs de  $D(k)$  constitue la période fondamentale  $T_0$ . Donc

$$F_0 = \frac{1}{T_0}$$

#### Le niveau 0

Pour ce niveau les personnes sont presque normaux parlants.

Nous allons analyser la variation de la fréquence fondamentale pour deux personnes de niveau 0.

La fréquence fondamentale est entre 70 et 250 Hz dans les zones voisées. Pour une personne normale la variation de la fréquence de parole ne dépasse pas les 10% [GV00].

Sur un fichier de parole pour une personne donnée, nous détectons les zones voisées, puis nous relevons la valeur moyenne, la valeur minimale, la valeur maximale et l'écart-type.

Nous donnons des exemples de chaque niveau de handicaps.

Le tableau suivant indique quelques statistiques sur la fréquence fondamentale sur un fichier d'une minute de parole pour la personne 1 :

	Fréquence
Moyenne	111 Hz
Ecart-type	11.6 Hz
Coeff. de variation	11%
valeur minimale	96 Hz
valeur maximale	127 Hz

TAB. 4.1 – Quelques statistiques sur la fréquence fondamentale pour la personne 1

Le coefficients de variation de la fréquence fondamentale est de même ordre pour les personnes normaux parlants, ce qui montre que la personne est légèrement handicapée.

La variation de la fréquence fondamentale dans les zones voisées dépend des sons voisés produits.

Pour la personne 2 de niveau 0 nous avons obtenu les résultats suivants :

	Fréquence
Moyenne	119 Hz
Ecart-type	13.3 Hz
Coeff. de variation	11%
valeur minimale	80 Hz
valeur maximale	189 Hz

TAB. 4.2 – Quelques statistiques sur la fréquence fondamentale pour la personne 2

Le coefficients de variation indique le niveau de perturbation pour une personnes handicapée. Ces types de personnes (de niveau 0) présente des handicaps pas profond qui ne peut pas les empêcher d'utiliser les systèmes classiques de reconnaissance de la parole.

La variation de la fréquence fondamentale est plus importante dans des zones ou la personne présente des difficultés pour la production de la parole.

La production de la parole continue nécessite du souffle (silence), ce silence dépend d'une personne à une autre. Ce silence peut agir sur les performances des systèmes de reconnaissance.

Nous avons mesuré le taux de silence dans les enregistrements de 20 personnes de niveau 0.

Le tableau suivant donne le taux de silence de quelques fichiers de niveau 0 :

Numéro de fichier	1	2	3	4	5	6	7	8
Pourcentage de silence (en %)	32	34	31	33	34	35	33	34

TAB. 4.3 – Mesure de taux de silences des fichiers enregistrés pour le niveau 0

Le tableau suivant donne la vitesse d'élocution en phonèmes par seconde de quelques fichiers de niveau 0 :

Nom de fichier	1	2	3	4	5	6	7	8
Vitesse moyenne d'élocution (en phonèmes/seconde)	25	20	18	22	24	24	20	25

TAB. 4.4 – Mesure de vitesse d'élocution des fichiers enregistrés pour le niveau 0

### Le niveau 1

Pour ce niveau les personnes ont eu un handicap plus important que celles de niveau 0.

Nous allons analyser la fréquence fondamentale pour deux personnes de ce niveau de handicap.

Le tableau suivant indique quelques statistiques sur la fréquence fondamentale sur le fichier de parole enregistrée pour la personne 1 :

	Fréquence
Moyenne	160 Hz
Ecart-type	21 Hz
Coeff. de variation	13.2%
valeur minimale	110 Hz
valeur maximale	190 Hz

TAB. 4.5 – Quelques statistiques sur la fréquence fondamentale pour la personne 1

Le coefficient de la variation de la fréquence fondamentale est plus important que celle de niveau 0, ces personnes présente des perturbation plus important au niveau des cordes vocales. Ces perturbation sont liés aux problèmes neurologique.

Pour la personne 2 nous avons obtenu les résultats suivants :

	Fréquence
Moyenne	174 Hz
Ecart-type	20 Hz
Coeff. de variation	11.5%
valeur minimale	102 Hz
valeur maximale	198 Hz

TAB. 4.6 – Quelques statistiques sur la fréquence fondamentale pour la personne 2

Nous donnons ici les taux de silences des enregistrements des personnes de niveau 1 :

Nom de fichier	1	2	3	4	5
Taux de silence (en %)	36	41	39	41	37

TAB. 4.7 – Taux de silence pour des fichiers enregistrés pour le niveau 1

Le tableau suivant donne la vitesse moyenne d'élocution des enregistrements des personnes de niveau 1 :

Nom de fichier	1	2	3	4	5
Vitesse d'élocution ( Phonèmes/seconde )	17	14	13	13	10

TAB. 4.8 – Vitesse d'élocution pour des fichiers enregistrés pour le niveau 1

## Le niveau 2

Pour ce niveau les personnes ont un handicap lourd mais la parole reste intelligible.

Le tableau suivant indique quelques statistiques sur la fréquence fondamentale sur un fichier d'une minute de parole pour une personne qui a un niveau de handicap 2 :

	Fréquence
Moyenne	113 Hz
Ecart-type	19 Hz
Coeff. de variation	17%
valeur minimale	90 Hz
valeur maximale	132 Hz

TAB. 4.9 – Quelques statistiques sur la fréquence fondamentale pour une personne de niveau 2 (exemple 1)

Le tableau suivant donne les taux de silence de quelques fichiers pour des personnes de niveau 2 :

Numéro de fichier	1	2	3	4	5	6
Taux de silence en (en %)	41	45	39	41	39	43

TAB. 4.10 – Taux de silence des fichiers enregistrés pour le niveau 2

Le tableau suivant donne la vitesse moyenne d'élocution des fichiers des personnes de niveau 2 :

Numéro de fichier	1	2	3	4	5	6
Vitesse moyenne d'élocution (en phonèmes/seconde)	11	8	12	14	14	13

TAB. 4.11 – Vitesse d'élocution des fichiers enregistrés pour le niveau 2

### Le niveau 3

Pour ce niveau les personnes ont un handicap plus lourd que le précédent, mais il reste toujours intelligible pour être traité.

Le tableau suivant indique quelques statistiques sur la fréquence fondamentale sur un fichier d'une minute et 30 secondes de parole pour une personne de niveau de handicap 3 :

	Fréquence
Moyenne	199 Hz
Ecart-type	52 Hz
Coeff. de variation	26%
valeur minimale	116 Hz
valeur maximale	230 Hz

TAB. 4.12 – Quelques statistiques sur la fréquence fondamentale pour une personne de niveau 3

Le tableau suivant donne les taux de silence de quelques fichiers pour des personnes de niveau 3 :

Numéro de fichier	1	2	3	4	5	6
Taux de silence en (en %)	39	49	46	44	43	44

TAB. 4.13 – Taux de silence des fichiers enregistrés pour le niveau 3

Le tableau suivant donne les vitesses moyennes d'élocution des fichiers des personnes de niveau 3 :

Numéro de fichier	1	2	3	4	5	6
Vitesse d'élocution (en phonèmes/seconde)	9	12	11	8	12	7

TAB. 4.14 – Vitesse moyenne d'élocution des fichiers enregistrés pour le niveau 3

### 4.3.2 Commentaires

Nous avons analysé pour chaque niveau les caractéristiques vocales de plusieurs personnes ( de 10 à 20 personnes ) pour résumer enfin les résultats obtenus dans le tableau suivant :

	Coeff de variation de F0	Vitesse d'élocution	Pourcentage de silence
Niveau 0	10-13%	17-26 phonèmes par seconde	28-36%
Niveau 1	11-14.5%	10-19 phonèmes par seconde	33-41%
Niveau 2	14-19%	8-15 phonèmes par seconde	38-45%
Niveau 3	20-29%	7-13 phonèmes par seconde	38-50%

TAB. 4.15 – Tableau récapitulatif

Les résultats obtenus nous permettent de vérifier les faits suivants [GV00] :

- 1 - Les personnes qui ont des handicaps plus lourds mettent plus de temps pour l'enregistrement, ce qui indique que la personne handicapée a plus de difficultés à enregistrer une base de données qu'une personne normale. Nous avons remarqué aussi que plus la personne est handicapée plus elle fait plus de pauses (silences) lors de l'enregistrement.
- 2 - La perturbation de la fréquence fondamentale est proportionnelle au degré du handicap.
- 3 - La vitesse d'élocution diminue avec le degré de le handicap.

## 4.4 Conclusion

La voix de la personne handicapée présente quelques déformations fréquentielles et temporelles par rapport à celle d'une personne normale. La variation de la fréquence fondamentale est plus importante pour les personnes handicapées. Une personne handicapée met plus de temps qu'une personne normale pour l'enregistrement d'une base de données.

Nous avons remarqué aussi que les personnes handicapées font plus d'hésitations que les personnes normales lors de l'enregistrement de la base de données.

Certaines personnes handicapées présentent des difficultés assez nettes pour la production de quelques phonèmes.

Dans le chapitre suivant nous allons exposer notre approche de système de reconnaissance de mots isolés avec apprentissage dynamique en utilisant la DTW.



## Chapitre 5

# Apprentissage dynamique

### 5.1 Introduction

Deux mots ou deux phrases ne peuvent pas être prononcés avec la même vitesse d'élocution par deux locuteurs différents ou même par le même locuteur. La comparaison de deux mots ne peut pas être faite point à point. La programmation dynamique fournit une solution de comparaison de deux occurrences pour le traitement de la parole [Bel57].

Les systèmes de reconnaissance de la parole qui utilisent la programmation dynamique nécessitent généralement une base de données d'apprentissage importante pour apprendre correctement les paramètres des unités de reconnaissance. Les performances de ces systèmes sont liées à la qualité de la construction de cette base de données qui représente une tâche lourde, coûteuse et difficile.

L'approche que nous proposons dans ce chapitre consiste à adapter le système de reconnaissance au fur et à mesure de son utilisation. Ainsi l'utilisateur n'est pas obligé d'enregistrer les données plusieurs fois. Durant l'utilisation du système de reconnaissance, si un mot est bien reconnu, il va s'ajouter à la base des références. Cette base s'enrichit automatiquement.

L'utilisation de la DTW (Dynamic Time Warping) comme méthode de reconnaissance s'impose pour deux raisons :

- La base de données de départ contient seulement une ou deux occurrences des commandes ce qui réduit considérablement la pénibilité pour un handicapé de devoir constituer une base de parole de longue durée.
- Les modèles HMM sont mal adaptés aux personnes ayant des difficultés articulatoires.

Dans ce qui suit nous décrivons d'abord un système de reconnaissance de mots isolés, puis nous détaillons notre approche, ensuite nous présentons la paramétrisation du signal de parole et la méthode de comparaison dynamique. Enfin nous analysons les résultats obtenus.

## 5.2 Description de la méthode

D'une part, les personnes handicapées ont du mal à enregistrer une importante base donnée, d'autre part les handicaps articulatoires diffèrent d'une personne à une autre. Ces facteurs rendent difficile voire même impossible la construction d'un système de reconnaissance de la parole de mots isolés pour des personnes handicapées. Une solution pour les systèmes de reconnaissance de la parole de mots isolés pour des personnes handicapées consiste à adapter le système de reconnaissance au fur et à mesure de son utilisation. un système de reconnaissance de la parole utilisant la DTW nécessite au moins une occurrence de chaque mot des références. De là une personne handicapée doit enregistrer au moins une occurrence de chaque mot de référence. Lors de l'utilisation du système, si un mot est bien reconnu, il va s'ajouter à la base des références. Pour ne pas avoir une grande base de références, qui peut encombrer le système, nous avons procédé à une décomposition temporelle.

Ainsi la méthode de reconnaissance est constituée des étapes suivantes :

- Enregistrement de la base de références avec une ou deux occurrences de phrases à reconnaître.
- Extractyion des caractéristiques du signal parole.
- Reconnaissance et ajout du signal reconnu à la base de données d'apprentissage si l'utilisateur est satisfait du résultat de reconnaissance.

La figure 5.1 présente le principe de cette méthode.

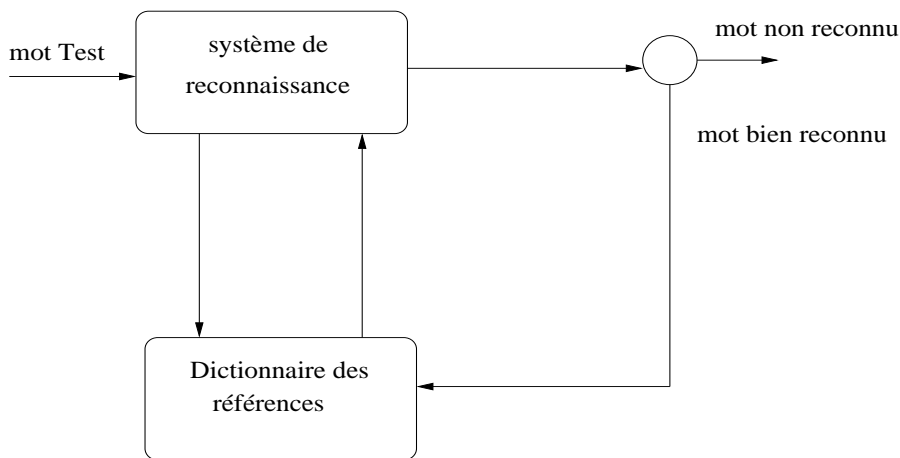


FIG. 5.1 – Mise à jour dynamique du dictionnaire des références

## 5.3 Paramétrisation du signal parole

Une fois numérisé, le signal parole subit une opération de pré-accentuation qui consiste en un filtrage de type passe-haut qui relève le niveau des aigus. En pratique, on utilise simplement un filtre de réponse impulsionnelle finie (1,a) avec  $a = -0.95$ . Si  $X(n)$  désigne le signal de parole et  $X_P(n)$  le signal pré-accentué on a :

$$X_P(n) = X(n) - 0.95 \times X(n - 1).$$

### 5.3.1 Fenêtrage et normalisation du signal

Le signal de parole n'est pas stationnaire, c'est pourquoi on applique un fenêtrage au signal en utilisant une fenêtre de Hamming de 20 ms avec un recouvrement de 10 ms. La fréquence d'échantillonnage est de  $F_e=16$  Khz. l'opération de fenêtrage se fait de la façon suivante :  $X_W(n) = X_P(n) \times W(n)$ .

Où  $W(n) = 0.54 - 0.46 \cos(2\pi n/(N-1))$

On utilise l'énergie du signal pour effectuer la normalisation pour permettre d'analyser le signal de parole indépendamment du niveau du signal.

L'énergie d'un signal X dans chaque fenêtre à normaliser est calculée avec la formule suivante :

$$E(fenetre) = \sum_{i=1}^N X(i)^2.$$

où N est la longueur de la fenêtre et  $X(i)$  la valeur du signal à l'instant i.

### 5.3.2 Coefficients MFCC

Les paramètres représentant le signal parole sont les coefficients MFCC (Mel-scaled Frequency Cepstral Coefficients). Après fenêtrage, nous effectuons sur chaque trame une transformée de Fourier discrète (TFD) :

$$Y(k) = \sum_{i=0}^{L-1} X_W(i) \exp(-2j\pi ki/L)$$

Avec  $L=1024$ .

Nous partitionnons ensuite l'échelle de fréquence entre 0 et  $F_e/2$  en 24 bandes correspondant dans la bande  $[0, 8 \text{ Khz}]$  à l'échelle logarithmique mel donnée par :

$$mel(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

Où f est la fréquence en Hz.

Nous sommions les valeurs de la transformée de Fourier discrète sur chacun des  $B=24$  bandes, en pratique les filtres sont triangulaires de largeur de bande constante et régulièrement espacés sur l'échelle Mel, exemple espacement 150 mels et largeurs 300 mels, Nous prenons le logarithme. Ce qui donne :

$$Y_m(p) = \log(\sum_{k \in B_p} (|Y(k)|)) \quad \text{pour } p \in 0, 1, \dots, B-1$$

Nous effectuons une transformée inverse en cosinus sur ces B valeurs :

$$z_m = C_B Y_m$$

où  $Y_m$  désigne le vecteur de composantes  $Y_m(p)$  et  $C_B$  la matrice d'éléments :

$$[C_B]_{k,n} = \sqrt{\frac{2}{B}} c(k) \cos\left(\frac{\pi k(n+0.5)}{B}\right) \quad \text{pour } 0 \leq k, \quad n \leq B-1$$

avec  $c(0) = \frac{1}{\sqrt{2}}$  et  $c(k)=1$  pour le reste.

Nous conservons pour chaque trame, les (D-1) premières valeurs de  $z_m$ , après les avoir centrées, ainsi que l'énergie de chaque trame.

Typiquement nous prenons  $D=13$  (12 premiers coefficients avec l'énergie). La figure 5.2 explique le principe de calcul des MFCC :



FIG. 5.2 – Calcul des coefficients MFCC

En considérant un mot d'une durée de 0.5s, on obtient une suite de 50 trames et chaque trame contient 13 coefficients, donc le mot est présenté par une matrice de dimension  $50 \times 13$ .

## 5.4 L'algorithme de comparaison dynamique

Compte tenu de la forte variabilité chez un handicapé ayant des difficultés articulatoires, les prononciations d'un même mot se réalisent acoustiquement et dans le temps de manière différente. Nous observons entre autres, une distorsion temporelle (les échelles temporelles de deux occurrences du même mot ne coïncident pas). Nous ne pouvons donc comparer point à point les formes acoustiques issues de la paramétrisation. Il est nécessaire de procéder à un alignement dit temporel.

Cette comparaison s'effectue par programmation dynamique. Elle est fondée sur les travaux de R. Bellman en 1957 [Bel57] pour la recherche de la trajectoire optimale. Étant données deux séquences acoustiques :

$$X = \{x_i, 1 \leq i \leq I\}$$

et

$$Y = \{y_j, 1 \leq j \leq J\}$$

de longueurs respectives I et J, la comparaison entre X et Y revient à rechercher le chemin optimal qui minimise la distance cumulée entre X et Y en respectant les contraintes suivantes :

- faire coïncider les extrémités : le chemin doit commencer en  $(x_1, y_1)$  et finir en  $(x_I, y_J)$ .
- respecter la continuité du chemin.

L'application de ces contraintes permet de respecter la dimension temporelle des signaux.

De plus s'ajoutent des contraintes locales dites de cheminement liées aux connaissances a priori sur le débit oral donné par la figure 5.3 :

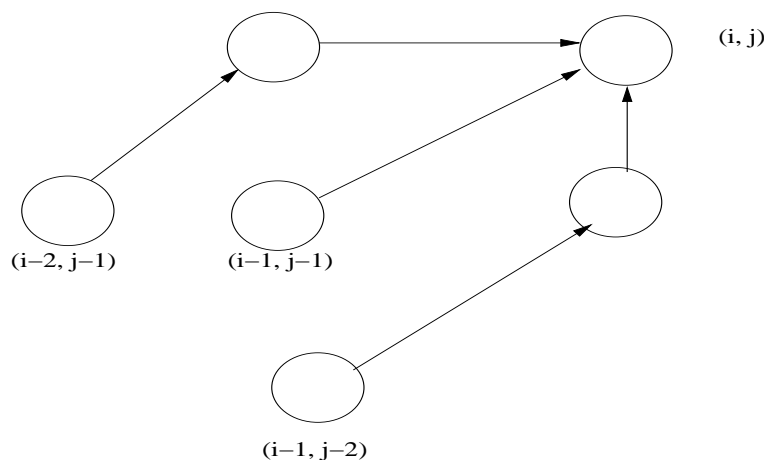
Le coût d'un chemin est obtenu par la formule récurrente de Sakoe et Shiba [SS78] appliquée ici à la contrainte représentée à la figure 5.3.

- Initialisation :

$$D(1, 1) = d(1, 1) \quad \text{et} \quad D(1, j) = D(j, 1) = \inf$$

- Récurrence :

$$D(i_x, j_y) = \min \begin{cases} D(i_x - 1, j_y - 1) + 2 \times d(i_x, j_y) \\ D(i_x - 1, j_y - 2) + d(i, j - 1) + d(i_x, j_y) \\ D(i_x - 2, j_y - 1) + d(i - 1, j) + d(i_x, j_y) \end{cases}$$

FIG. 5.3 – *contrainte de cheminement utilisée*

Avec  $d(i,j)$  la distance euclidienne entre le  $i$ ème vecteur de  $X$  et le  $j$ ème vecteur de  $Y$  et  $D(i,j)$  la distance cumulée.

- coût du chemin optimal :  $D$

$$D = \frac{D(I, J)}{I + J}$$

## 5.5 Résultats expérimentaux et discussions

Pour tester un système de reconnaissance de mots isolés fondé sur cette méthode, un ensemble de références est constitué lors de la phase d'apprentissage par une suite de vecteurs acoustiques  $X$ . Lors de la phase de reconnaissance, pour chaque mot à reconnaître  $Y$ , nous calculons toutes les distances entre l'observation  $Y$  et les références  $X$  par comparaison dynamique. Le mot reconnu est celui qui correspond à la référence pour laquelle la distance cumulée à  $Y$  est minimale.

### 5.5.1 Procédure expérimentale

Pour le système de reconnaissance utilisé, la base des références est constituée des commandes utilisées dans un environnement domestique.

Nous avons enregistré 40 commandes comme références (ouvrir la porte, fermer la porte, monter le rideau, allumer la télé etc...).

La méthode a été évaluée sur 5 phrases enregistrées 10 fois chacune (démarrer le lave-linge, arrêter le lave-linge, ouvrir la porte, fermer la porte et appeler la boîte vocale).

Les évaluations sont faites d'abord avec un seul enregistrement de référence puis avec la mise à jour dynamique du dictionnaire des références lors de la phase de reconnaissance.

### 5.5.2 Résultats et discussion

Les résultats obtenus sont expliqués de la façon suivante :

- Chaque phrase correspond à une lettre parmi les suivantes :

A : démarrer le lave-linge

B : arrêter le lave-linge

C : ouvrir la porte

D : ouvrir la fenêtre

E : appeler la boîte vocale

- Chaque évaluation correspond à une lettre aussi avec :

a : avec un seul enregistrement

b : avec mise à jour dynamique de 5 occurrences

c : avec mise à jour dynamique de 10 occurrences

- Les valeurs correspondent à un taux de reconnaissance.

Le système a été testé en monolocuteur.

Le tableau suivant récapitule les résultats obtenus :

	A	B	C	D	E	moyenne
a	50%	40%	90%	30%	60%	54%
b	60%	60%	100%	70%	60%	70%
c	90%	80%	100%	90%	80%	88%

TAB. 5.1 – Évolution de taux de reconnaissance au cours de l'apprentissage dynamique

Les performances des systèmes de reconnaissance de la parole dépendent de la base de données d'apprentissage. Dans notre approche nous nous intéressons à l'évolution des taux de reconnaissance en fonction de l'utilisation, ce qui revient à dire l'évolution du taux de reconnaissance avec la base de données d'apprentissage. Avec l'évolution du taux de reconnaissance de 54% avec un seul enregistrement à 70% après 5 utilisations, puis à 88% après 10 utilisations, notre système peut atteindre les performances des systèmes de reconnaissance de la parole classique rapidement. Cette solution est assez importante pour les personnes handicapées qui leur permet de construire une base de données d'apprentissage propre à eux en utilisant en même temps le système.

La base des références s'enrichit au fur et à mesure de l'utilisation du système ce qui augmente nécessairement les performances de système de reconnaissance.

### 5.5.3 Taux de confiance

Nous définissons le taux de confiance pour un mot test comme

$$\frac{D}{N} \times 100 \quad (5.1)$$

avec :

- D : distance minimale entre le mot test et les mots de références
- N : distance entre le mot test et le mot de référence qui lui correspond réellement.

Ce taux de confiance est de 100% en cas de bonne reconnaissance.

Il illustre les performances d'un système de reconnaissance. En fait autant ce taux de confiance est faible autant le dictionnaire des références est moins représentatif.

Le calcul de taux de confiance permet d'évaluer la qualité du dictionnaire des références.

Le tableau 5.2 illustre l'évolution de taux de confiance pour la phrase "ouvrir la fenêtre" (initiale : chaque mot du dictionnaire contient une seule occurrence, finale : après l'application de notre approche).

numéro du présentation de l'utilisation	Taux de confiance initial	Taux de confiance finale
1	86%	86%
2	100%	100%
3	100%	100%
4	95%	100%
5	82%	100%
6	100%	100%
7	89%	100%
8	97%	100%
9	62%	100%
10	90%	100%

TAB. 5.2 – Évolution de taux de confiance au cours de l'apprentissage dynamique pour la phrase D

L'évolution de taux de confiance n'est possible que si le système reconnaît bien une phrase test comme le montre l'exemple de la présentation N 1. Le taux de confiance initiale n'a pas changé parce que le dictionnaire des références n'a pas changé.

Le taux de confiance initiale augmente ou au pire des cas reste le même si on augmente le nombre des occurrences des bases des références.

Même si nous avons un taux de confiance initiale assez faible comme le montre la présentation N 9 de cette occurrence et avec l'augmentation des nombres des occurrences ( il passe d'une occurrence initialement à 8 occurrences ), le mot peut être bien reconnu par le système.

Le tableau 5.3 illustre l'évolution pour la phrase "arrêter le lave-linge" au cours de l'utilisation de notre système.

numéro du présentation de l'utilisation	Taux de confiance initial	Taux de confiance finale
1	100%	100%
2	74%	85%
3	80%	100%
4	85%	100%
5	100%	100%
6	100%	100%
7	69%	95%
8	81%	100%
9	87%	100%
10	70%	100%

TAB. 5.3 – Évolution de taux de confiance au cours de l'apprentissage dynamique pour la phrase B

La notion de taux de confiance permet de voir l'évolution du système de reconnaissance au niveau de l'approchement à la bonne reconnaissance. Les performances des systèmes de reconnaissance sont calculées à partir des phrases bien reconnues et ne tiennent pas compte des autres niveaux de reconnaissance.

Si par exemple pour un système de reconnaissance de phrases isolés le système propose les trois occurrences les plus proches et l'utilisateur peut choisir, les performances du système de reconnaissance doivent être meilleures, ce qui explique l'importance de taux de confiance

L'évolution de taux de confiance est due essentiellement à l'évolution du dictionnaire des références.

## 5.6 Conclusion

Les résultats obtenus pour un système de reconnaissance de la parole à apprentissage dynamique montrent que notre approche donne des résultats évolutifs au cours de l'utilisation du système en même temps que la base d'apprentissage augmente.

Cette approche présente une solution de construction d'une base de données pour les personnes handicapées qui représente généralement un handicap majeur pour la construction des systèmes de reconnaissance de la parole.

L'utilisation de la programmation dynamique est l'une des plus ancienne méthode de reconnaissance de la parole qui nécessite un calcul assez important. Dans le chapitre suivants nous allons exposer une méthode de reconnaissance qui utilise les modèles de Markov cachés en se basant sur une sgmentation automatique indépendante de la langue (ALISP (Automatic Langage Independant Speech Processing)).



# Chapitre 6

## Utilisation d'ALISP

### 6.1 Introduction

Les systèmes classiques de reconnaissance de la parole utilisent des unités linguistiques telles que phonèmes, diphtonges, mots etc... Ces systèmes nécessitent une segmentation linguistique de la base de données.

L'enregistrement d'une base de données d'apprentissage nécessite une étude préalable de sa constitution et la population nécessaire pour le faire. La segmentation de cette base de données est la tâche la plus difficile, en effet elle nécessite une connaissance phonétique. Même la segmentation automatique nécessite une intervention humaine. En plus de ce problème s'ajoutent les erreurs de segmentation. L'approche que nous proposons dans ce chapitre consiste à utiliser la correspondance entre une segmentation phonétique et une segmentation automatique (ALISP) [Cer98] afin de construire un système de reconnaissance de parole pour des personnes handicapées en utilisant la segmentation automatique de la base d'apprentissage.

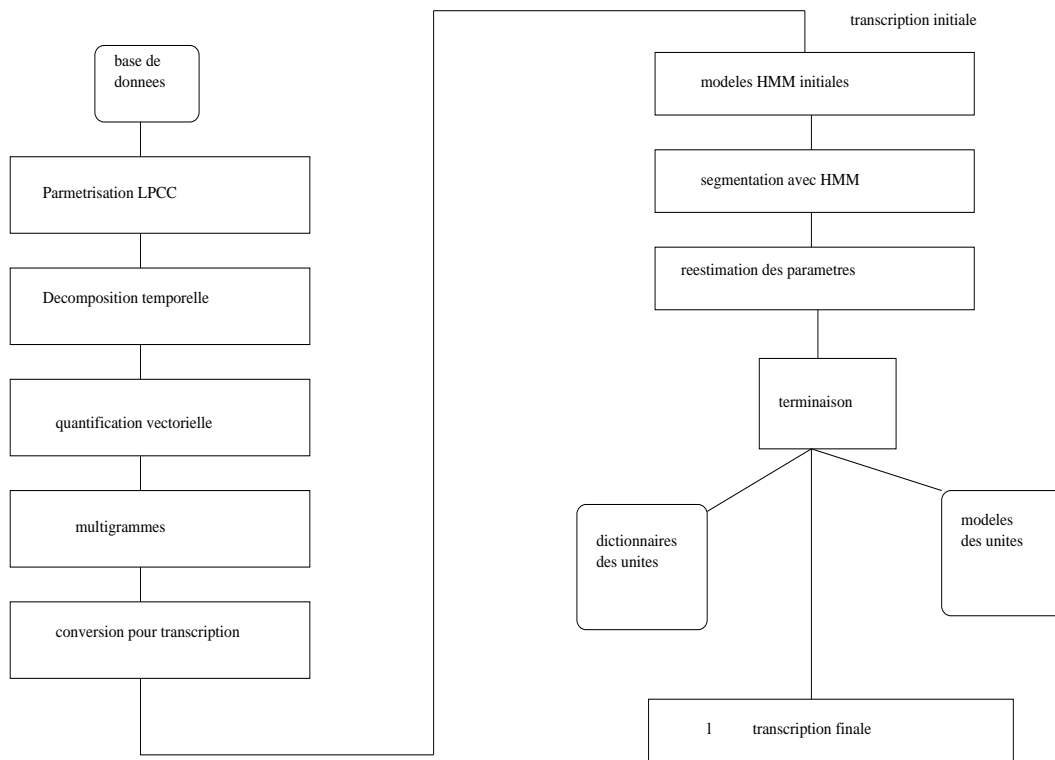
Le chapitre est structuré de la façon suivante : dans la section 2, nous présentons les étapes pour la segmentation ALISP. Dans la section 3 nous examinons la correspondance ALISP-phonétique. La section 4 est consacrée à la construction du système de reconnaissance de la parole.

### 6.2 Principe de la segmentation ALISP

La segmentation ALISP s'effectue en deux étapes [Cer98] :

1. Une segmentation initiale du corpus de parole en utilisant une décomposition temporelle et une quantification vectorielle.
2. Une segmentation statistique en utilisant des modèles de segments initiaux.

La figure 6.1 explique le principe de la segmentation ALISP :

FIG. 6.1 – *Principe de la segmentation ALISP*

### 6.2.1 Segmentation initiale

La segmentation initiale s'effectue en trois étapes :

1. Paramétrisation du signal parole.
2. Décomposition temporelle
3. Quantification vectorielle

1. Paramétrisation du signal de parole : A partir du signal de parole numérisé, nous allons extraire les coefficients LPCC (Linear Predictive Coefficient Cepstraux). Les étapes à suivre sont les suivantes :

- la pré-accentuation : une fois le signal numérisé, il subit un filtrage de type passe-haut qui relève le niveau des aigus :

$$\hat{S}(n) = s(n) - 0.95s(n-1)$$

- le fenêtrage : le signal pré-accentué est divisé en trames de N échantillons avec un recouvrement de N/2 échantillons. Généralement N=512(20 ms) avec une fréquence d'échantillonnage de 16 KHz.

$$s_w(n) = \hat{s}(n) \times (0.54 - 0.46\cos(2\pi n/(N-1)))$$

- l'analyse LPC : elle consiste à effectuer une modélisation AR (modélisation Auto-Régressive) du signal fenêtré. Nous cherchons à représenter le signal de parole sous

une forme AR :

$$s(n) = - \sum_{i=1}^P a(i)s(n-1) + e(n).$$

Avec  $P$  : l'ordre de filtre (qui est à l'ordre de 16).  $e(n)$  est un processus aléatoire (bruit blanc) qui prend en compte les erreurs du modèle

- la conversion des paramètres LPC en LPCC : les coefficients LPCC sont obtenus à partir des coefficients LPC de la façon suivante :

$$c_i = -a(i) - \sum_{n=1}^{i-1} -a(i-n)c_n$$

2. Décomposition Temporelle (DT) : nous appliquons la décomposition temporelle sur les vecteurs LPCC. Cette méthode utilise le critère de stabilité spectrale et les transitions entre ses vecteurs LPCC pour déterminer la segmentation de la parole. La décomposition temporelle approche une matrice de paramètres Cepstraux par des vecteurs cibles et des fonctions d'interpolation [DM88a]. La recherche des vecteurs cibles et des fonctions d'interpolation se fait par la décomposition en valeurs singulières à court terme d'une sous matrice  $Y$  de la matrice des coefficients Cepstraux  $X$  :

$$Y^t = U^t D V$$

Nous assemblons ensuite les lignes de la matrice  $U$  pour trouver une fonction d'interpolation concentrée sur une fenêtre rectangulaire. La ré-estimation de la fonction d'interpolation et l'adaptation de la fenêtre sont itérées pour affiner la recherche de la fonction d'interpolation [DM88a].

3. La quantification vectorielle : les segments trouvés après la DT subissent une classification non supervisée. Cette classification est faite en utilisant la méthode LBG(Lino Buzo Gray) avec des éclatements successifs du codebook : 1, 2, 4...64. L'ensemble d'apprentissage est constitué des vecteurs cepstraux originaux situés aux centres de gravité des fonctions d'interpolation.

## 6.2.2 Segmentation statistique

Les unités trouvées par la combinaison décomposition temporelle suivie d'une quantification vectorielle, sont modélisées par des modèles de Markov cachés. Cette étape contribue à un affinement des unités par des itérations de segmentation du corpus de parole et de ré-estimation des paramètres des modèles.

L'apprentissage des HMM se fait sur le même corpus que celui utilisé pour apprendre la décomposition temporelle et quantification vectorielle. L'initialisation des HMM prend en compte les transcriptions initiales obtenues par la décomposition temporelle et la quantification vectorielle. Nous reprenons ensuite, les étapes de segmentation à l'aide des modèles préalablement appris et de ré-estimation des paramètres de ces modèles.

## 6.3 Transcription phonétique d'une base de données

Quand nous faisons de la phonétique, il faut laisser de côté tout l'aspect orthographique de la langue. Ce n'est pas la forme orthographique qui prime sur la prononciation, mais plutôt le contraire. Par conséquent, il vaut mieux agir comme si nous ne savons pas écrire quand on fait de la phonétique.

Mais il faut quand même un mécanisme pour représenter les sons. L'alphabet normal convient assez mal à cette tâche, puisqu'une seule lettre peut correspondre à plus d'un son (t en français) et puisqu'un seul son peut se représenter au moyen de plus d'une lettre (son [s] en français).

Cette segmentation phonétique de la parole se fait par plusieurs méthodes, mais la méthode la plus utilisée est le TTS (Text To Speech). Cette segmentation utilise la parole et le texte dit par le locuteur pour faire l'alignement entre le texte et la parole à l'aide des modèles statistiques utilisés pour la reconnaissance de la parole. La méthode utilisée est automatique, mais elle nécessite une intervention humaine (un phonéticien) pour les corrections, et cette intervention est coûteuse.

### 6.3.1 Voix d'essai

Pour chaque locuteur nous avons besoin d'une base de données pour apprendre les modèles de phonèmes utilisés lors de la phase de l'alignement texte-parole. Cette base de données est dite base d'essai, qui est transcrite manuellement. Les étapes d'enregistrement de cette base d'essai sont les suivantes :

1- Mise en forme des données acoustiques : numérisation du signal, découpage des enregistrements et description phonétique des diphtongues.

2- Sélection du locuteur : sur une base d'un test d'écoute un expert fait la sélection du locuteur.

3-Enregistrement de la voix finale : le locuteur doit enregistrer la base d'essai dans des conditions idéales.

### 6.3.2 Mise en forme du signal acoustique

La segmentation phonétique est partiellement automatique, elle aligne la transcription phonétique et la forme acoustique du signal de parole, à l'aide des méthodes statistiques (HMM). Mais elle requiert les compétences du phonéticien à deux niveaux :

- pour bien fonctionner, le système a besoin de paramètres de description de tous les phonèmes de la langue (voyelles, semi-voyelles, consonne, plosive, fricative, nasale, liquide, voisé etc.).

- une écoute de l'ensemble des logatomes et une vérification manuelle sont nécessaires, pouvant conduire à la correction des unités.

## 6.4 Étude des correspondances entre une segmentation phonétique et une segmentation ALISP

### 6.4.1 Caractéristiques de la base de données

Cette base de données est construite à la société ELAN informatique. cette base de données est un extrait de journaux français, elle couvre 1060 diphtonges et 10000 triphthonges. Elle les caractéristiques suivants :

- Type de ressource : enregistrement de parole
- Mode de parole : Lu
- Conditions d'enregistrement : studio d'enregistrement professionnel
- Langue : Français
- Heures d'enregistrement : 9 heures
- Taille (en heure, vocabulaire) : 1 heure 30 minutes, 900 phrases
- Locuteur : une speakerine professionnelle
- Résolution : 16 bits
- Canal : mono
- Fréquence d'échantillonnage : 16 KHz

### 6.4.2 Correspondances ALISP-phonétique

Pour un corpus de parole nous allons comparer la segmentation acoustico-phonétique et celle obtenue par la segmentation ALISP. Nous définissons pour cela le recouvrement relatif que nous appelons  $R(A_{j,k}, P_i)$  comme le recouvrement relatif entre la kème occurrence de l'unité ALISP  $A_j$  et le phonème  $P_i$ . Pour cela nous définissons le recouvrement absolu  $RA(A_{j,k})$  qui est la longueur de la kème unité ALISP  $A_j$

Nous définissons le recouvrement entre l'unité ALISP  $A_j$  et le phonème  $p_i$ , par la quantité  $x_{j,i}$  telle que :

$$x_{j,i} = \sum_{k=1}^N \frac{R(A_{j,k}, P_i)}{RA(A_{j,k})}$$

Avec N est le nombre d'occurrences de l'unité ALISP  $A_j$

Nous avons  $\sum_{i=1}^I x_{j,i} = N$  avec I est le nombre total de phonèmes

Nous définissons  $T_{j,i}$  le taux de correspondance entre le phonème  $P_i$  et l'unité ALISP  $A_j$  telle que :

$$T_{j,i} = \frac{x_{j,i}}{N}$$

Nous avons  $\sum_{i=1}^I T_{i,j} = 1$ .

Le schéma 6.2 explique comment on détermine  $R(A_{j,k}, P_i)$  :

A1			A2		unites ALISP
R					
P1	P2				phonemes

FIG. 6.2 – *La correspondance ALISP-phonèmes*

La correspondance ALISP-phonétique permet de voir les éventuelles composantes phonétiques d'un segment ALISP qui nous conduit à faire une étude de la correspondance ALISP-polysons. Un polyson est une suite de phonèmes (diphone ou triphone).

### 6.4.3 Étude de la correspondance ALISP-polyson

Pour la même base de données nous examinons le contenu phonétique de chaque classe. Pour mieux étudier cette correspondance nous déterminons la probabilité de chaque correspondance d'un segment ALISP avec un polyson. Pour cela nous déterminons :

1. Tous les polysons qui composent un segment ALISP donné sur toute la base de données
2. La probabilité d'avoir chaque polyson sachant le segment ALISP donné sur toute la base de données.

#### Détermination des polysons d'un segment ALISP

Pour déterminer ces polysons, nous devons parcourir toute la base de données et chaque fois que nous trouvons le segment ALISP cherché, nous analysons le polyson qui lui correspond.

$$C_{A_{i,j}} = \sum_{k=\min_{A_{i,j}}}^{\max_{A_{i,j}}} P_K$$

Avec  $\min_{A_{i,j}}$  la borne inférieure de  $A_{i,j}$  et  $\max_{A_{i,j}}$  la borne supérieure de  $A_{i,j}$ .

Nous avons donc l'ensemble des polysons qui correspond a un segment ALISP  $A_i$  :

$$C_{A_i} = C_{A_{i,j}} \quad 1 \leq j \leq N$$

Avec N est le nombre de segment ALISP  $A_i$

A1			A2			segments ALISP
P1	P2	P3		P4	P5	Phonèmes

FIG. 6.3 – *La correspondance ALISP-polyson*

Par exemple le segment ALISP A1 correspond au polyson P1P2P3 et le segment ALISP A2 correspond au polyson P3P4P5.

## 6.5 Résultats expérimentaux et commentaires

Dans l'étude de la correspondance, nous développons deux approches : la première correspond à l'étude de cette correspondance en utilisant les classes classiques et la deuxième en utilisant les sous-classes.

### 6.5.1 Correspondance utilisant les classes ALISP

Pour notre étude nous avons utilisé la base de données ELAN, qui est une base de données monolucuteur et bien segmentée phonétiquement.

La notation des segments ALISP est présentée de la façon suivante : chaque segment est composé de la lettre H suivi de l'une des lettres A..Z, a..z ou des chiffres de 0..9 ou des symboles @ et \$.

L'attribution des noms des segments est faite au cours de la phase d'apprentissage. Au début, nous avons calculé le nombre de segments dans la base de données d'apprentissage (nous avons utilisé 1 heure pour l'apprentissage).

Le tableau suivant récapitule le nombre de chaque segment dans cette base de données :

HA	HB	HC	HD	HE	HF	HG	HH
844	894	749	993	956	1134	1629	969
HI	HJ	HK	HL	HM	HN	HO	HP
768	737	481	978	946	869	1565	562
HQ	HR	HS	HT	HU	HV	HW	HX
928	663	774	940	1143	1143	947	1068
HY	HZ	H0	H1	H2	H3	H4	H5
515	1351	1447	661	940	1090	525	807
H6	H7	H8	H9	Ha	Hb	Hc	Hd
532	764	981	854	1125	1519	1272	1081
He	Hf	Hg	Hh	Hi	Hj	Hk	Hl
728	953	882	1170	1194	821	958	520
Hm	HN	Ho	Hp	Hq	Hr	Hs	Ht
458	532	593	683	433	657	598	474
Hu	Hv	Hw	Hx	Hy	Hz	H@	Hl
1085	600	528	357	643	456	482	428

TAB. 6.1 – Nombre de segments ALISP dans la base de données ELAN

Pour une heure de parole nous avons eu 61000 segments ALISP. Nous pouvons dire alors que la longueur moyenne d'un segment ALISP est de 60 ms. Nous savons aussi que la longueur moyenne d'un phonème est de 30 ms, donc un segment ALISP correspond en moyenne à un diphone. Pour les expériences nous utilisons les classes ALISP comme références pour les correspondances.

### Correspondance ALISP-phonétique

Pour cette expérience, nous avons essayé de déterminer le taux de correspondance entre les phonèmes et les classes ALISP. Nous avons remarqué que cette correspondance est assez diverse parce qu'une classe ALISP peut correspondre jusqu'à 7 phonèmes voire beaucoup plus. La classe HE par exemple correspond aux phonèmes K, G, P, D et N, certes avec des proportions différentes par exemple HE correspond à 60% à K et à 12% à D etc..

Le tableau suivant présente quelques correspondances entre les phonèmes et les segments ALISP :

	A	#	M	E	N	C	ON	AN	autres
HA	0%	30%	0%	0%	5%	0%	0%	5%	60%
HJ	30%	0%	0%	10%	0%	10%	0%	0%	50%
HL	10%	10%	5%	0%	20%	0%	15%	10%	30%

TAB. 6.2 – Quelques correspondances entre les phonèmes et les segments ALISP

La correspondance ALISP-phonétique avec les classes n'a pas donné des résultats



bien précis qui risque d'alourdir les traitements du système de reconnaissance de la parole. En fait, la présence pour chaque segment ALISP de 4 à 8 même plus de correspondants phonétiques, rend les volumes de traitement assez importants ce qui agit aussi sur les scores de reconnaissance. Nous ajoutons aussi le fait qu'un segment ALISP est en moyenne deux fois plus long qu'un phonème donc un segment ALISP peut correspondre à une suite de deux ou trois phonèmes.

Donc nous avons opté pour la correspondance ALISP-Polyson.

### **Correspondance ALISP-polyson**

Nous avons utilisé la même base de données (ELAN). Nous cherchons pour chaque segment ALISP le polyson (suite de phonèmes) qui lui correspond.

Cette correspondance nous paraît plus pratique parce que la correspondance ALISP-phonétique ne permet pas de déterminer une correspondance exacte.

Chaque segment ALISP correspond à plusieurs polysons, mais en fait nous avons remarqué qu'un segment ALISP correspond généralement avec un pourcentage important à un polyson donné. Mais cette correspondance nous paraît pas précise.

La plupart des segments correspondent à plus de 10 polysons. Les résultats obtenus nous amènent à chercher d'autres solutions dont l'utilisation des sous classes

#### **6.5.2 Correspondance utilisant les sous-classes**

Pour avoir des résultats plus précis, nous avons procédé à l'utilisation des sous classes. Pour ALISP nous avons 64 classes soit  $64 \times 64$  sous classes c'est à dire 3096 sous classes. Nous définissons une sous classe A1A2 comme la classe A1 suivie de la classe A2. La détermination des sous classes est faite automatiquement pendant la reconnaissance. La figure 6.4 explique le principe des sous-classes :

	sous-classe A1A2	sous-classe A2A3	
A1	A2	A3	A4

FIG. 6.4 – *Définition des sous-classes*

La correspondance ALISP-polyson nous permettra de construire notre système de reconnaissance. Nous avons analysé cette correspondance sur la même base de données ELAN. Nous avons remarqué qu'avec cette base de données, 40% des sous-classes sont vides.

Nous commençons par une analyse générale sur toutes les sous-classes puis nous donnons en annexe quelques exemples de classes.

Les classes correspondent à 8 jusqu'à 18 polysons (diphones et triphones)

90% des sous-classes correspondent à 2 jusqu'à 4 polysons.

10% des sous-classes correspondent à 5 jusqu'à 7 polysons.

90% des sous-classes correspondent à des diphones.

7% des sous-classes correspondent à des triphones.

Ces résultats montrent que les résultats sont plus précis avec les sous-classes.

## 6.6 Utilisation de la segmentation ALISP pour la reconnaissance de la parole

L'avantage principal de la segmentation ALISP est qu'elle est indépendante de la langue. Cet avantage nous permet de l'utiliser pour n'importe quelle personne. En fait, l'utilisation de cette segmentation nous permet d'utiliser une base de données existante comme base d'apprentissage.

Nous avons utilisé la segmentation ALISP pour les systèmes de reconnaissance de mots isolés et la correspondance ALISP-polyson pour le système de reconnaissance de la parole continue.

### 6.6.1 Utilisation de la segmentation ALISP pour la reconnaissance des mots isolés

Pour ce système de reconnaissance de mots isolés, nous avons utilisé une grammaire de quelques mots, chaque mot est transcrit en segments ALISP.

La transcription des mots de la base des références est faite automatiquement en utilisant les modèles de segments ALISP. L'utilisateur est censé enregistrer une fois

tous les mots du dictionnaire qui nous permettent d'avoir une transcription ALISP des bases de références.

Nous donnons un exemple simple de reconnaissance de 5 villes de France par exemple (Paris, Marseille, Bordeaux, Lyon et Lille).

L'utilisateur doit enregistrer les noms des cinq villes. Nous aurons donc un dictionnaire formé comme suit :

Paris HaHbH1

Marseille H1HCH@H0

Bordeaux H2HcHM

Lyon HKHnH9

Lille H8HJH1

La grammaire est la suivante :

\$ville = Paris | Marseille | Lyon | Bordeaux | Lille ;

((#) \$ville (#))

Nous devons avoir alors des modèles de segments ALISP et en utilisant la grammaire et le dictionnaire nous construisons le système de reconnaissance de mots isolés.

Ce système de reconnaissance de mots isolés, utilise les modèles statistiques et donne des résultats qui dépassent les 80%. Nous avons remarqué aussi que le système proposé dans le chapitre précédent est plus lourd du point de vu calcul.

Nous avons mené des expériences avec une petite base de données de personnes handicapées de 4 phrases pour chacune de deux personnes.

Pour la personne 1, nous avons pris les phrases suivantes :

Ouvrir la porte

Fermer la fenêtre

Allumer la télé

Éteindre la télé

Nous avons procédé à segmenter une seule occurrence de chaque phrase en segments ALISP pour définir notre dictionnaire. Nous avons défini une grammaire de 12 phrases comme suit :

\$verbe = ouvrir | fermer | éteindre | allumer

\$complément = la porte | la fenêtre | la télé

Phrase = ((#)\$verbe \$complément(#))

Les résultats de reconnaissance sont résumés dans le tableau suivant :

Phrase	Taux de reconnaissance
phrase1	80%
phrase2	90%
phrase3	100%
phrase4	70%

TAB. 6.3 – Taux de reconnaissance pour le locuteur 1

Pour la personne 2, nous avons pris les phrases suivantes :

Ouvrir la porte

Fermer la fenêtre

Ouvrir la fenêtre

Fermer la porte

Nous avons procédé de la même manière que précédemment.

Les résultats de reconnaissance sont résumés dans le tableau suivant :

Phrase	Taux de reconnaissance
phrase1	100%
phrase2	100%
phrase3	100%
phrase4	90%

TAB. 6.4 – Taux de reconnaissance pour le locuteur 2

Ces résultats sont certes intéressants mais reste à vérifier sur une base de données plus importante. Les résultats obtenus sont proches de l'état de l'art pour les personnes normales.

Le problème de cette méthode est dans la transcription des mots en segments ALISP qui n'est pas universelle comme la transcription phonétique.

Cette méthode a l'avantage d'ignorer les problèmes articulatoires lors de la production des phonèmes par les personnes handicapées.

Pour un système de reconnaissance de parole continue, le dictionnaire contient des milliers. L'utilisation de cette approche pour la parole continue est impossible d'où l'utilisation de la correspondance ALISP-phonétique.

### 6.6.2 Utilisation de la correspondance ALISP-polyson pour la reconnaissance de la parole continue

Pour un système de reconnaissance de la parole continue, la grammaire utilise un très grand nombre de mots pour en construire une qui dépasse les 20 000 mots pour un système dit de reconnaissance de parole continue.

La méthode précédente nous ne permet pas de construire une telle grammaire pour un système de reconnaissance de la parole continue. Nous utilisons la correspondance ALISP-polyson pour construire un tel système.

Nous définissons la même grammaire phonétique que celle pour un système classique de reconnaissance de la parole continue. A ce système nous ajoutons une couche intermédiaire qui utilise la reconnaissance de la parole sous forme d'une suite de segments ALISP. Pour donner le résultat final nous utilisons la correspondance ALISP-polyson et la grammaire.

Afin de construire un système de reconnaissance de parole continue basé sur la correspondance ALISP-phonétique nous procédons comme suit :

- Définir la grammaire du système de reconnaissance
- Déterminer la correspondance ALISP-phonétique
- Faire la reconnaissance en utilisant les modèles des segments ALISP

- Faire la correspondance pour déterminer les possibilités de la transcription phonétique de la parole test.
- Utiliser la grammaire pour déterminer le résultat final de reconnaissance.

La mise en oeuvre de cette méthode nécessite une grammaire assez significative et un dictionnaire assez important. Nous expliquons cette méthode théoriquement.

L'étude précédente sur la correspondance ALISP-phonétique nous a permis de constater que pour la correspondance des sous classes ALISP avec les polysons nous donne des correspondances assez précises qui nous permettent de les exploiter lors de la reconnaissance.

### **Étapes de la reconnaissance**

Pour construire le système de reconnaissance il faut passer par trois étapes :

- 1- Reconnaissance de la parole test sous forme de suite de segments ALISP : cette étape revient à utiliser les modèles des segments ALISP pour trouver une séquence de segments ALISP du fichier test.
- 2- Détermination de toutes les correspondances ALISP-polyson pour les segments trouvés.

La figure 6.5 explique cette étape : avec 1, 2 et 3 sont les segments ALISP du mot

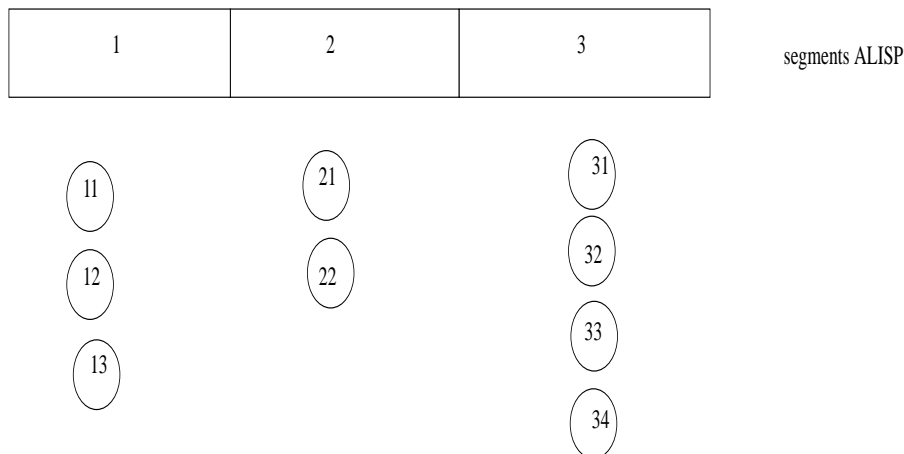


FIG. 6.5 – *Détermination de la correspondance ALISP-polyson pour la reconnaissance*

test et 11, 12, 13, 21,22, 31, 32, 33 et 34 leurs correspondants successifs en polysons.

3- La reconnaissance phonétique : cette reconnaissance utilise le dictionnaire et le modèle de langage et l'étape précédente pour déterminer le résultat de reconnaissance. Cette étape comporte 2 sous étapes :

- Utilisation de l'étape précédente et le dictionnaire pour déterminer les possibilités de suite de mots comme le montre la figure 6.6 :

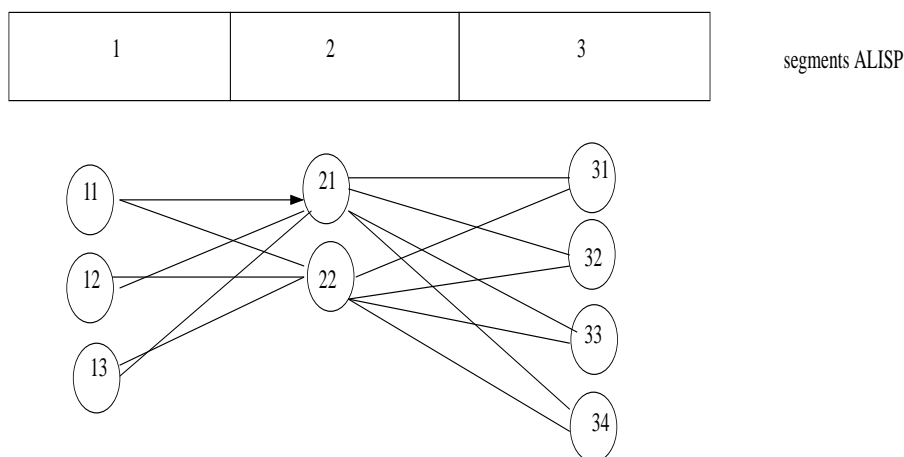


FIG. 6.6 – *Détermination hypothèses de mots du dictionnaire*

- Utilisation de la grammaire pour déterminer le résultat final de la reconnaissance. La figure suivante explique cette sous étape :

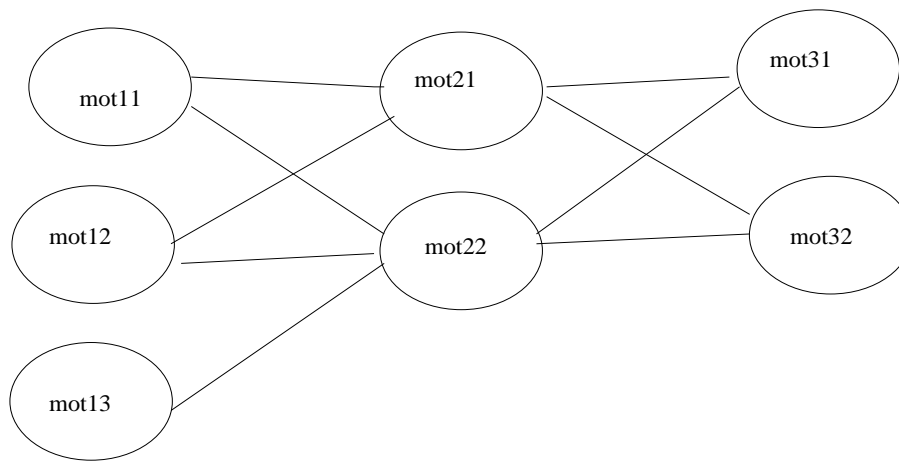


FIG. 6.7 – *Détermination du résultat final de la reconnaissance*

Le système proposé est un système de parole continue classique mais avec une étape cachée qui est l'utilisation de la correspondance ALISP-phonétique.

## 6.7 Conclusion

Une segmentation automatique indépendante de la langue peut servir pour avoir un système de reconnaissance de mots isolés, et l'utilisation d'un système plus compliqué permet d'avoir un système de reconnaissance de parole continue.

Pour le système de reconnaissance de mots isolés, notre approche nous permet d'utiliser un système statistique qui est plus facile à implémenter. Ce système donne des résultats plus performants sans avoir des modèles propres aux personnes handicapées. C'est que la segmentation utilisée est une segmentation indépendante de langue et des difficultés articulatoires chez une personne handicapée.

Pour le système de reconnaissance de parole continue, l'approche proposée peut donner des résultats satisfaisants, mais qui n'atteignent pas les performances des systèmes classiques. Deux solutions sont possibles :

- 1- La mise à jour de cette correspondance au cours de l'utilisation du système de reconnaissance.
- 2- Pour une base de données, nous segmentons phonétiquement une partie de la base de données pour étudier la correspondance ALISP-polysons et nous utilisons toute la base de données pour l'apprentissage des segments ALISP.

Dans le chapitre suivant nous allons exposer l'approche utilisation des modèles statistiques avec adaptation dynamique pour la reconnaissance de parole handicapée.





# Chapitre 7

## Adaptation dynamique

### 7.1 Modèles de Markov cachés

Depuis leur introduction en traitement de la parole [Bak75], 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 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 [Rab89a].

#### 7.1.1 Modélisation du signal parole

Nous expliquons cette étape par un exemple de reconnaissance de mots isolés.

Dans le but de réaliser une reconnaissance de mots isolés, on suppose qu'une machine de Markov modélise un mot du vocabulaire. Dans un cas plus général, ces modèles de mots sont eux-mêmes construits par concaténation d'unités acoustiques de base (phonèmes). Les états peuvent être interprétés comme des configurations de l'appareil phonatoire, et les observations émises lors de l'arrivée dans un état correspondent aux trames acoustiques. Ces trames sont usuellement représentées par des vecteurs de paramètres continus. Il est nécessaire soit de se ramener au cas de l'émission de symboles discrets par quantification vectorielle, soit de modéliser les probabilités d'émission par des densités de probabilité continues [Bar96].

#### Modèles de mots

Les modèles de Markov utilisés pour représenter la parole sont des modèles "gauche-droite", c'est-à-dire qu'il n'existe pas de cycle dans le graphe orienté engendré par les transitions entre états. Ceci traduit la causalité du processus de production de la parole. Leurs probabilités de transition vérifient :

$$a_{ij} = 0 \quad i \leq j$$

Des modèles gauche-droite ont été introduit par R. Bakis qui autorisent le bouclage de l'état courant, le passage à l'état suivant ou le saut d'un état. Le nombre d'états d'un modèle de mot est généralement proportionnel à la durée moyenne de ce mot. Les observations émises lors des transitions représentent la succession des trames acoustiques au cours de la prononciation du mot. Ces observations dans un espace généralement continu peuvent être décrites par un nombre fini de symboles au moyen de la quantification vectorielle.

### Observations continues

Le principe de l'émission de symboles discrets peut se généraliser au cas continu. Les probabilités d'émission discrètes  $b_j(k)$  sont alors remplacées par des densités de probabilité continues dans l'espace de représentation. On utilise une combinaison linéaire de gaussiennes :

$$b_j(o) = \sum_{k=1}^R g_k N(o, \mu_k, \sigma_k) \quad (7.1)$$

Avec  $\mu_k$  la moyenne de la gaussienne,  $\sigma_k$  est la matrice de covariance de la gaussienne et  $g_k$  est la pondération affectée à cette gaussienne.

La densité de probabilité d'une loi normale de moyenne  $\mu$  et de covariance  $\sigma$  est :

$$N(o, \mu, \sigma) = \frac{1}{(2\pi)^{d/2} \sigma^{1/2}} \exp\left(-\frac{1}{2}(o - \mu)^t \sigma^{-1} (o - \mu)\right) \quad (7.2)$$

L'hypothèse d'une indépendance entre les  $d$  dimensions de l'espace autorise l'utilisation de matrices de covariance diagonales. Cela limite le nombre de paramètres à estimer et simplifie les calculs.

### Flux de données

Les vecteurs de paramètres extraits du signal de parole contiennent souvent des données hétérogènes : les coefficients cepstraux, l'énergie, les coefficients différentiels... Il est possible de découper les vecteurs d'observation en plusieurs groupes de coefficients appelés des flux de données et de modéliser ces flux comme des observations indépendantes. La probabilité d'émission pour l'observation complète  $o_t = (o_t^1 \dots o_t^f)$  composée de  $f$  flux est le produit de la probabilité d'émission pour chacun des flux :

$$b_j(o_t) = \prod_{k=1}^f b_j^k(o_t^k) \quad (7.3)$$

Le principe est applicable pour des observations discrètes et continues. K.F. Lee utilise par exemple des modèles discrets et fabrique un premier dictionnaire pour les coefficients cepstraux, un deuxième pour les coefficients différentiels et un troisième pour l'énergie [Lee88]. Dans le cas d'observations continues où les densités sont représentées par une gaussienne à matrice de covariance diagonale, tous les coefficients sont indépendants, ce qui revient formellement à définir un flux par coefficient.

### 7.1.2 Reconnaissance d'un modèle

Connaissant une suite d'observations, la règle de décision bayésienne désigne le modèle qui les a émis. Il est nécessaire pour cela de calculer la probabilité d'émission de la suite des observations par chaque modèle. L'évaluation de cette probabilité est compliquée par le fait que le chemin parcouru n'est pas connu, mais un algorithme récursif efficace peut être utilisé. De plus, certaines variables introduites ici servent pour l'apprentissage des modèles.

#### Décision Bayésienne

Soit  $O = (o_1 \dots o_T)$  des trames acoustiques d'un mot test. On cherche le mot  $\hat{m}$  parmi les mots de dictionnaire le plus probable.

$$\hat{m} = \arg \max_{m \in E_m} P(m|O) = \arg \max_{m \in E_m} P(O|m)P(m) \quad (7.4)$$

Chaque mot  $m$  est modélisé par un modèle  $\lambda_m$ . Donc

$$P(O|m) = P(O|\lambda_m) \quad (7.5)$$

Donc on peut écrire

$$\hat{m} = \arg \max_{m \in E_m} P(O|\lambda_m)P(m) \quad (7.6)$$

La résolution de cette équation nécessite l'estimation de la probabilité  $P(O|\lambda_m)$  pour chaque mot de dictionnaire.

#### Probabilité d'observation

La probabilité que la suite d'observations  $O = (o_1 \dots o_T)$  n'est pas directement calculable, car le chemin emprunté est a priori inconnu. Mais elle peut être reformulée comme la somme des probabilités conjointes de l'observation  $O$  et du chemin  $q$  pour l'ensemble  $Q$  de tous les chemins possibles de longueur  $T$  :

$$P(O|\lambda) = \sum_{q \in Q} P(O, q|\lambda) \quad (7.7)$$

Soit  $a_{q_{t-1}q_t}$  la probabilité de passer de  $q_{t-1}$  à  $q_t$ .

On pose que  $q = (q_1 \dots q_T)$  donc

$$P(q|\lambda) = \prod_{t=1}^T a_{q_{t-1}q_t} \quad (7.8)$$

et la probabilité d'avoir émis les observations  $O$  en suivant ce chemin est :

$$P(O|q, \lambda) = \prod_{t=1}^T b_{q_t}(o_t) \quad (7.9)$$

Donc la probabilité conjointe de chemin et des observations

$$P(O, q | \lambda) = P(q | \lambda) \times P(O | q, \lambda) \quad (7.10)$$

$$= \prod_{t=1}^T a_{q_{t-1}q_t} \times b_{q_t}(o_t) \quad (7.11)$$

Donc

$$P(O | \lambda) = \sum_{q \in Q} \prod_{t=1}^T a_{q_{t-1}q_t} \times b_{q_t}(o_t) \quad (7.12)$$

Un algorithme rapide, dit "Forward" ou avant (estimation directe) permet de calculer récursivement cette quantité en maîtrisant l'explosion combinatoire

### Calcul Forward

Une variable intermédiaire notée  $\alpha_t(i)$  est introduite pour le calcul de la probabilité d'émission. C'est la probabilité que les observations jusqu'à l'instant  $t$  aient été émises par le modèle  $\lambda$  à  $N$  états, et que l'état à cet instant soit l'état  $i$  :

$$\alpha_t(i) = P(o_1 \dots o_t, q_t = s_i | \lambda) \quad (7.13)$$

$$\alpha_{t+1}(j) = P(o_1 \dots o_{t+1}, q_{t+1} = s_j) \quad (7.14)$$

.

$$\begin{aligned} P(o_1 \dots o_{t+1}, q_{t+1} = s_j) &= \sum_k P(o_1 \dots o_t, o_{t+1}, q_t = s_k, q_{t+1} = s_j) \\ &= \sum_k P(o_1 \dots o_t, q_t = s_k) \times P(o_{t+1}, q_{t+1} = s_j | o_1 \dots o_t q_t = s_k) \\ &= \sum_k P(o_1 \dots o_t, q_t = s_k) \times P(o_{t+1}, q_{t+1} = s_j | q_t = s_k) \\ &= \sum_k P(o_1 \dots o_t, q_t = s_k) \times P(o_{t+1} | q_{t+1} = s_j, q_t = s_k) \\ &\quad \times P(q_{t+1} = s_j | q_t = s_k) \\ &= \sum_k \alpha_t(k) \times a_{jk} \times b_j(o_{t+1}) \end{aligned}$$

Donc

$$\alpha_{t+1}(j) = \left( \sum_k \alpha_t(k) \times a_{jk} \right) \times b_j(o_{t+1}) \quad (7.15)$$

Donc l'algorithme Forward est le suivant :

Initialisation :

$$\alpha_1(i) = \pi_i \times b_i(o_1)$$

Induction :

$$\alpha_{t+1}(j) = [\sum_{i=1}^N \alpha_t(i) \times a_{ij}] \times b_j(o_{t+1}), \quad \text{pour} \quad 1 \leq t \leq T \quad \text{et} \quad 1 \leq j \leq N$$

Terminaison :

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i).$$

### Calcul Backward

Soit  $\beta_t(j)$  la probabilité que les observations après l'instant  $t$  soient émises en partant de l'état  $i$  :

$$\beta_t(i) = P(o_{t+1} \dots o_T | q_t = s_i, \lambda) \quad (7.16)$$

Donc

$$\beta_{t+1}(j) = P(o_{t+2} \dots o_T | q_{t+1} = s_j)$$

$$\begin{aligned} \beta_t(i) &= \sum_k P(o_{t+1}, o_{t+2} \dots o_T | q_t = s_i, q_{t+1} = s_k) \\ &= \sum_k P(o_{t+1}, o_{t+2} \dots o_T | q_{t+1} = s_k) \times P(q_{t+1} = s_k | q_t = s_i) \\ &= \sum_k P(o_{t+2} \dots o_T | q_{t+1} = s_k) \times P(o_{t+1} | q_{t+1} = s_k) \times P(q_{t+1} = s_k | q_t = s_i) \end{aligned}$$

Donc

$$\beta_t(j) = \sum_{k=1}^N a_{jk} \times b_j(o_{t+1}) \times \beta_{t+1}(k) \quad (7.17)$$

Donc l'algorithme backward est :

Initialisation :

$$\beta_T = 1$$

Induction :

$$\beta_t(i) = \sum_{j=1}^N a_{ij} \times b_j(o_{t+1}) \times \beta_{t+1}(j), \quad \text{pour} \quad 1 \leq t \leq T-1 \text{ et } 1 \leq i \leq N$$

Terminaison

$$P(O|\lambda) = \beta_0(1)$$

### 7.1.3 Recherche des états cachés

La procédure Forward-Backward fournit la probabilité d'émission des observations cumulées sur toutes les séquences d'états possibles, sans choisir un chemin particulier. Il est parfois utile de connaître la séquence d'états qui a émis les observations. L'algorithme de Viterbi cherche la séquence d'états cachés la plus probable et calcule la probabilité d'émission le long de ce chemin. La probabilité ainsi estimée néglige les chemins moins probables, et une reconnaissance à partir de cette probabilité est sous-optimale par rapport à l'estimation Forward-Backward. Mais la procédure fournit une segmentation du signal qui peut être exploitée pour l'apprentissage initial des modèles ainsi que pour le décodage de la parole continue.

### Chemin optimal

En vu des observations  $O$  émises par le modèle  $\lambda$ , la séquence d'états  $q$  la plus probable ayant pu émettre ces observations est donnée par l'équation 7.20 :

$$\hat{q} = \arg \max_{q \in Q} P(q|O, \lambda) \quad (7.18)$$

La règle de Bayes nous permet d'écrire :

$$\hat{q} = \arg \max_{q \in Q} P(q, O|\lambda) \quad (7.19)$$

Donc

$$P(O, \hat{q}|\lambda) = \max_{q \in Q} \prod_{t=1}^T a_{q_{t-1}q_t} \times b_{q_t}(o_t) \quad (7.20)$$

L'algorithme de Viterbi qui est une application de la programmation dynamique comme l'algorithme de DTW , résout ce problème de manière récursive.

### Algorithme de Viterbi

Suivant un critère d'optimisation globale : on cherche la séquence la plus probable c'est à dire  $P(O, q | \lambda) = P(q | O, \lambda) \times P(q | \lambda)$ .

Pour cela, nous définissons la probabilité maximale à un instant  $t$  qui se termine dans l'état  $s_i$ .

$$\delta_t(i) = \max_{q_1 \dots q_{t-1}} P(q_1 q_2 \dots q_t = s_i, o_1 \dots o_t | \lambda)$$

par récurrence  $\delta_{t+1}(i)$  s'écrit suivant l'équation 7.23 :

$$\delta_{t+1}(i) = [\max_j \delta_t(j) \times a_{ij}] \times b_j(o_{t+1}) \quad (7.21)$$

On conserve le max de la fonction donnée par (6.25) à chaque instant  $t$  et chaque état  $i$ .

Algorithme de Viterbi

Initialisation :

$$\delta_1(i) = \phi_i b_i(o_1) \text{ et } \phi_1(i) = 0$$

Récursion :

$$\delta_t(i) = \max_{1 \leq j \leq N} [\delta_{t-1}(j) \times a_{ij}] \times b_j(o_t) \quad 2 \leq t \leq T$$

$$\phi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) \times a_{ij}] \quad 1 \leq j \leq T$$

Fin :

$$\hat{p} = \max_{1 \leq i \leq N} \delta_T(i)$$

$$\hat{q}_T = \arg \max_{1 \leq i \leq N} \delta_T(i)$$

Meilleurs séquences :

$$\hat{q}_t = \phi_{t+1}(\hat{q}_{t+1}), \quad t = T-1, T-2 \dots 1;$$

### 7.1.4 Apprentissage d'un modèle

La reconnaissance d'un mot prononcé est rendue possible par l'évaluation de la probabilité d'émission des observations par tous les modèles de mots. Cela suppose l'existence d'un modèle au moins pour chaque mot, éventuellement construit par concaténation de modèles acoustiques plus courts, et l'apprentissage des paramètres de ces modèles. Ces paramètres sont les probabilités de transition entre états et les probabilités d'émission associées aux états, car la topologie du modèle, à savoir le nombre d'états des modèles, les transitions autorisées entre ces états et l'alphabet des symboles émis, sont supposés fixée à priori. Ainsi, connaissant une suite d'observations émises par un modèle, il est possible de modifier les paramètres du modèle de manière à rendre plus probable l'émission des observations par le modèle. Il s'agit d'une estimation sur le critère du maximum de vraisemblance (Maximum Likelihood Estimation ou MLE) qui est réalisée par l'algorithme de Baum-Welch. D'autres critères d'apprentissage existent, comme les critères MAP (Maximum A Posteriori) ou MMI (Maximum Mutual Information), mais leur mise en oeuvre est généralement plus difficile.

#### Maximum de vraisemblance

L'estimation par maximum de vraisemblance consiste à choisir les paramètres du modèle  $\lambda$  afin de rendre maximale la probabilité d'émission des observations  $O$  par le modèle :

$$\hat{\lambda} = \arg \max_{\lambda} P(O|\lambda) \quad (7.22)$$

Une résolution analytique directe n'est pas possible, mais les formules de Baum-Welch permettent une ré-estimation itérative des paramètres  $a_{ij}$  et  $b_j(k)$  du modèle en appliquant ce critère [Baum, 1972]. A la suite de la ré-estimation des paramètres du modèle  $\hat{\lambda}_n$ , le nouveau modèle  $\hat{\lambda}_{n+1}$  vérifie :

$$P(O|\hat{\lambda}_{n+1}) \geq P(O|\hat{\lambda}_n) \quad (7.23)$$

La convergence vers un optimum local est démontrée, mais les valeurs initiales des paramètres A et B sont cruciales pour assurer une convergence correcte et rapide le plus près possible du maximum global. L'algorithme de Viterbi réalisant le décodage peut servir à l'initialisation des modèles.

#### Ré-estimation des modèles

Pour estimer les paramètres on utilise une procédure itérative qui fournit un maximum locale de  $P(O|\lambda)$  : l'algorithme EM.

Soit  $\xi_t(i, j) = P(q_t = s_i, q_{t+1} = s_j | O, \lambda)$  la probabilité de transiter de  $i$  vers  $j$  sachant l'observation  $O$  et le modèle.

nous avons :

$$\begin{aligned}\xi_t(i, j) &= \frac{P(q_t = s_i, q_{t+1} = s_j, O|\lambda)}{P(O|\lambda)} \\ &= \frac{\alpha_t(i) \times a_{ij} b_j(o_{t+1}) \times \beta_{t+1}(j)}{\sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) \times a_{ij} \times b_j(o_{t+1}) \times \beta_{t+1}(j)}.\end{aligned}$$

Notons :

$$\gamma_t(i) = P(q_t = s_i | O, \lambda), \quad d'ou : \gamma_t(i) = \sum_{j=1}^N \xi_t(i, j).$$

$\sum_{t=1}^{T-1} \gamma_t(i)$  = l'espérance de la probabilité de transition depuis  $s_i$  sachant  $O$  et le modèle.

$\sum_{t=1}^{T-1} \xi_t(i, j)$  = le nombre de transition depuis  $s_i$  vers  $s_j$  sachant  $O$  et le modèle.

Donc nous pouvons calculer les paramètres du modèle de la façon suivante :

$\bar{\pi}_i$  = nombre de fois au temps 1 où on est en  $s_i = \delta_1(i)$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (7.24)$$

$$\bar{b}_j(k) = \frac{\sum_{t=1}^T o_t^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \quad (7.25)$$

Algorithme EM :

Dans les modèles de Markov on définit la fonction auxiliaire  $A$  :

$$A(\lambda, \lambda') = \sum_{q \in Q} P(O, q | \lambda') \times \log[P(O, q | \lambda)].$$

et

$$\hat{\lambda} = \arg \max_{\lambda \in Q} E_{\lambda}(\log P(O, q | \lambda')) \quad (7.26)$$

Posons  $q = q_1 \dots q_T$  et  $O = o_1 \dots o_T$ .

$$P(O, q | \lambda) = \pi_{q_1} \times b_{q_1} \times \prod_{t=2}^T a_{q_{t-1} q_t} \times b_{q_t}(o_t).$$

Donc

$$\begin{aligned}A(\lambda, \lambda') &= \sum_{q \in Q} [\log(\pi_{q_1}) + \log(b_1(o_1)P(O, q | \lambda'))] + \\ &\quad \sum_{q \in Q} [\sum_{t=2}^T \log(a_{q_{t-1} q_t})] P(O, q | \lambda') + \sum_{q \in Q} [\sum_{t=2}^T \log(b_{q_t}(o_t))] P(O, q | \lambda')\end{aligned}$$

Pour connaître les valeurs des probabilités initiales, il suffit de dériver la fonction  $A(\lambda, \lambda')$  par rapport à chaque  $\pi_i$ .



Nous avons la contrainte  $\sum_{j=1}^N \pi_j = 1$  que l'on peut intégrer dans la maximisation en introduisant le multiplicateur de Lagrange ( $\mu$ ).

$$\frac{\delta}{\delta \pi_i} \left( \sum_{i=1}^N \log(\pi_i P(O, q_1 = i | \lambda)) - \mu \left( \sum_{j=1}^N \pi_j - 1 \right) \right)$$

Donc

$$\frac{P(O, q_1 = i | \lambda)}{\pi_i} - \mu = 0, \quad i \in [1, N]$$

Soit

$$\pi_i = \frac{P(O, q_1 = i | \lambda)}{\mu}$$

Or

$$\sum_{i=1}^N \pi_i = 1 \rightarrow \mu = \sum_{i=1}^N P(O, q_1 = i | \lambda).$$

d'où

$$\pi_i = \frac{P(O, q_1 = i | \lambda)}{\sum_i P(O, q_1 = i | \lambda)} = \gamma_1(i).$$

## 7.2 Adaptation des HMM

Pour les systèmes de reconnaissance de la parole indépendant du locuteur, la modélisation de la variabilité interlocuteurs nécessite une large population de locuteurs. Cette stratégie d'apprentissage entraîne une variance dans les modèles acoustiques, réduisant les capacités des systèmes à différencier les phonèmes. Pour augmenter les performances de ces systèmes de reconnaissance, on emploie la technique d'adaptation pour permettre de mieux prendre en compte les caractéristiques acoustiques de nouveau locuteur où les nouvelles caractéristiques du nouvel environnement.

### 7.2.1 Adaptation de la moyenne (MLLR : Processus d'adaptation de régression linéaire de maximum de vraisemblance)

MLLR [DOM01] est une méthode d'adaptation qui permet de modifier les moyennes des mélanges de gaussiennes afin d'être plus proche de l'espace acoustique du nouveau locuteur. Cette méthode d'adaptation ne nécessite pas une grande base d'apprentissage (adaptation à posteriori non supervisée).

Dans la technique MLLR, on cherche à estimer une transformation  $W$  linéaire qui modélise les différences entre les conditions d'apprentissages et les conditions de tests. Le critère d'estimation est le maximum de vraisemblance. Le système adapté est obtenu en appliquant cette transformation à un ensemble de paramètres de modèles d'apprentissage.

En MLLR nous supposons que la moyenne initiale est transformée en moyenne adaptée de la façon suivante :

$$\hat{\mu} = W\mu$$

### Calcul en données complètes

Dans ce cas on suppose que la séquence d'observations et la séquence d'états associés sont connues.

soit  $O = o_1 \dots o_T$  : suite d'observations de test.

$q = q_1 \dots q_T$  : états cachés qui peuvent expliquer O

Nous avons

$$P(O, q) = P(O|q) \times P(q) = p(q_1 = i) \times p(o_1|q_1 = i) \times \prod_{t=1}^T \left[ \sum_{j=1}^K P(o_t, j) \mathbb{1}(q_t = j) \right] \times P(q)$$

$$\begin{aligned} \log(P(O, q)) &= \sum_{t=1}^T \log \left[ \sum_{j=1}^K P(o_t, j) \times \mathbb{1}(q_t = j) \right] + \sum_{t=1}^T \log(P(q_t)) \\ &+ \log(p(q_1 = i)) + \log(p(o_1|q_1 = i)). \\ &= \sum_{t=1}^T \left[ \sum_{j=1}^K \log(b_j(o_t)) \right] + \sum_{t=1}^T \log(P(q_t)). \end{aligned}$$

Nous avons

$$b_j(o_t) = \frac{1}{(2\pi)^{d/2} (\det(\Sigma))^{1/2}} \exp(-1/2 \times (o_t - \mu_j)^t \Sigma^{-1} \times (o_t - \mu_j)) \quad (7.27)$$

Donc

$$\log(b_j(o_t)) = -1/2 \times \log(\det(\Sigma)) - 1/2 \times (o_t - \mu_j)^t \Sigma^{-1} \times (o_t - \mu_j) \times \log((2\pi)^{d/2}).$$

Donc

$$\begin{aligned} \frac{\delta(\log(b_j(o_t)))}{\delta \mu_j} &= (o_t - \mu_j)^t \Sigma^{-1}. \\ \frac{\delta(P(O, q))}{\delta \mu_j} &= \sum_{t=1}^T (o_t - \mu_j)^t \Sigma^{-1} \mathbb{1}(q_t = j). \\ \hat{\mu}_j &= \frac{1}{N_j} \sum_{t=1}^T o_t \mathbb{1}(q_t = j | \lambda). \end{aligned}$$

avec  $N_j$  : nombre de fois où apparaît l'état j.

Donc

$$\begin{aligned} \log(b_j(o_t)) &= -1/2 \times \log(\det(\Sigma)) - 1/2 \times (o_t - W\mu_j)^t \times \Sigma^{-1} \times (o_t - W\mu_j) \\ &= -1/2 \times \log(\det(\Sigma)) - 1/2 \times [(o_t)^t \Sigma^{-1} (o_t) - (o_t)^t W\mu_j - (W\mu_j)^t \Sigma^{-1} o_t \\ &+ (W\mu_j)^t \Sigma^{-1} W\mu_j] \end{aligned}$$

Donc

$$\frac{\delta(\log(b_j(o_t)))}{\delta W} = -\Sigma_j^{-1} o_t \mu_j^t + \Sigma_j^{-1} W \mu_j \mu_j \mu_j^t$$

Donc

$$\log P(O, q) = \sum_{t=1}^T \sum_{j=1}^K \log b_j(o_t) \mathbb{1}(q_t = j)$$

Donc

$$\frac{\delta(\log P(O, q))}{\delta W} = \sum_{t=1}^T \sum_{j=1}^K (-\Sigma_j^{-1} o_t \mu_j^t + \Sigma_j^{-1} W \mu_j \mu_j \mu_j^t) \mathbb{1}(q_t = j) \quad (7.28)$$

Pour trouver la valeur souhaitée de W il suffit de résoudre l'équation

$$\frac{\delta(\log P(O, q))}{\delta W} = 0$$

### Calcul en données incomplètes

Dans ce cas on ne connaît que les séquences d'observations. Pour utiliser les résultats obtenus avec des données complètes, il faut effectuer une "statistiques" combinatoire sur toute l'ensemble des séquences d'états cachés possibles associé à chaque observation. On note  $E^p(\mathbb{1}(q_t = j))$  l'espérance à posteriori de la fonction indicatrice. Donc on a :

$$E(\mathbb{1}(q_t = j)) = E(\mathbb{1}(q_t = j) | o_t) = P(q_t = j | o_t)$$

Donc

$$\begin{aligned} E(\log(P(O, q))) &= \sum_{t=1}^T \left[ \sum_{j=1}^K \log(b_j(o_t)) \right] P(q_t | o_t) \\ &= \sum_{t=1}^T \left[ \sum_{j=1}^K \left( \frac{1}{2} \log(\det(\Sigma)) - \frac{1}{2} (o_t - \mu_j)^t \Sigma^{-1} (o_t - \mu_j) \right) \times P(q_t | o_t) \right] \end{aligned}$$

de la même manière qu'avec les données complètes nous trouvons que

$$\frac{\delta(E(\log P(O, q)))}{\delta W} = \sum_{t=1}^T \sum_{j=1}^K (-\Sigma_j^{-1} o_t \mu_j^t + \Sigma_j^{-1} W \mu_j \mu_j \mu_j^t) P(q_t | o_t) \quad (7.29)$$

Pour trouver la valeur souhaitée de W il suffit de résoudre l'équation

$$\frac{\delta(\log(E(P(O, q))))}{\delta W} = 0$$

### 7.2.2 Adaptation de la covariance

Pour l'adaptation de la covariance nous allons chercher à estimer une transformation qui modélise la transformation de la covariance initiale  $C_m$  en covariance adaptée  $\hat{C}_m$ . Dans notre cas, on suppose que la matrice de covariance est diagonale [Mok01].

La matrice de covariance peut se décomposer sous la forme :

$$C_m = C_m^{1/2} \times C_m^{1/2}$$

pour la transformation linéaire, on suppose que

$$\hat{C}_m = C_m^{1/2} \times H_m \times C_m^{1/2}$$

avec  $H_m$  la matrice de transformation des covariances à estimer à partir des données d'adaptation. Cette matrice est diagonale pour maintenir la diagonalité de la matrice de covariance.

Soit  $O_k$  la suite des observations des données d'adaptation avec  $1 \leq k \leq N$  et  $\lambda_m$  les paramètres du modèle associé à  $m$ . Dans la MLLR la matrice  $H_m$  est estimée en se basant sur le maximum de vraisemblance tel que :

$$H_m = \arg \max_H \prod_k P(O_k | \lambda_m)$$

La matrice d'adaptation est estimée selon la règle de Bayes :

$$H_m = \arg \max_H \prod_k P(O_m^k | \lambda_m) \times P(H)$$

Où  $P(H)$  est la probabilité à priori de  $H$ . l'utilisation de  $P(H)$  à priori nous permet d'incorporer la connaissance antérieure dans le processus d'adaptation de la covariance par une statistique bayésienne. Mais la solution MAP(Maximum A Posteriori) est préférable

Dans les systèmes de reconnaissance de la parole les matrices de covariances sont diagonales avec  $C_m = \text{diag}[\sigma_1^2, \dots, \sigma_D^2]$  et nous considérons que  $H_m = \text{diag}[H_1^2, \dots, H_D^2]$ , dont la distribution a priori est  $P(H)$ .

On suppose que  $P(H)$  est une loi de probabilité normale avec un vecteur moyen  $\mu_H = [\mu_H(1), \dots, \mu_H(D)]^t$  et une variance  $C_H = [\sigma_H^2(1), \dots, \sigma_H^2(D)]$ .

On définit la fonction

$$Q(\lambda, \hat{\lambda}) = -\frac{1}{2} \sum_t \sum_n \sum_m \gamma_t(n, m) [\log |\hat{C}_m| + (O_t - \mu_m)^t \hat{C}_m^{-1} (O_t - \mu_m)] + \log(P(H))$$

avec  $\gamma_t(n, m)$  est la probabilité d'observer  $O_t$  à l'instant  $n$  et au mélange  $m$ . Nous avons donc

$$\log |\hat{C}_m| + (O_t - \mu_m)^t \hat{C}_m^{-1} (O_t - \mu_m) = \sum_{i=1}^D [\log(\sigma_m^{(i)}) + \log(H_i^2) + H_i^{-2} (O_t^{(i)} - \mu_m^{(i)})^2]$$

Donc

$$Q(\lambda, \hat{\lambda}) = \quad (7.30)$$

$$-\frac{1}{2} \sum_t \sum_n \sum_m \gamma_t(n, m) \left[ \sum_{i=1}^D [\log(\sigma_m^{(i)}) + \log(H_i^2) + H_i^{-2} (O_t^{(i)} - \mu_m^{(i)})^2] - \right.$$

$$\left. \frac{1}{2} \log(2\pi\sigma_H^2(i)) - \frac{H_i - \mu_H(i)}{2\sigma_H^2(i)} \right].$$

Pour trouver la valeur optimal de H il suffit de résoudre l'équation suivante :

$$\frac{\delta(Q(\lambda, \hat{\lambda}))}{\delta(H_i)} = 0$$

Ce qui donne :

$$\sum_t \sum_n \sum_m \gamma_t(n, m) \left[ -H_i^{-1} + H_i^{-3} \frac{(O_t^{(i)} - \mu_m^{(i)})^2}{\sigma_m^{(i)}} \right] - \frac{H_i}{\sigma_H^2(i)} + \frac{\mu_H(i)}{\sigma_H^2(i)} = 0 \quad (7.31)$$

## 7.3 Reconnaissance de la parole continue

Les modèles de Markov cachés permettent une formalisation globale de toute la chaîne de reconnaissance. Leur attrait principal provient de leur simplicité structurelle et du cadre théorique rigoureux qui assure la convergence des modèles vers une solution acceptable lorsqu'une quantité suffisante de données d'apprentissage est disponible. Il existe cependant un grand nombre de paramètres qui doivent être choisis par le concepteur d'un système de reconnaissance de la parole continue et dont l'impact sur les performances du système est essentiel :

- l'espace de représentation du signal et les pré-traitements associés ;
- les unités acoustiques modélisées ;
- la structure des machines modélisant ces unités ;
- le réseau syntaxique et le modèle de langage. Pour les systèmes de reconnaissance de parole continue à grand vocabulaire, il n'est pas réaliste d'apprendre un modèle pour chaque mot. L'utilisation d'unités acoustiques sub-lexicales est nécessaire. De plus, la démarche proposée pour l'identification de mots isolés qui consiste à tester chacun des mots possibles n'est plus adaptée à la reconnaissance de la parole continue. En effet, le nombre de phrases possibles par enchaînement de mots et donc le nombre de modèles à évaluer est virtuellement infini. La reconnaissance est donc réalisée avec une variante de l'algorithme de Viterbi déjà présenté pour le décodage, en respectant la syntaxe définie par un réseau de mots. Enfin, la procédure d'apprentissage est elle aussi modifiée pour permettre l'apprentissage des modèles sur des phrases non segmentées.

### 7.3.1 Modèles acoustiques

L'unité acoustique modélisée, le nombre d'états du modèle, les transitions autorisées entre les états et le type de densité de probabilité associé à un état ou à

une transition doivent être choisis par le concepteur du système. En effet, les procédures d'apprentissage des modèles portent uniquement sur la valeur des probabilités d'émission et de transition. L'ensemble de ces choix est le plus souvent guidé par une approche heuristique, même si quelques procédures automatiques ont été proposées pour résoudre certains problèmes.

### Unités sub-lexicales

La modélisation par mots est très efficace pour des applications de reconnaissance de mots isolés ou en vocabulaire limité [Rab89a]. Pour la reconnaissance de parole continue à grand vocabulaire, il est impossible d'apprendre un modèle pour chacun des mots du dictionnaire. L'utilisation des unités plus courtes est nécessaire. Ces unités peuvent être des phonèmes indépendants du contexte, des modèles phonétiques dépendants du contexte, des diphones, des syllabes, ou encore des unités acoustiques sans signification phonologique comme les fénones. L'unité la plus utilisée est la suivante :

- Le phonème semble un choix naturel pour l'unité de base, car il a l'avantage de ne nécessiter qu'un nombre très réduit de modèles, au plus quelques dizaines. Alors que la variabilité de la coarticulation n'est pas prise en compte, et les systèmes de reconnaissance utilisant des unités phonétiques sont de ce fait moins performants que les systèmes utilisant des modèles de mots.

### Modélisation phonétique

Pour la modélisation d'une unité acoustique courte comme un phonème ou un de ses allophones, le nombre d'états est nécessairement réduit. Un seul état suffirait à modéliser un son stationnaire. En raison des phénomènes de coarticulation, le début et la fin d'une réalisation phonétique peuvent présenter des caractéristiques acoustiques différentes du centre supposé plus stationnaire. Le choix le plus simple est un modèle gauche-droite à trois états émetteurs, mais rien n'oblige à priori tous les phonèmes à être modélisés de la même façon. La modélisation des probabilités d'émission est aussi l'occasion de choix souvent arbitraires. Les densités de probabilité continues sont usuellement préférables à une quantification vectorielle des paramètres. Une combinaison linéaire de gaussiennes à matrice de covariance diagonale est le plus souvent retenue pour modéliser la distribution des paramètres. Encore faut-il vérifier la pertinence de cette modélisation pour les paramètres utilisés et choisir un nombre adapté de gaussiennes.

### 7.3.2 Modèles de langage

La modélisation acoustique permet à elle seule de réaliser la transcription phonétique d'une phrase. En absence de contraintes d'ordre linguistique, qu'elles soient lexicales, syntaxiques ou sémantiques, il est cependant probable que la suite de phonèmes ainsi obtenue n'aura qu'un lointain rapport avec la chaîne attendue. Cela est dû en partie à la faiblesse des systèmes actuels, mais aussi à la difficulté de l'identification phonétique dans la parole continue en raison des phénomènes de

coarticulation. L'auditeur humain exploite les niveaux supérieurs pour interpréter l'ambiguïté intrinsèque de nombreux contextes phonétiques, et même compenser une dégradation limitée de l'information acoustique. De plus, une reconnaissance acoustique même parfaite ne suffit pas pour obtenir une transcription correcte de la phrase. J. Mariani rapporte qu'une suite particulière de 9 phonèmes peut être transcrite en Français en 32000 suites de mots différentes orthographiquement correctes; quelques unes seulement sont des phrases syntaxiquement correctes. Il est donc indispensable, autant que possible, d'introduire dans les systèmes de RAP des connaissances sur les niveaux supérieurs du langage.

### Modèles probabilistes

Dans le cadre markovien, l'utilisation de modèles de langages probabilistes est naturelle. La probabilité d'une suite de  $n$  mots est exprimée comme le produit des probabilités conditionnelles d'un mot sachant tous les mots précédents :

$$P(m_1 \dots m_n) = P(m_1) \prod_{i=2}^n P(m_i | m_1 \dots m_{i-1}) \quad (7.32)$$

En pratique, la manipulation des probabilités correspondant à toutes les suites de mots de longueur quelconque est impossible. Il est nécessaire d'apporter des simplifications à ce modèle. Les modèles  $k$ -grammes de langage sont ainsi caractérisés par le fait que la probabilité d'apparition d'un mot (ou plus généralement d'un symbole de ce langage) ne dépend que des  $k-1$  mots précédents :

$$P(m_i | m_1 \dots m_{i-1}) \simeq P(m_i | m_1 \dots m_{i-k+1}) \quad (7.33)$$

### 7.3.3 Apprentissage en parole continue

La ré-estimation d'un modèle par l'algorithme de Baum-Welch optimise la probabilité d'émission d'un ensemble de séquences acoustiques par un modèle. Lorsque les modèles représentent des unités sub-lexicales, il n'est pas possible de prononcer ces unités de manière isolée, et il faut les isoler par une segmentation des phrases prononcées. Or, la segmentation manuelle est un travail fastidieux qui doit être réalisé par un expert. On ne peut espérer obtenir de cette manière de très nombreuses prononciations d'une unité nécessaire à l'estimation robuste du modèle. Une ré-estimation des modèles avec des phrases non segmentées est heureusement possible en utilisant les modèles connectés.

Connaissant la transcription phonétique d'une phrase d'apprentissage, une machine composite est fabriquée en mettant les modèles correspondants en séquence, il n'est pas donc nécessaire de disposer d'une segmentation de la phrase d'apprentissage ce qui permet l'exploitation de grandes bases de données. la transcription phonétique manuelle d'une phrase est plus facile à réaliser que sa segmentation. Il est même possible d'utiliser une simple transcription lexicale de la phrase, et de constituer le modèle de la phrase par concaténation de modèles de mots, eux-mêmes décrits par des modèles phonétiques.

La ré-estimation des modèles connectés pour chaque phrase se fait de la façon suivante :

- Construction d'un modèle de la phrase avec des modèles acoustiques
- Estimation du modèle sur la phrase par l'algorithme de Baum-Welch
- Mise à jour des paramètres de tous les modèles. De manière assez similaire, la construction d'un modèle composite pouvant produire toutes les phrases possibles permet le décodage de la parole continue.

### 7.3.4 Reconnaissance de la parole continue

En reconnaissance de la parole continue la suite de mots recherchée est celle qui maximise l'équation suivante :

$$\hat{m} = \arg \max_{\hat{m} \in E_m} P(O|m)P(m) \quad (7.34)$$

Pour résoudre cette équation, il n'est pas possible de construire un modèle pour chacune des phrases pouvant être prononcées puis de comparer tous ces modèles avec la phrase à identifier. L'alternative consiste à construire un modèle unique pouvant émettre toutes les phrases syntaxiquement correctes du langage. Le décodage de la phrase prononcée est déduit du meilleur chemin dans ce modèle obtenu par une variante de l'algorithme de Viterbi. La modélisation linguistique peut être intégrée au processus de décodage ou ajoutée en post-traitement. Le modèle unique utilisé en reconnaissance de la parole continue est un réseau des modèles correspondant aux mots du vocabulaire. Ce réseau permet l'enchaînement des mots en respectant la syntaxe du langage, ce qui correspond à un modèle de langage où toutes les phrases syntaxiquement correctes sont à priori équiprobables. Dans le cas le plus simple, n'importe quelle suite de mots est autorisée, et le modèle composite est constitué de modèles de mots mis en parallèle, avec un bouclage de l'état final des modèles vers l'état initial de n'importe quel autre modèle pour permettre l'enchaînement de plusieurs mots. Les modèles de mots peuvent eux-mêmes être des modèles composites fabriqués à partir de modèles sub-lexicaux. Il peut être pratique, pour la fabrication du réseau, de rajouter certains états qui n'émettent pas d'observations, mais servent uniquement à simplifier la représentation des transitions d'un modèle à un autre. La suite d'états optimale trouvée par l'algorithme de Viterbi dans ce réseau fournit un décodage de la phrase en mots ou en phonèmes, ainsi qu'une segmentation du signal acoustique. Cependant, l'algorithme recherche la meilleure suite d'états dans le réseau et non pas la meilleure suite de modèles.

### Intégration de modèle de langage

Un modèle de langage simple, de type bigramme, s'intègre facilement au décodage par l'algorithme de Viterbi. En effet, la transition entre deux modèles, au lieu d'être équiprobable, est affectée de la probabilité du bigramme. Dans le cas d'une tâche de décodage acoustico-phonétique, si  $\lambda_i$  et  $\lambda_j$  sont les modèles de phonèmes  $\phi_i$  et  $\phi_j$  la probabilité de passer à l'intérieur du modèle composite de  $\lambda_{i-1}$  à  $\lambda_j$  est



celle estimée par un modèle de bigramme phonétique :

$$\hat{P}(\lambda_j|\lambda_i) = \frac{f(\phi_i, \phi_j)}{f(\phi_i)}. \quad (7.35)$$

Le décodage recherche alors le chemin optimal, en maximisant la probabilité conjointe des observations acoustiques et du modèle de langage plutôt que la seule probabilité acoustique. Lorsque le modèle de langage est plus complexe, son intégration au module acoustique devient très coûteuse.

### Taux de reconnaissance

Le décodage d'une phrase fournit une suite d'unités acoustiques. Pour évaluer la qualité du décodage et quantifier les différences avec le résultat idéalement attendu, il faut disposer d'un étiquetage de référence des phrases de test. Les deux chaînes de mots ou de symboles sont alignées par un algorithme de programmation dynamique, tel que l'algorithme de Wagner et Fisher [Wagner et Fisher, 1974]. A partir de cette mise en correspondance entre la chaîne de référence et la chaîne décodée, il est possible de compter le nombre d'unités correctement identifiées, omises, substituées par une autre ou encore le nombre d'unités insérées au cours du décodage. Soit :

- $N_{ref}$  : Nombre d'unités de références
- $N_{ide}$  : Nombre d'unités identifiées
- $N_{omi}$  : Nombre d'unités omises
- $N_{sub}$  : Nombre d'unités substituées
- $N_{ins}$  : Nombre d'unités insérées

Alors nous avons :

$$T_{ide} = N_{ide}/N_{ref} = 100\% - (T_{subs} + T_{omis}) \quad (7.36)$$

Le taux d'identification ne tient pas compte de taux d'insertion, donc

$$T_{rec} = T_{ide} - T_{ins} \quad (7.37)$$

## 7.4 Système de commande vocale de l'environnement domestique pour les handicapés

Nous avons procédé à un système de reconnaissance de parole pour la commande vocale de l'environnement domestique pour des personnes handicapées. Nous avons évalué ce système avec deux personnes meyopates qui ont des difficultés articulaires.

Nous avons utilisé la base de données BREF pour l'apprentissage du modèle du monde. Cette base de données a les caractéristiques suivantes :

- contient 100 heures de parole
- enregistrée par 120 locuteurs (80 hommes 40 femmes)
- textes extraits du journal "le monde"
- contient un grand vocabulaire de mots (plus de 20 000 mots)
- contient plus de 1115 diphtongues
- contient plus de 17500 triphongues.

### 7.4.1 Description des bases de données de tests

Nous avons utilisé une base de données des personnes meyopates. Cette base de données a été enregistrée à l'hôpital de Garche par deux locuteurs meyopates. Les enregistrements sont faits dans les laboratoires de l'hôpital avec une fréquence d'échantillonnage de 16 Khz sur 16 bits. Chaque enregistrement a une durée de 6 à 20 secondes.

Le premier locuteur a enregistré 10 fois les phrases suivantes :

Ouvrir la porte  
Fermer le porte  
Allumer la télé  
Éteindre la télé

Le deuxième locuteur a enregistré 10 fois les phrases suivantes :

Ouvrir la porte  
Fermer le porte  
Ouvrir la fenêtre  
Fermer le fenêtre

Pour comparer les résultats obtenus avec des locuteurs normaux nous avons procédé à enregistrer les mêmes bases de données avec des locuteurs normaux parlant.

### 7.4.2 Description du système de reconnaissance de la parole

Le système de reconnaissance est un système à petit vocabulaire qui utilise le modèle de langage suivant :

La grammaire utilisée :

```

action1 = ouvrir | fermer ;
appareil1 = porte | chambre-enfant | chambre-parent | fenêtre ;
action2 = allumer | éteindre ;
appareil2 = chambre-enfant | chambre-parent | lampe ;
action3 = commencer | pause | arrêt | rapide | normal ;
appareil3 = lave-linge | vitesse ;
action4 = allumer | éteindre | volume-plus | volume-moins | pause ;
appareil4 = tele | frigo | chaine-hifi ;
action5 = lecture-avant | lecture-arriere | avance | retour | enregistrer ;
appareil5 = cassette | vidéo ;
action6 = raccrocher | décrocher | appeler | enregistrer ;
appareil6 = téléphone ;
iy = la | le ;
phrase1 = ((action1)(iy)(appareil1)) ;
phrase 2 = ((action2)(iy)(appareil2)) ;
phrase3 = ((action3)(iy)(appareil3)) ;
phrase4 = ((action4)(iy)(appareil4)) ;
phrase5 = ((action5)(iy)(appareil5)) ;
phrase6 = ((action6)(iy)(appareil6)) ;

```

( (sil) phrase1 | phrase 2 | phrase3 | phrase4 | phrase5 | phrase6 (sil) )

Nous avons analysé les résultats obtenus en utilisant :

- les modèles de monde
- les modèles de monde adaptés avec adaptation des moyennes
- les modèles de monde adaptés avec adaptation des moyennes et des covariances

### 7.4.3 Utilisation du modèle du monde

Cette expérience correspond à utiliser le modèle du monde appris sur BREF pour la reconnaissance de la parole pour des locuteurs handicapés.

Les mêmes phrases tests du locuteur 1 sont enregistrées par un locuteur normal.

Les résultats de reconnaissance pour le locuteur 1 sont résumés dans le tableau suivant :

Phrase	T.rec pour le locuteur normal	T.rec pour le locuteur handicapé
phrase 1	100%	30%
phrase 2	100%	30%
phrase 3	100%	10%
phrase 4	100%	20%

TAB. 7.1 – T.rec avec les modèles du monde pour le locuteur 1

Pour un système de reconnaissance de parole continue à petit vocabulaire indépendant du locuteur, le taux de reconnaissance de 100% est atteint par les systèmes de reconnaissance de mot isolés actuels sur la marché ce qui explique le taux de reconnaissance obtenu par le locuteur normale (100%), mais d'avoir un taux de reconnaissance assez faible (moins de 30%) pour le locuteur handicapé explique le fait que ce locuteur à un degré de handicap articulatoire assez lourd.

Les mêmes phrases tests de le locuteur 2 sont enregistrées par un locuteur normal. Les résultats sont résumés dans le tableau 7.2.

Phrase	T.rec pour locuteur normale	T.rec pour locuteur handicapée
phrase 1	100%	40%
phrase 2	100%	50%
phrase 3	100%	40%
phrase 4	100%	60%

TAB. 7.2 – T.rec avec les modèles du monde pour le locuteur 2

Le taux de reconnaissance obtenu montre qu'un système de reconnaissance de la parole pour des locuteurs normaux ne donne pas des résultats satisfaisants pour des locuteurs handicapés, ce qui prouve que les utilisateurs handicapés articulatoires

ont beaucoup de difficultés à utiliser des systèmes de reconnaissance de la parole grand public.

Pour essayer d'adapter au mieux les modèles du monde aux locuteurs handicapés, nous avons adapté ces modèles en utilisant une partie de la base de test comme base d'adaptation.

#### 7.4.4 Adaptation des modèles (adaptation des moyennes) du monde aux locuteurs handicapés

Nous avons utilisé le modèle du monde et la méthode MLLR pour l'adaptation des moyennes. Nous avons utilisé pour chaque type de phrase 5 phrases pour l'adaptation et les 5 autres pour le test. Et pour étudier au mieux le processus d'adaptation nous avons étudié l'évolution du taux de reconnaissance en fonction de l'évolution de la base d'adaptation.

Les résultats obtenus pour le locuteur 1 sont résumés dans le tableau 7.3.

Phrase	T.rec après 2 adaptations	T.rec après 5 adaptations
phrase 1	20%	60%
phrase 2	40%	80%
phrase 3	20%	40%
phrase 4	20%	60%

TAB. 7.3 – T.rec avec les modèles adaptés (adaptation des moyennes) pour le locuteur 1

L'adaptation des modèles phonétiques pour la voix d'un locuteur donné même s'il a un handicap lourd augmente les performances du système de reconnaissance.

Pour le locuteur 2 les résultats sont résumés dans le tableau 7.4.

Phrase	T.rec après 2 adaptations	T.rec après 5 adaptations
phrase 1	40%	60%
phrase 2	20%	60%
phrase 3	20%	40%
phrase 4	20%	40%

TAB. 7.4 – T.rec avec les modèles adaptés pour le locuteur 2

L'augmentation de taux de reconnaissance avec la taille de la base d'adaptation montre que plus la base d'adaptation est plus importante, plus que les modèles sont mieux adaptés aux locuteurs handicapés.

Nous avons procédé à une autre expérience, en utilisant les 10 répétitions de la phrase “ ouvrir la porte “ comme base d'adaptation et les 10 répétitions de la phrase “ fermer la porte “ comme base de test et inversement. Nous avons obtenu les résultats donnés par le tableau 7.5.

Phrase	T.rec après 5 adaptations	T.rec après 10 adaptations
phrase 1	40%	60%
phrase 2	50%	80%

TAB. 7.5 – T.rec avec les modèles partiellement adaptés

L'adaptation de certains modèles de reconnaissance aux locuteurs handicapés augmente le taux de reconnaissance. Cette augmentation est expliquée par le fait qu'un mot est une suite de phonème, donc en utilisant la transcription phonétique d'un mot, ce dernier va devenir une suite d'états de ces phonèmes, donc même une adaptation partielle peut avoir une influence sur le taux de reconnaissance.

Les résultats obtenus montrent que le taux de la reconnaissance augmente avec l'adaptation des moyennes. Alors que le taux de reconnaissance n'atteint pas les 100% pour un système de reconnaissance de 40 phrases, d'où l'utilisation de l'adaptation de la covariance pour essayer d'augmenter le taux de reconnaissance.

#### 7.4.5 Adaptation des modèles (adaptation des moyennes et des covariances) du monde aux locuteurs handicapés

Nous avons utilisé le modèle du monde appris sur BREF. Nous avons appliqué l'adaptation de la moyenne et de la covariance et nous avons fait les mêmes manipulations qu'avec l'adaptation des moyennes.

Les résultats obtenus pour le locuteur 1 sont résumés dans le tableau suivant :

Phrase	T.rec après 3 adaptations	T.rec après 5 adaptations
phrase 1	60%	100%
phrase 2	80%	100%
phrase 3	60%	100%
phrase 4	60%	100%

TAB. 7.6 – T.rec avec les modèles adaptés (adaptation des moyennes et des covariances) pour le locuteur 1

L'adaptation des moyennes et des covariances pour les modèles phonétiques est meilleur pour un locuteur handicapé ce qui explique le taux de reconnaissance de 100% après les 5 phrases d'adaptations.

Les résultats obtenus pour le locuteur 2 sont résumés dans le tableau 7.7.

Phrase	T.rec après 3 adaptations	T.rec après 6 adaptations
phrase 1	60%	100%
phrase 2	60%	80%
phrase 3	40%	80%
phrase 4	60%	100%

TAB. 7.7 – T.rec avec les modèles adaptés pour le locuteur 2

Même pour un locuteur qui a un handicap articuloire lourd, l'augmentation de taux de reconnaissance avec l'adaptation des moyennes et des covariances est assez intéressante.

Nous avons procédé à une autre expérience, en utilisant les 10 répétitions de la phrase “ ouvrir la porte “ comme base d'adaptation et les 10 répétitions de la phrase “ fermer la porte “ comme base de test et inversement.

Nous avons obtenus les résultats suivants :

Phrase	T.rec après 5 adaptations	T.rec après 10 adaptations
phrase 1	60%	100%
phrase 2	70%	100%

TAB. 7.8 – T.rec avec les modèles partiellement adaptés

Les résultats obtenus avec l'adaptation des moyennes et des covariances sont meilleurs qu'avec l'adaptation des moyennes, le taux de reconnaissance atteint des valeurs proche de 100% pour une grammaire de 40 phrases. Donc l'adaptation de la covariance est plus significative que l'adaptation des moyennes.

#### 7.4.6 Utilisation des modèles adaptés aux locuteurs handicapés pour la reconnaissance de la parole des locuteurs normaux

Cette expérience consiste à adapter le modèle du monde au locuteur handicapé, puis utiliser ces modèles pour reconnaître la parole d'un locuteur normal, puis faire la même expérience avec l'autre locuteur handicapé. l'expérience consiste à augmenter chaque fois le nombre de phrases d'adaptation et de voir son effet sur le taux de reconnaissance.

Le tableau suivant résume les résultats obtenus pour le locuteur normale 1.

Phrase	T.rec après 2 adaptations	T.rec après 5 adaptations	T.rec après 10 adaptations
phrase 1	80%	40%	00%
phrase 2	90%	40%	00%
phrase 3	60%	10%	00%
phrase 4	70%	30%	00%

TAB. 7.9 – T.rec avec les modèles adaptés au locuteurs handicapés pour le locuteur normale 1

Les modèles adaptés au locuteur handicapé est propre à lui ce qui explique les taux de reconnaissance obtenus pour le locuteur normal.

Le tableau suivant résume les résultats obtenus pour le locuteur normale 2.

Phrase	T.rec après 3 adaptations	T.rec après 5 adaptations	T.rec après 10 adaptations
phrase 1	80%	30%	00%
phrase 2	70%	20%	00%
phrase 3	60%	20%	00%
phrase 4	90%	30%	00%

TAB. 7.10 – T.rec avec les modèles adaptés au locuteurs handicapés pour le locuteur normale 2

La parole du locuteur handicapé présente des différences par rapport à celui du locuteur normale.

Les résultats obtenus montrent que les locuteurs handicapés ont des modèles propres à eux, ce qui les empêche à utiliser les systèmes de reconnaissance de parole grand public. L'adaptation d'un modèle initial (du monde) est une solution pour que les locuteurs handicapés utilisent quelques systèmes de reconnaissance de la parole de mots isolés grand public.

Même les locuteurs handicapés n'ont pas le même handicap articulatoire, ce qui nous amène à étudier chaque handicap à part.

## 7.5 Système de reconnaissance de parole continue pour des parkinsoniens

### 7.5.1 Système de référence

Lors du développement d'un système de RAP, il est nécessaire d'évaluer les performances de ce système pour le comparer aux systèmes existants. De plus, les améliorations apportées à ce système au cours de son développement doivent être quantifiées, afin de juger objectivement de leur pertinence. Deux approches sont

envisageables : tester le système sur une base de données de référence admise par la communauté, ou comparer dans les mêmes conditions de test les performances du système développé à celles d'un système de référence. Dans notre application, nous avons utilisé les mêmes conditions de test qu'avec le système de référence. Nous avons utilisé des modèles de phonèmes appris sur BREF, et un modèle de langage de BREF. Nous avons fait les tests avec une base d'évaluation de 10 locuteurs de BREF, nous avons obtenus les résultats résumés dans le tableau 7.11.

Locuteurs	Taux de reconnaissance
locuteur 1	23.36%
locuteur 2	23.93%
locuteur 3	37.5%
locuteur 4	21.21%
locuteur 5	15.22%
locuteur 6	18.75%
locuteur 7	22.50%
locuteur 8	23.27%
locuteur 9	15.14%
locuteur 10	17.91%

TAB. 7.11 – Taux de reconnaissance avec les modèles du BREF pour la base de données BREF

## 7.5.2 Procédure expérimentale

Nous avons analysé notre approche d'adaptation dynamique sur une base de données de parkinsoniens avec un système de reconnaissance de parole continue.

### Base de données

La base a été enregistrée dans l'hôpital d'Aix en Provence. Ces enregistrements sont fournis par Monsieur Bernard TESTON Ingénieur de Recherche au CNRS au Laboratoire Parole et Langage de l'université de Provence.

Cette base de données comprend les enregistrements de 200 parkinsoniens, chaque enregistrement à une durée de 49 secondes à 4 minutes suivant le degré du handicap de la personne.

Chaque locuteur est censé d'enregistrer une partie de "LA CHÈVRE DE MONSIEUR SEGUIN" qui est :

*M. Séguin n'avait jamais eu de bonheur avec ces chèvres. Il les perdait toutes de la même façon : un beau matin, elles cassaient leur corde, s'en allaient dans la montagne, et là haut le loup les mangeait. Ni les caresses de leurs maître, ni la peur du loup, rien ne les retenait. c'était paraît-il, des chèvres indépendantes, voulant à tout prix le grand air et la liberté. Le brave M. Séguin, qui ne comprenait rien au caractère de ses bêtes, était consterné. Il disait :*



*C'est fini ; les chèvres s'ennuient chez moi, je ne garderai pas une. Cependant, il ne découragea pas, et, après avoir perdu six chèvres de la même manière, il en acheta une septième ; seulement, cette fois, il eut soin de la prendre toute jeune, pour qu'elle s'habitât à demeurer chez lui.*

*Ah ! Gringoire, qu'elle était jolie la petite chèvre de M. Séguin ! qu'elle était jolie avec ses yeux doux, sa barbiche de sous-officier, ses sabots noirs et luisants, ses cornes zébrées et ses long poils blancs qui lui faisaient une houppe !*

Les personnes enregistrées ont été classées en sept niveaux de handicaps (0, 0.5, 1, 1.5, 2, 2.5, 3, ). Cette classification est faite à l'aide des tests mis à disposition dans l'hôpital par des cliniciens. Le tableau suivant indique le nombre de locuteurs pour chaque niveau :

Niveaux	Nombre de locuteurs
niveau 0	39
niveau 0.5	21
niveau 1	84
niveau 1.5	8
niveau 2	30
niveau 2.5	3
niveau 3	13

TAB. 7.12 – Nombre de locuteurs par niveau

Nous avons regroupé les handicaps en 4 degrés (degré 0 (0 + 0.5), degré 1 (1 + 1.5), degré 2 (2) et degré 3 (2.5 + 3)).

### Description de la procédure expérimentale

L'adaptation est la solution la plus utilisée pour les systèmes de reconnaissance de la parole continue grand public, mais cette solution n'est pas très efficace pour les personnes handicapées.

L'utilisation de la mémoire de parole de la personne acquise lors de l'utilisation de système est l'approche originale que nous proposons.

Nous avons analysé notre approche avec trois systèmes :

- 1- Système monolocuteur : pour chaque locuteur, nous avons segmenté la base de données en 5 bases : 4 bases pour l'adaptation et une base de test. Nous avons analysé l'évolution des taux de reconnaissance avec l'adaptation successive avec 4 bases d'adaptation pour chaque locuteur.
- 2- Système multilocuteurs pour chaque degré de handicap : pour chaque degré de handicap, nous avons pris des locuteurs pour l'adaptation et d'autre pour l'évaluation et nous avons analysé l'évolution de taux de reconnaissance en fonction de la base d'adaptation.
- 3- Système indépendant de locuteur pour les parkinsoniens : ce système repose sur le même principe que le précédent mais pour toute la base des parkinsoniens.

### Système monolocuteur avec adaptation dynamique

Pour chaque degré de handicap nous avons choisi 10 locuteurs. Pour chaque locuteur nous avons choisi 4 fichiers pour l'adaptation et un fichier pour le test.

L'adaptation du système à une personne handicapée est plus lente que son adaptation à une personne normale parce que nous partons d'un système de reconnaissance de parole pour des personnes normales. L'approche d'adaptation dynamique est une solution originale qui permet aux personnes handicapées d'utiliser les systèmes grand publics. Certes au départ les performances ne seront pas satisfaisantes, mais il vont augmenter au fur et à mesure de l'utilisation du système.

Les résultats de reconnaissance obtenus avec les modèles de BREF sont les suivants :

Locuteurs	T.rec pour le degré 0	T.rec pour le degré 1	T.rec pour le degré 2	T.rec pour le degré 3
locuteur 1	5.12%	5.12%	2.56%	2.56%
locuteur 2	7.69%	2.56%	5.12%	2.56%
locuteur 3	5.12%	7.89%	5.12%	00%
locuteur 4	2.56%	2.56%	2.56%	5.12%
locuteur 5	5.12%	5.12%	00%	2.56%
locuteur 6	5.12%	5.12%	7.69%	5.12%
locuteur 7	7.89%	00%	2.56%	2.56%
locuteur 8	5.12%	7.89%	5.12%	00%
locuteur 9	7.89%	5.12%	5.12%	5.12%
locuteur 10	12.82%	7.89%	5.12%	5.12%

TAB. 7.13 – Taux de reconnaissance pour les locuteurs handicapés avec les modèles de BREF

Les résultats obtenus montrent qu'un système pour des personnes normales parlant ne donne pas des résultats satisfaisants avec des personnes qui ont un handicap articulaire.

Ces faibles performances sont dû aux deux causes suivantes :

- Les modèles utilisés sont des modèles appris sur une base de données de personnes normales parlant qui ne tient pas compte des difficultés articulatoires des personnes handicapées.
- Le lexique utilisé n'est pas adapté aux enregistrements de test. Le lexique est pris des textes de journal le monde tandis que les phrases tests sont des phrases d'un texte littéraire.

Maintenant nous analysons l'évolution des performances avec l'adaptation successive des 4 bases d'adaptation.

Les résultats obtenus pour chaque degré sont résumé dans les tableaux suivants :

Locuteurs	T.rec avec les modèles de BREF	T.rec après 1 adaptation	T.rec après 2 adaptations	T.rec après 3 adaptations	T.rec après 4 adaptations
locuteur 1	5.12%	5.12%	7.69%	10.25%	12.82%
locuteur 2	7.69%	10.25%	10.25%	12.82%	15.38%
locuteur 3	5.12%	7.69%	10.25%	12.82%	15.38%
locuteur 4	2.56%	5.12%	5.12%	7.69%	10.25%
locuteur 5	5.12%	5.12%	7.69%	7.69%	10.25%
locuteur 6	5.12%	10.25%	12.82%	12.82%	15.58%
locuteur 7	7.69%	10.25%	12.82%	15.38%	15.38%
locuteur 8	5.12%	7.69%	7.69%	12.82%	15.38%
locuteur 9	7.69%	10.25%	12.82%	15.38%	15.38%
locuteur 10	12.82%	17.94%	17.94%	17.94%	23.07%

TAB. 7.14 – Évolution de taux de reconnaissance pour le degré 0

Locuteurs	T.rec avec les modèles de BREF	T.rec après 1 adaptation	T.rec après 2 adaptations	T.rec après 3 adaptations	T.rec après 4 adaptations
locuteur 1	5.12%	7.69%	7.69%	10.25%	10.25%
locuteur 2	2.56%	5.12%	7.69%	7.69%	10.25%
locuteur 3	7.69%	10.25%	12.82%	12.82%	15.38%
locuteur 4	2.56%	2.56%	5.12%	5.12%	7.69%
locuteur 5	5.12%	7.69%	10.25%	10.25%	12.82%
locuteur 6	5.12%	7.69%	7.69%	10.25%	10.25%
locuteur 7	00%	00%	2.56%	5.12%	7.69%
locuteur 8	7.69%	10.25%	10.25%	12.82%	15.38%
locuteur 9	5.12%	5.12%	7.69%	10.25%	12.82%
locuteur 10	7.69%	10.25%	10.25%	12.82%	15.38%

TAB. 7.15 – Évolution de taux de reconnaissance pour le degré 1

Locuteurs	T.rec avec les modèles de BREF	T.rec après 1 adaptation	T.rec après 2 adaptations	T.rec après 3 adaptations	T.rec après 4 adaptations
locuteur 1	2.56%	5.12%	5.12%	7.69%	10.25%
locuteur 2	5.12%	4.69%	7.69%	10.25%	12.82%
locuteur 3	5.12%	5.12%	7.69%	10.25%	12.82%
locuteur 4	2.56%	5.12%	7.69%	7.69%	10.25%
locuteur 5	00%	2.56%	5.12%	7.69%	7.69%
locuteur 6	7.69%	7.69%	10.25%	12.82%	12.82%
locuteur 7	2.56%	5.12%	5.12%	7.69%	10.25%
locuteur 8	5.12%	7.69%	10.25%	10.25%	12.82%
locuteur 9	5.12%	7.69%	7.69%	7.69%	10.25%
locuteur 10	5.12%	7.69%	10.25%	12.82%	15.38%

TAB. 7.16 – Évolution de taux de reconnaissance pour le degré 2

Locuteurs	T.rec avec les modèles de BREF	T.rec après 1 adaptation	T.rec après 2 adaptations	T.rec après 3 adaptations	T.rec après 4 adaptations
locuteur 1	2.56%	5.12%	5.12%	7.69%	7.69%
locuteur 2	2.56%	2.56%	5.12%	7.69%	7.69%
locuteur 3	00%	2.56%	2.56%	5.12%	5.12%
locuteur 4	5.12%	7.69%	10.25%	10.25%	12.82%
locuteur 5	2.56%	7.69%	10.25%	10.25%	12.82%
locuteur 6	5.12%	7.69%	7.69%	10.25%	10.25%
locuteur 7	2.56%	5.12%	7.69%	7.69%	10.25%
locuteur 8	00%	00%	2.56%	5.12%	7.69%
locuteur 9	5.12%	7.69%	7.69%	10.25%	10.25%
locuteur 10	5.12%	7.69%	10.25%	12.82%	12.82%

TAB. 7.17 – Évolution de taux de reconnaissance pour le degré 3

L'adaptation dynamique permet d'augmenter les performances de système de reconnaissance. La base d'adaptation utilisée ne dépasse pas quelques minutes. Si nous supposons que la personne utilise son système de reconnaissance 30 minutes par jour, donc pendant un mois nous avons 15 heures de parole pour cette personne qui nous permet d'avoir des résultats beaucoup plus intéressants.

### **Système indépendant de locuteur pour chaque degré de handicap avec adaptation dynamique**

Pour ce type de système nous avons analysé un système de reconnaissance de parole pour chaque degré de handicap (0, 1, 2 et 3).

Pour chaque degré de handicap nous avons pris 10 locuteurs pour l'adaptation et 10 locuteurs pour le test. Nous avons adapté les modèles du monde avec les 10 locuteurs d'adaptation, puis nous avons fait les tests de reconnaissance sur les 10 autres locuteurs, pour le niveau 3 nous avons utilisé 8 locuteurs pour l'adaptation et 8 locuteurs pour les tests.

Les résultats obtenus sont résumés dans les tableaux suivants :

Locuteurs	T.rec avec les modèles de BREF	T.rec avec adaptation dans le même degré de handicap
locuteur 1	5.12%	12.82%
locuteur 2	7.69%	12.82%
locuteur 3	5.12%	15.38%
locuteur 4	2.56%	10.25%
locuteur 5	5.12%	10.25%
locuteur 6	5.12%	12.82%
locuteur 7	7.69%	15.38%
locuteur 8	5.12%	12.82%
locuteur 9	7.69%	15.38%
locuteur 10	12.82%	17.94%

TAB. 7.18 – Évolution de taux de reconnaissance pour le degré 0

Locuteurs	T.rec avec les modèles de BREF	T.rec avec adaptation dans le même degré de handicap
locuteur 1	5.12%	10.25%
locuteur 2	2.56%	7.69%
locuteur 3	7.69%	12.82%
locuteur 4	2.56%	7.69%
locuteur 5	5.12%	10.25%
locuteur 6	5.12%	10.25%
locuteur 7	00%	5.12%
locuteur 8	7.69%	12.82%
locuteur 9	5.12%	12.82%
locuteur 10	7.69%	12.82%

TAB. 7.19 – Évolution de taux de reconnaissance pour le degré 1

Locuteurs	T.rec avec les modèles de BREF	T.rec avec adaptation dans le même degré de handicap
locuteur 1	2.56%	10.25%
locuteur 2	5.12%	10.25%
locuteur 3	5.12%	12.82%
locuteur 4	2.56%	10.25%
locuteur 5	00%	7.69%
locuteur 6	7.69%	12.82%
locuteur 7	2.56%	7.69%
locuteur 8	5.12%	10.25%
locuteur 9	5.12%	10.25%
locuteur 10	5.12%	7.69%

TAB. 7.20 – Évolution de taux de reconnaissance pour le degré 2

Locuteurs	T.rec avec les modèles de BREF	T.rec avec adaptation dans le même degré de handicap
locuteur 1	2.56%	7.69%
locuteur 2	2.56%	10.25%
locuteur 3	00%	5.12%
locuteur 4	5.12%	10.25%
locuteur 5	2.56%	7.69%
locuteur 6	5.12%	10.25%
locuteur 7	2.56%	5.12%
locuteur 8	00%	2.56%

TAB. 7.21 – Évolution de taux de reconnaissance pour le degré 3

L'augmentation de taux de reconnaissance avec l'adaptation des locuteurs de même degré de handicap augmente le taux de reconnaissance, ce qui signifie qu'on peut avoir un système de reconnaissance de la parole pour des personnes qui ont le même degré de handicap.

### Système indépendant de locuteur pour des parkinsoniens avec adaptation dynamique

Nous avons utilisé 40 locuteurs (10 locuteurs pour chaque degré de handicap) pour l'adaptation et pour chaque degré de handicap 10 locuteurs pour les évaluations. Les résultats obtenus sont résumés dans les tableaux suivants :

Locuteurs	T.rec avec les modèles de BREF	T.rec avec adaptation dans le même type de handicap
locuteur 1	5.12%	15.38%
locuteur 2	7.69%	15.38%
locuteur 3	5.12%	15.38%
locuteur 4	2.56%	12.82%
locuteur 5	5.12%	12.82%
locuteur 6	5.12%	20.51%
locuteur 7	7.69%	20.51%
locuteur 8	5.12%	15.38%
locuteur 9	7.69%	17.94%
locuteur 10	12.82%	25.64%

TAB. 7.22 – Évolution de taux de reconnaissance pour le degré 0

Locuteurs	T.rec avec les modèles de BREF	T.rec avec un système de reconnaissance de la parole parkinsonien
locuteur 1	5.12%	12.82%
locuteur 2	2.56%	10.25%
locuteur 3	7.69%	17.94%
locuteur 4	2.56%	12.82%
locuteur 5	5.12%	12.82%
locuteur 6	5.12%	15.38%
locuteur 7	00%	10.25%
locuteur 8	7.69%	15.38%
locuteur 9	5.12%	15.38%
locuteur 10	7.69%	17.94%

TAB. 7.23 – Évolution de taux de reconnaissance pour le degré 1

Locuteurs	T.rec avec les modèles de BREF	T.rec avec un système de reconnaissance de la parole parkinsonien
locuteur 1	2.56%	12.82%
locuteur 2	5.12%	15.38%
locuteur 3	5.12%	17.94%
locuteur 4	2.56%	12.82%
locuteur 5	00%	10.25%
locuteur 6	7.69%	15.38%
locuteur 7	2.56%	12.82%
locuteur 8	5.12%	12.82%
locuteur 9	5.12%	15.38%
locuteur 10	5.12%	10.25%

TAB. 7.24 – Évolution de taux de reconnaissance pour le degré 2

Locuteurs	T.rec avec les modèles de BREF	T.rec avec un système de reconnaissance de parole parkinsonien
locuteur 1	2.56%	10.25%
locuteur 2	2.56%	12.82%
locuteur 3	00%	10.25%
locuteur 4	5.12%	15.38%
locuteur 5	2.56%	10.25%
locuteur 6	5.12%	12.82%

TAB. 7.25 – Évolution de taux de reconnaissance pour le degré 3



Pour des personnes qui ont les mêmes types de handicaps articulatoires, un système de reconnaissance de parole continue est réalisable. Les difficultés articulatoires d'une personne donnée affectent en fait la production de quelques phonèmes.

une étude plus précise sur les phonèmes affectés chez ces locuteurs, nous permet de définir une base d'adaptation plus orientée qui peut donner des résultats plus intéressants. Cette étude nous permet aussi de définir les modèles phonétiques à adapter.

## 7.6 Conclusion

L'adaptation dynamique est une approche assez compatible avec les personnes handicapées. Le système de reconnaissance de la parole s'adapte au fur et à mesure aux utilisateurs. L'augmentation de taux de reconnaissance en cours de l'utilisation du système encourage les utilisateurs à l'utiliser de mieux en mieux. Après certaines utilisations les performances du système se stabilisent ce qui nous permet d'arrêter l'adaptation.

Le regroupement de personnes par type de handicap articulatoire nous permet de construire un système de reconnaissance de la parole pour chaque type de handicap. La classification de type de handicap nous permet d'orienter les recherches selon le type de handicap tout en partant des mêmes principes de solutions.



## Chapitre 8

# Conclusions et perspectives

La reconnaissance de parole pour les personnes qui ont des handicaps articulatoires nécessite une étude de chaque type de handicap ce qui peut compliquer la tâche d'avoir un système de reconnaissance de la parole pour des personnes handicapées. Les approches que nous avons proposées dans notre travail sont des approches indépendantes du types de handicaps articulatoires.

L'approche d'un système de reconnaissance de parole en utilisant la programmation dynamique avec apprentissage dynamique pour un système de reconnaissance de mots isolés donne des résultats assez importants. Le taux de reconnaissance doit atteindre les performances des systèmes classiques après certaines utilisation du système par la personne handicapée. Cette solution est faite pour les systèmes à petits vocabulaires parce que nous ne disposons pas de bases de données des personnes handicapées qui nous permet d'utiliser les HMM.

L'utilisation d'une segmentation non phonétique pour la reconnaissance de la parole pour les personnes handicapées est une approche prometteuse. En effet, cette segmentation ne tient pas compte des difficultés articulatoires chez une personne donnée ce qui nous permet d'utiliser des bases de données des personnes normales pour l'apprentissage de modèles des segments ALISP. Le système de reconnaissance de parole à petit vocabulaire qui utilise cette méthode donne des résultats assez intéressants. Cette approche à l'avantage par rapport à l'approche précédente d'utiliser les modèles statistiques.

L'utilisation de la segmentation ALISP pour construire un système de reconnaissance de parole continue pour des personnes handicapées est dans la bonne voie, mais faute d'avoir une base de données de personnes handicapées assez significative ne nous a pas permis de mener nos expériences jusqu'au bout.

L'approche d'adaptation dynamique des modèles statistiques a donné des résultats assez intéressants. Cette approche peut donner naissance à des systèmes de reconnaissances de parole indépendants du locuteur pour chaque type de handicap articulatoire.

La première perspective de notre travail est d'utiliser l'aspect non supervisé. En fait pour nos deux approches d'apprentissage dynamique et adaptation dynamique nous utilisons l'aspect supervisé c'est à dire l'utilisateur doit valider pour que l'adaptation ou l'apprentissage dynamique soit réaliser, ce qui peut gêner l'utilisa-

teur, même si la plupart des systèmes de commande vocale par exemple demandent la confirmation avant l'exécution de la commande.

La deuxième perspective est de coupler l'utilisation de la segmentation ALISP et l'adaptation dynamique des modèles statistiques pour construire le système de reconnaissance de parole continue pour des personnes handicapées.

La troisième perspective est de construire des bases de données de personnes handicapées, vue que les performances des systèmes de reconnaissance dépendent fortement des bases de données d'apprentissage.

## Annexes



## Annexe A

### correspondance sous-classes-ALISP

Les résultats de correspondance pour quelques sous-classes.

1 - La sous classe de HB : La classe HB correspond à 14 polysons (12 dipphones et 2 triphones).

La classe HB contient 11 sous-classes vides.

40 sous-classes de HB correspondent de 2 à 4 polysons.

Le tableau 6.3 donne l'exemple des sous-classes de la classe HB

	W A	W AN	W OU	V AN	V O	F O	T O	R EU	R O	R U
H6HB	49	3	4	0	0	0	0	0	0	0
HdHB	8	0	0	12	2	0	0	0	0	0
HIHB	0	0	0	0	0	32	3	0	0	0
HnHB	0	0	0	0	0	0	0	48	8	30

TAB. A.1 – Correspondances de quelques sous-classes de HB avec des polysons

La sous-classe H6HB correspond à 88% au diphone W A.

La sous-classe HdHB correspond à 55% au diphone V AN et avec 36% au diphone W A.

La sous-classe HIHB correspond avec 95% au diphone F O.

La sous-classe HnHB correspond avec 50% au diphone R EU, avec 40% au diphone R U et avec 10% au diphone R O.

2 - La classe HC : Elle correspond à 16 polysons (14 dipphones et 2 triphones)

La classe HC contient 17 sous-classes vides

41 sous-classes correspondent à 2 jusqu'à 4 polysons.

Le tableau suivant donne l'exemple des sous-classes de la classe HC

	An #	AN P	AN S	OU R	AU K	OU R	U R	R O	K AN	P OU R	P ON
HIHC	17	11	12	0	0	0	0	0	0	0	0
H6HC	12	0	0	25	3	0	0	0	0	0	0
H7HC	0	0	0	0	0	10	3	0	0	0	0
HeHC	0	0	0	0	0	0	0	8	18	0	8
HhHC	0	0	0	0	0	0	0	0	0	8	4

TAB. A.2 – Correspondances de quelques sous-classes de HC avec des polysons

Pour la sous-classe HBHC correspond avec 100% au diphone R O (98 segments)

La sous-classe HIHC correspond avec 45% au diphone AN #, 30% au diphone AN P et 25% au diphone AN S.

La sous-classe H6HC correspond avec 65% au diphone OU R, avec 28% au diphone AU # et avec 10% au diphone AU K.

La sous-classe H7HC correspond avec 80% au diphone OU R et avec 20% au diphone U R.

La sous-classe HeHC correspond avec 50% au diphone K AN, avec 25% au diphone P ON et avec 25% au diphone R O.

La sous-classe HhHC correspond avec 65% au triphone P OU R et avec 35% au diphone P ON.

La sous-classe HzHC correspond avec 100% au diphone S AN.

3 - La classe HD : Elle correspond à 12 polysons (11 diphones et 1 triphone)

La classe HD contient 21 sous-classes vides.

40 sous classes de HD correspondent à 2 jusqu'à 4 polysons.

Le tableau suivant donne l'exemple des sous-classes de la classe HD

	A #	A R	A P	O R	EU R	E R	U R	R A	R EU	R #
HGHD	0	0	0	0	50	14	8	0	0	0
HKHD	40	30	8	3	0	0	0	0	0	0
HeHD	0	0	0	0	0	0	0	18	4	20
HnHD	0	0	0	0	0	0	0	0	8	18

TAB. A.3 – Correspondances de quelques sous-classes de HD avec des polysons

La sous-classe HGHD correspond avec 70% au diphone EU R, avec 20% au diphone E R et avec 10% au diphone U R.

La sous-classe HKHD correspond avec 50% au diphone A #, avec 40% au diphone A R et avec 10% au diphone A P.

La sous-classe HeHD correspond avec 50% au diphone R #, avec 40% au diphone R A et avec 10% au diphone R EU

La sous-classe HnHD correspond avec 80% au diphone R EU et avec 15% au diphone R A

4 - La sous classe HG : Elle correspond à 11 polysons (11 diphones)

La classe HG contient 22 sous-classes vides.

41 sous-classes correspondent à 2 jusqu'à 4 polysons.



Le tableau suivant donne l'exemple des sous-classes de la classe HG

	N O	N O T	N O N	EI R	EU R	T EU	S EU	F EU	T O
HQHG	48	20	3	0	0	0	0	0	0
H0HG	0	0	0	16	4	0	0	0	0
HkHG	0	0	0	0	0	44	20	18	6

TAB. A.4 – Correspondances de quelques sous-classes de HG avec des polysons

La sous-classe HQHG correspond avec 70% au diphone N O et avec 25% au triphone N O T.

La sous-classe H0HD correspond avec 80% au diphone EI R et avec 20% au diphone EU R.

La sous-classe HkHG correspond avec 50% au diphone T EU, avec 25% au diphone S EU et avec 20% au diphone F EU.

5 - La sous classe HH : Elle correspond à 15 polysons (13 diphones et 3 triphones)

La classe HH contient 17 sous-classes vides.

43 sous-classes correspondent à 2 jusqu'à 4 polysons.

Le tableau suivant donne l'exemple des sous-classes de la classe HH

	E #	R #	R AI	A V	A P	A T	IN #	AI R V	IN F	AI R	A V	R A
HDHH	30	18	4	0	0	0	0	0	0	0	0	0
HLHH	0	0	0	64	30	18	24	0	0	0	0	0
HOHH	0	0	0	0	0	0	0	34	20	28	24	0
HfHH	0	15	16	0	0	0	0	0	0	0	0	12

TAB. A.5 – Correspondances de quelques sous-classes de HH avec des polysons

La sous-classe HDHH correspond avec 60% au diphone E #, avec 35% au diphone R # et avec 5% au diphone R AI.

La sous-classe HLHH correspond avec 45% au diphone A V, avec 23% au diphone A P, avec 18% au diphone IN # et avec 14% au diphone A T.

La sous-classe HOHH correspond avec 35% au triphone AI R V, avec 20% au diphone IN F et avec 25% au diphone AI R.

La sous-classe HfHH correspond avec 36% au diphone R #, avec 38% au diphone R AI et avec 28% au diphone R A.

6 - La sous classe HI : Elle correspond à 13 polysons (13 diphones )

La classe HI contient 27 sous-classes vides.

33 sous-classes correspondent à 2 jusqu'à 4 polysons.

Le tableau suivant donne l'exemple des sous-classes de la classe HI

La sous-classe HAHl correspond avec 40% au diphone AN #, avec 25% au diphone AN N, avec 20% au diphone AN N et avec 10% au diphone AN S.

La sous-classe HMHI correspond avec 50% au diphone M AN, avec 25% au diphone L AN et avec 20% au diphone L AN.

	AN #	AN M	AN N	AN S	M AN	Y ON	T AN	L AN	# ON	R AN
HAHI	18	16	12	6	0	0	0	0	0	0
HMHI	0	0	0	0	16	2	6	8	0	0
H8HI	16	0	0	0	0	0	0	0	8	0
HeHI	0	0	0	0	3	0	0	0	0	15

TAB. A.6 – Correspondances de quelques sous-classes de HI avec des polysons

La sous-classe H8HI correspond avec 72% au diphone AN # et avec 28% au diphone # ON.

La sous-classe HeHI correspond avec 84% au diphone R AN et 16%.

7 - La sous classe HJ : Elle correspond à 10 polysons (10 diphones)

La classe HJ contient 15 sous-classes vides.

45 sous-classes correspondent à 2 jusqu'à 4 polysons.

Le tableau suivant donne l'exemple des sous-classes de la classe HJ

	ON P	ON B	ON M	ON N	ON L	AN D	AN T	AN #
HAHJ	30	25	30	20	15	0	0	0
HMHI	0	0	0	0	0	105	25	20

TAB. A.7 – Correspondances de quelques sous-classes de HJ avec des polysons

La sous-classe HAHJ correspond avec 25% au diphone ON P, avec 20% au diphone ON B, avec 25% au diphone ON M, avec 15% au diphone ON N et avec 10% au diphone ON L.

La sous-classe HIHJ correspond avec 65% au diphone AN D, avec 16% au diphone AN T et avec 19% au diphone AN #.

8 - La sous classe HK : Elle correspond à 11 polysons (11 diphones)

La classe HK contient 37 sous-classes vides.

19 sous-classes correspondent à 2 jusqu'à 4 polysons.

Le tableau suivant donne l'exemple des sous-classes de la classe HK

	W A	W IN	A #	R EU	R A	M A
HAHK	25	8	1	0	0	0
HMHK	0	0	0	0	0	40
HfHK	0	0	0	0	28	0
HnHK	3	0	0	6	6	0

TAB. A.8 – Correspondances de quelques sous-classes de HK avec des polysons

La sous-classe HAHK correspond avec 75% au diphone W A et avec 25% au diphone W IN.

La sous-classe HMHK correspond avec 100% au diphone M A (40).

Le sous-classes HfHK correspondent avec 100% au diphone R A.

La sous-classe HAHK correspond avec 40% au diphone R EU, avec 40% au diphone R A et avec 20% au diphone W A.

9 - La sous classe HL : Elle correspond à 13 polysons (10 diphones et 3 triphones)

La classe HL contient 30 sous-classes vides.

29 sous-classes correspondent à 2 jusqu'à 4 polysons.

Le tableau suivant donne l'exemple des sous-classes de la classe HL

	A R	A P	A D	AI R	OE R	P A R	P A	P A P	# A	# A P	# R
HGHL	25	20	6	0	0	0	0	0	0	0	0
HNHL	55	0	0	40	20	0	0	0	0	0	0
HWHL	3	0	0	0	15	7	0	0	0	0	0
HbHL	0	0	0	0	0	18	13	14	0	0	0
HjHL	0	0	0	0	0	0	0	0	20	10	3

TAB. A.9 – Correspondances de quelques sous-classes de HL avec des polysons

La sous-classe HGHL correspond avec 50% au diphone A R, avec 40% au diphone A P et avec 10% au diphone A D.

La sous-classe HGHL correspond avec 48% au diphone A R, avec 35% au diphone AI R et avec 18% au diphone OE R.

La sous-classe HWHL correspond avec 60% au triphone P A R, avec 28% au diphone P A et avec 12% au diphone A R.

La sous-classe HbHL correspond avec 40% au triphone P A R, avec 30% au triphone P A P et avec 30% au diphone P A.

La sous-classe HjHL correspond avec 60% au diphone # A, avec 30% au triphone # A P et avec 10% au diphone # R.

11 - La sous classe HM : Elle correspond à 14 polysons (13 diphones et 1 triphone)

La classe HM contient 32 sous-classes vides.

29 sous-classes correspondent à 2 jusqu'à 4 polysons.

Le tableau suivant donne l'exemple des sous-classes de la classe HM

	O M	O L	O N	EU M	A M	A B	AI M	AI B	A P	# M	V A
HBHM	30	18	10	13	0	0	0	0	0	0	0
HNHM	0	0	0	0	40	28	16	10	6	0	0
HcHM	0	0	0	0	0	0	0	0	0	14	33
HSBM	0	0	0	22	0	0	0	0	0	0	0

TAB. A.10 – Correspondances de quelques sous-classes de HM avec des polysons

La sous-classe HBHM correspond avec 42% au diphone O M, avec 25% au diphone O L, avec 14% au diphone O N et avec 18% au diphone EU M.

La sous-classe HNHM correspond avec 40% au diphone A M, avec 28% au diphone A B, avec 16% au diphone AI M et avec 10% au diphone AI B.

La sous-classe HcHM correspond avec 70% au diphone # M et avec 30% au diphone V A.

La sous-classe HSHM correspond avec 100% au diphone EU M

# Bibliographie

- [And88] R. André. A new statistical approach for the automatic segmentation of continous speech signals. *IEEE Trans, Acoustic, speech and signal processing*, 36 :29–40, 1988.
- [AOR92] C. Barras M. Caraty P. Delèglise C. Montacié R. André-Obrecht and X. Rodet. Décomposition temporelle et ruptures des modèles pour le décodage acoustico-phonétique. *JEP*, pages 335–340, 1992.
- [Ata83] B-S. Atal. Efficient coding of lpc parameters by temporal decomposition. *IEEE ICASSP*, pages 81–84, 1983.
- [Ave87] Averbuch. Experiemnts with the tangora 20 000 word speech recognizer. *IEEE ICASSP*, 1987.
- [BA91] F. Bimbot and B-S. Atal. Une évaluation de la décomposition temporelle. *Séminaire Greco-SFA*, 1991.
- [Bak75] J.K Baker. The dragon system : an overview. *IEEE Trans, Acoustic, speech and signal processing*, 23 :24–29, 1975.
- [Bar96] C. Barras. Reconnaissance de la parole continue : adaptation du locuteur et contrôle temporel des les modèles de markov cachés. *Thèse à l'université Paris VI*, 1996.
- [BC97] M. Bensaid J. Schoentgen F. Bucella and S. Ciocea. Estimation by means of wavelet analysis of the signal to noise ratio of dysphonic voice. *ProRISC workshop on circuits, systems and signal processing*, pages 49–53, 1997.
- [Bel57] R. Bellman. Dynamic programming. *Princetopn university Press*, pages 745–748, 1957.
- [BG00] L. Bréhélin and O. Gascuel. Modèles de markov cachés et apprentissage de séquences. *Ecole thématique document et évolution*, pages 407–421, 2000.
- [Bim88] F. Bimbot. Synthèse de la parole : des segments aux règles, avec utilisation de la décomposition temporelle. *Thèse à l'ENST de Paris*, 1988.
- [BJ96] M. BRAHIM and B-H. JUANG. Signal bias removal by maximum likelihood estimation for robust téléphone speech recognition. *IEEE transaction speech and audio processing*, 4 :19–30, 1996.
- [BL98] A. Behrman C. Agresti E. Blumstein and N. Lee. Micophone and egg data from dysphonic patients : type 1, 2 and 43 signals. *Journal of voice*, 12 :249–260, 1998.

- [BM00] A. Bonneau and P. Mokhtari. Un diagnostic phonétique pour les déficiences auditives. *XXIII ème journées d'Etude sur la parole Aussois*, pages 389–392, 2000.
- [BN90] J.R. Bellagarda and D. Nahamoo. Tied mixture continuous parameter modeling for speech recognition. *IEEE Trans, Acoustic, speech and signal processing*, 38 :2033–2045, 1990.
- [Cas90] F. Casacuberta. Some relations among stochastic finite state networks used in automatic speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 12 :691–695, 1990.
- [Cer98] J. Cernocky. Speech processing using automatically derived segmental units : applications to very low rate coding and speaker verification. *Thèse à l'université Paris VI Orsay*, 1998.
- [CH92] D. Cairns, , and J.H.L. Hansen. Nonlinear analysis and detection of speech under stressed conditions. *The journal of the acoustic society of America*, 96 :3392–3400, 1992.
- [DB96] S. Bielałowicz J. Kreiman B-R. Gerrat M-S. Dauer and G-S. Berke. Comparison of voice analysis systems for perturbation measurement. *Journal of speech and hearing research*, 39 :126–134, 1996.
- [DEM01] J. DEMOULIN. Intégration des technologies du traitement des langues naturelles dans des applications technologiques. *quatrième rencontre des jeunes chercheurs*, 2001.
- [Den92] L. Deng. A generalized hidden markov model with state-conditioned trend functions of time for the speech signal. *Signal processing*, 27 :65–78, 1992.
- [DM80] Davis and Mermelstein. Comparaison of parametric representations for monosyllabic word recognition in continuous spoken sentences. *IEEE ASSP*, 28, 1980.
- [DM88a] F. Bimbot G. Chollet P. Deleglise and C. Montacie. Temporal decomposition and acoustic-phonetic decoding of speech. *IEEE ICASP*, pages 445–448, 1988.
- [DM88b] L-R. Bahl P-F. Brown P. Desouza and R. Mercer. A new algorithm for the estimation of hidden markov model parameters. *ICASSP*, pages 493–496, 1988.
- [DOM01] L. FABRICE I. IRINA F. DOMINIQUE. Adaptation mllr pour les hmms. *quatrième rencontre des jeunes chercheurs*, 2001.
- [Eme77] F. Emerard. Synthèse par diphone et traitement de la prosodie. *Thèse à l'université des langues et lettres, Grenoble*, 1977.
- [FH90] S. Feijoo and C. Hernandez. Short-term stability measures for the evaluation of vocal quality. *J. speech and hearing research*, 33 :324–334, 1990.
- [Fur86] S. Furui. Speaker independent isolated word recognizer using dynamic features of speech spectrum. *IEEE Trans, Acoustic, speech and signal processing*, 34 :52–59, 1986.

- [GL96] L. Gu and R. Liu. The application of optimazation in feature extraction of speech recognition. *International conference on signal processing*, 1996.
- [GM86] G. Chollet Y. Grenier and S-M. Marcus. Temporel decomposition and non-stationnary modeling of speech. *Eurasip*, pages 365–368, 1986.
- [GV00] B. Teston A. Ghio and F. Viallet. Evaluation objective de la dysprosodie des pathologies neurologique : critères de différenciation diagnostique et suivi longitudinal des prises en charge thérapeutiques. *XXIII ème journées d’Etude sur la parole Aussois*, pages 441–444, 2000.
- [HA95] L. Hansen and L. Arslan. Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit-card corpus. *IEEE Trans on speech audio processing*, 3 :169–184, 1995.
- [Haf94] P. Haffner. Apprentissage connexioniste et statistique global pour la reconnaissance de la parole. *Thèse à l’ENST de Paris*, 1994.
- [HC91] J.H.L. Hansen and M.A Clements. Constrained iterative speech enhancement with application to speech recognition. *IEEE transaction signal processing*, 39 :795–805, 1991.
- [HCD91] B. Merialdo H. Cerf-Danaon, M. El-Bèze. Reconnaissance automatique de la parole. *Informatique et santé*, 4 :84–100, 1991.
- [Her90] Hermansky. Perceptual linear predective (plp) analysis of speech. *Journal acoustique*, 87 :1738–1752, 1990.
- [HH96] J. Hillenbrand and R-A. Houde. Acoustic correlates of breathly vocal quality : Dysphonic voices and continous speech. *Journal of speech and hearing research*, 39 :311–321, 1996.
- [HK71] M-H-L. Hecker and M-J. Kreul. Description of the speech of patients with cancer of the vocal folds. part i : measures of fundamental frequency. *Journal of the acoustical society of America*, 49 :1275–1282, 1971.
- [HL00] R. Boite H. Boulard T. Dutoit J. Hancoq and H. Leich. Traitement de la parole. *Presses polytechniques et universitaires romandes*, 2000.
- [HM99] Y. Qi R-E. Hillman and C. Milstein. The estimation of signal to noise ratio in continous speech for disordered voices. *Journal of the acoustic society of america*, 105 :537–543, 1999.
- [Hol75] J-H. Holland. Adaptation in natural and artificial systems. *University of Michigan Press*, 1975.
- [HS87] I-R. Titze Y. Horii and R-C. Scherer. Some technical consideration in voice perturbation measurements. *Journal of speech and hearing research*, 30 :252–260, 1987.
- [JM95] C. Mokbel D. Jouvét and J. Monne. Blind equalization using adaptative filtering for improving speech recognition over telephone. *Eurospeech’95*, pages 1987–1990, 1995.
- [Jou88] D. Jouvét. Reconnaissance de mots connectés indépendant du locuteur par des méthodes statistiques. *Thèse à l’ENST de Paris*, 1988.

- [JR90] B-H. Juang and L-R. Rabiner. The segmental k-means algorithm for estimating parameters of hidden markov models. *IEEE transactions on acoustic, speech, and signal processing*, 38 :16–39, 1990.
- [KBB92] S. Kadambe and G.F. Boudreaux-Bartels. Application of the wavelet transform for pitch detection of speech signals. *IEEE Transaction on information theory*, 38 :917–924, 1992.
- [KK88] M. Hirano S. Hibi T. Yoshida H. Kasuya and Y. Kikuchi. Acoustic analysis of pathological voice, some results of clinical application. *Acta Otolaryngologica*, 105 :432–438, 1988.
- [Kro94] G. De Krom. Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of speech and hearing research*, 37 :985–1000, 1994.
- [LC92] F. Bimbot L. Mathan A. De Lima and G. Chollet. Standard and target driven ar-vector models for speech analysis and speaker recognition. *IEEE ICASP*, 1992.
- [LC98] Y-F. Liao and S-H. Chen. An mrnn-based method for continuous mandarin speech recognition. *ICASSP'98*, pages 1121–1124, 1998.
- [Lee88] K.F Lee. Large vocabulary speaker independent continuous speech recognition. *PHD thesis, Carnegie Mellon university*, 1988.
- [LG89] A. De Lima and Y. Grenier. Identification des cibles spectrales par modèles à excitation multi-échelle. *GRETSI*, 1989.
- [Lip97] R-P. Lippmann. Speech recognition by machines and humans. *Speech communication*, 22 :1–15, 1997.
- [LM98] K. Laurila and M. Mettala. Noise robust voice activated dialling. *AVIOS'98*, pages 64–76, 1998.
- [Lok99] M.N Lokbani. La reconnaissance de la parole ... 20 ans après. *S.White, la recherche*, 1999.
- [Mas00] S. Masaki. The speech signal and its production model. *Handbook of neural network for speech processing*, 2000.
- [MC97] Magrin-Chagnolleau. Approches statistiques et filtrage vectoriel de trajectoires spectrales pour l'identification du locuteur indépendante du texte. *Thèse à l'ENST de Paris*, 1997.
- [MG02] J. Colomer J. Melendez and F. Gamero. Pattern recognition based on episodes and dtw. application to diagnosis of a level control system. *Sixteenth International Workshop on Qualitative Reasoning*, 2002.
- [Mok01] C. Mokbel. Online adaptation of hmms to real-life conditions : a unified framework. *IEEE Trans. on speech and audio processing*, 9 :342–357, 2001.
- [MR86] J. Laver S. Hiller J. Mackenzie and E. Rooney. An acoustic screening system for the detection of laryngeal pathology. *Journal of phonetics*, 14 :517–524, 1986.



- [NH00] C-T. Lin H-W. Nein and J-Y. Hwu. Ga-based noisy speech recognition using two dimensional cepstrum. *IEEE trans. on speech and audio processing*, 8 :664–675, 2000.
- [NT00] M. Lalain D. Demolin M. Habib N. Nguyen and B. Teston. Particularités articulatoires de la dyslexie développementale phonologique. *XXIII ème journées d’Etude sur la parole Aussois*, pages 405–408, 2000.
- [PJ96] V. Parsa and D-G. Jamieson. Acoustic discrimination of pathological voice : sustained vowels versus continous speech. *Journal of speech and hearing research*, 44 :337–339, 1996.
- [Pye94] D. Pye. Automatic recognition of continous speech and keyword-spotting. *Technical report, IDIAP*, 1994.
- [QZ03] X. Qi and T. Zhang. Isolated word speech recognition of the chinese digits. *Automatic speech processing*, 2003.
- [Rab89a] L-R. Rabiner. A tutorial on hidden markov model ans select application in speech recognition. *The proceding of the IEEE*, 77 :257–285, 1989.
- [Rab89b] L-R. Rabiner. A tutorial on hidden markov model ans select application in speech recognition. *The proceding of the IEEE*, 77 :257–285, 1989.
- [RH00] S. De Martino R. Espesser V. Rey and M. Habib. Dyslexie et déficit du traitement temporel : relation entre jugement d’ordre et durée des sons de parole. *XXIII ème journées d’Etude sur la parole Aussois*, pages 393–396, 2000.
- [SC97] M. Bensaid J. Schoentgen and S. Ciocea. Estimation of the formant frequencies by means of a wavelet transform of the speech spectrum. *Pro-RISC and IEEE-Benelux workshop on circuits, systems and signal processing*, pages 49–54, 1997.
- [SC00] G. Gravier M. Sigelle and G. Chollet. A markov random field model for automatic speech recognition. *ICPR’00*, pages 3258–3262, 2000.
- [Sch82] J. Schoentgen. Quantitative evaluation of the discriminaotion performance of acoustic features in detecting laryngeal pathology. *Speech communication*, 1, 1982.
- [Sch01] J. Schoentgen. Stochastic models of jitter. *Journal of the acoustical society of america*, 109 :1631–1650, 2001.
- [Spa99] A. Spalanzani. Algorithmes évolutionnaires de la robustesse des systèmes de reconnaissance automatique de la parole. *Thèse à L’université Joseph Fourier Grenoble 1*, 1999.
- [SS78] H. Sakoe and S. Shiba. Dynamic programming algorithm optimazation for spoken word recognition. *IEEE transactions on acoustic, speech and signal processing*, pages 143–165, 1978.
- [SW70] L-E. Baum T. Petrie G. Soules and N. Weiss. A maximisation technique occuring in statistical analysis of probabilistic functions in markov chains. *The annals of mathematical statistics*, 41 :164–171, 1970.

- 
- [Teb95] J. Tebelskis. Speech recognition using neural networks. *PHD thesis, Carnegie Mellon university*, 1995.
- [TK92] H. Szu B. Telfer and S. Kadambe. Neural network adaptative wavelets for signal representation and classification. *Journal of optical engineering*, 31 :1907–1916, 1992.
- [VM90] A-P. Varga and R-K. Moore. Hidden markov model decomposition of speech and noise. *ICASSP'90*, pages 845–863, 1990.
- [WH96] E-J. Wallen and J.H.L. Hansen. An objective quality measure based screening test for speech pathology. *Inter conference on spoken language processing*, 2 :776–779, 1996.
- [Zha97] Y. Zhang. A robust and fast endpoint detection algorithm for isolated word recognition. *IEEE conference*, 1997.