THÈSE
présentée pour obtenir le titre de Docteur
de l'Université de Paris VI

Spécialité: INFORMATIQUE

Sujet:

RECONNAISSANCE DE LA PAROLE CONTINUE: ADAPTATION AU LOCUTEUR ET CONTRÔLE TEMPOREL DANS LES MODÈLES DE MARKOV CACHÉS

Présentée par:

Claude BARRAS

Soutenue le 29 mai 1996 devant le jury composé de

Régine ANDRÉ-OBRECHT	Rapporteurs
Renato DE MORI	
Marie-José CARATY	Examinateurs
Marc EL-BÈZE	
Patrick GALLINARI	
Claude MONTACIÉ	
Xavier RODET	Directeur de Thèse

SOMMAIRE

INTRODUCTION

La reconnaissance automatique de la parole par les machines est depuis longtemps un thème de recherche qui fascine le public, mais qui demeure un défi pour les spécialistes. Pour le grand public, un archétype de la communication homme-machine reste probablement le dialogue avec l'ordinateur HAL mis en scène par Stanley Kubrick dans le film *2001, l'Odyssée de l'Espace* (1968): le dialogue en langage naturel avec une machine aussi intelligente que l'homme semble l'aboutissement normal des progrès technologiques. Dans le domaine de la recherche, il était possible d'imaginer à l'époque des progrès rapides des systèmes de reconnaissance automatique de la parole (RAP), en dépit de l'opinion sceptique de quelques spécialistes [Pierce, 1969]. Mais le projet de "compréhension de la parole" lancé en 1971 par le département de la défense américaine (ARPA Speech Understanding Project) [Klatt, 1977] a contraint les chercheurs à tempérer leur optimisme: malgré des directions prometteuses, le sujet nécessite un effort à long terme.

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

Notre étude s'intègre dans le cadre du développement d'un système de dictée vocale indépendant du locuteur [Caraty *et al.*, 1996]. La modélisation acoustique par les méthodes les plus performantes de l'état de l'art reste insuffisante; cette faiblesse est un facteur limitant des systèmes de RAP. Nous cherchons à améliorer la qualité de la modélisation acoustique, en intégrant certains traitements adaptés à un processus de décodage de la parole continue. Nous avons étudié en particulier les problèmes de l'adaptation d'un système de reconnaissance à un nouveau locuteur et d'une meilleure modélisation de la durée des événements acoustiques.

Le chapitre 1 présente les méthodes classiques employées en reconnaissance de la parole. Les difficultés rencontrées pour la mise au point des systèmes de RAP proviennent de la variabilité du signal de parole et de la continuité du processus de production. Parmi les méthodes développées, l'approche statistique par modèles de Markov cachés (HMM) semble la plus efficace. Elle est appliquée à une représentation centi-seconde du signal sous forme de vecteurs de paramètres dérivés d'une analyse spectrale. Le formalisme des HMM est présenté, ainsi que les algorithmes d'apprentissage et de décodage de la parole continue. Le continuum acoustique est modélisé comme une succession d'unités acoustiques courtes, typiquement des phonèmes. En intégrant une modélisation des niveaux lexicaux et syntaxiques, il devient possible de réaliser un système de dictée vocale à grand vocabulaire et indépendant du locuteur. Cependant, le développement de cette modélisation n'a pas été l'objet principal de notre étude, et le système de reconnaissance est évalué principalement par des expériences de décodage acoustico-phonétique.

Les HMM sont basés sur des hypothèses très simplificatrices, et nous cherchons à améliorer un système de reconnaissance standard en y intégrant des traitements adaptés. Nous avons décidé de réaliser un système de référence afin d'évaluer de manière objective la pertinence de certains traitements. Le système de référence que nous présentons dans le chapitre 2 est mis au point avec la plate-forme logicielle HTK [Young, 1992a] distribuée par l'Université de Cambridge, sur la base de données anglo-américaine TIMIT. Un système de reconnaissance de base est tout d'abord réalisé; il est constitué de modèles phonétiques indépendants du contexte. Le développement de ce système nous permet de valider un certain nombre de choix concernant la représentation du signal ou la topologie des modèles phonétiques. Un système de référence plus performant est ensuite construit à partir de ce système de base, en modélisant l'influence du contexte phonétique. Une procédure originale est proposée afin d'améliorer l'apprentissage des densités de probabilité associées aux modèles. Le système obtenu obtient des performances comparables à celles des meilleurs systèmes de reconnaissance testés sur cette base de données.

Le système de référence nous permet d'évaluer l'intérêt de certains traitements proposés afin d'améliorer la reconnaissance. La réduction de performances entre les systèmes mono-locuteur et les systèmes indépendants du locuteur pose le problème de l'adaptation au locuteur. Dans le chapitre 3, nous présentons les méthodes d'adaptation classiques, qui peuvent être par exemple supervisées, rapides, incrémentales… Nous nous intéressons à l'adaptation rapide d'un système de RAP à un nouveau locuteur, qui doit être réalisée au bout de quelques secondes de parole seulement. Nous proposons une classification des locuteurs d'apprentissage au moyen d'une distance inter-locuteurs initialement dédiée à l'identification des locuteurs. Pour un locuteur de test donné, une adaptation rapide consiste à combiner l'apprentissage réalisé sur les classes de locuteurs les plus proches au sens de la distance utilisée. Cette approche originale généralise l'adaptation par multi-modèles, qui consiste à choisir les modèles d'une seule classe de locuteurs. Nos expériences mettent en évidence la pertinence de la distance étudiée d'un point de vue acoustique.

Le deuxième point sur lequel nous avons cherché à améliorer notre système de référence concerne le contrôle temporel des modèles, abordé dans le chapitre 4. Certaines hypothèses faites lors de l'utilisation des modèles de Markov cachés ne sont pas valides en parole, et entraînent une modélisation peu vraisemblable de la durée de production des unités acoustiques. Des modèles plus réalistes existent, mais le surcoût de calcul nécessaire à leur mise en œuvre est généralement dissuasif. De plus, ils sont souvent basés sur une statistique *a priori*. Pourtant, des informations présentes dans le signal de parole peuvent être mises en évidence par des méthodes d'analyse spécialisées, comme l'analyse par ruptures de modèles qui détecte des événements infra-phonémiques. Nous avons proposé l'intégration de cette source d'information pour contraindre le décodage de la parole continue. Cette intégration est réalisée de manière naturelle en restant dans un cadre probabiliste, et consiste en une modification dynamique des probabilités de transition entre les unités acoustiques. Nous montrons l'utilité d'un tel contrôle au niveau phonétique lorsque la modélisation du niveau lexical est prise en compte.

En conclusion, nous avons développé le module acoustique d'un système de reconnaissance de la parole continue. Ce système est représentatif des meilleures techniques disponibles dans la discipline. Nous avons ensuite cherché à pallier certaines de ses faiblesses en explorant deux directions particulières: l'adaptation au locuteur et le contrôle temporel. De nombreuses études restent nécessaires pour apporter une meilleure robustesse à la modélisation acoustique. Le module développé ne prendra tout son sens qu'avec l'intégration d'un modèle de langage performant et des expériences sur des corpus plus importants, illustrant de manière plus complète toute la richesse et la variabilité de la parole.

1. Reconnaissance de la parole

La conception d'un système de reconnaissance automatique de la parole est rendue difficile par la complexité du signal de parole. La production de la parole est un processus continu, et l'identification univoque d'unités symboliques dans ce flux n'est pas toujours possible. Différentes approches ont été développées pour réaliser la reconnaissance de parole. L'identification de mots isolés en vocabulaire limité et pour un unique locuteur est un problème bien maîtrisé grâce à la technique de l'alignement temporel dynamique. Les méthodes les plus performantes actuellement sont des méthodes statistiques utilisant le formalisme des modèles de Markov cachés. Elles rendent concevable la reconnaissance de parole continue à grand vocabulaire et indépendamment du locuteur. Dans le cadre de ces méthodes, le signal de parole est classiquement représenté comme une suite de vecteurs de paramètres calculés de manière synchrone; les représentations les plus efficaces sont des représentations spectrales tenant compte de certaines connaissances acquises sur la production et la perception de la parole.

Nous présentons dans la suite de ce chapitre les problèmes posés par la reconnaissance de la parole, les recherches qui ont contribué aux progrès de la discipline, les méthodes d'analyse du signal choisies pour une bonne représentation de la parole et les techniques de reconnaissance les plus utilisées actuellement. Les algorithmes d'apprentissage et de reconnaissance exploitant les modèles de Markov cachés sont détaillés, dans l'objectif de réaliser un système de dictée vocale à grand vocabulaire. Notre étude concerne principalement la modélisation acoustique du signal de parole par des unités plus courtes que le mot; mais la modélisation des niveaux supérieurs du langage, à savoir les niveaux lexical, syntaxique ou sémantique, est indispensable à la réalisation d'un système complet. Parmi les articles ou les ouvrages ayant servi de base à cette présentation, nous renvoyons en particulier à [Calliope, 1988; Mariani, 1990; Haton *et al.*, 1991] qui peuvent être consultés pour plus de détails et qui sont accessibles à tous les francophones…

1.1. Introduction et méthodes de reconnaissance

Les difficultés rencontrées lors de la conception d'un système de reconnaissance automatique de la parole sont dues à la complexité intrinsèque du signal de parole mais aussi au niveau des performances demandées au système. L'identification de mots isolés est bien résolue par l'alignement temporel dynamique, et le décodage de la parole continue a pendant longtemps été abordé par une approche dite analytique. L'usage de méthodes statistiques globales est maintenant généralisé en raison de leur plus grande souplesse.

1.1.1. Complexité du signal de parole

La complexité du signal de parole provient de la combinaison de plusieurs facteurs, principalement la redondance du signal acoustique, la grande variabilité intra et inter-locuteurs, et les effets de la coarticulation en parole continue, qui doivent être pris en compte lors de la conception d'un système de RAP.

1.1.1.1. Redondance

Le signal acoustique présente, dans le domaine temporel, une redondance qui rend indispensable un traitement préalable à toute tentative de reconnaissance. Il existe en effet une grande disproportion entre le débit du signal enregistré et la quantité d'information cherchée pour une tâche de reconnaissance. Un signal échantillonné à 16 kHz sur 16 bits représente un débit de 256 kbit/s, alors qu'une tâche de reconnaissance phonétique recherche typiquement une dizaine de phonèmes à la seconde, soit une compression de près de 10^4 du débit initial. Ce constat motive la recherche d'une représentation plus compacte du signal. Il existe un grand nombre de paramètres possibles, souvent dérivés d'une analyse spectrale ou auto-régressive. Le choix de paramètres adaptés à un type de problème dépend des conditions d'enregistrement du signal et surtout de l'usage ultérieur qui doit en être fait, puisque ce qui n'est pas pertinent vis-à-vis du contenu lexical peut le devenir pour l'identification du locuteur. Le choix de cette représentation ne résout pas les difficultés provenant de la continuité de la production et plus généralement de la variabilité de la parole.

1.1.1.2. Continuité et coarticulation

Tout discours peut être retranscrit par des mots, qui peuvent à leur tour être décrits comme une suite de symboles élémentaires appelés phonèmes par les linguistes. Cela laisse supposer que la parole est un processus séquentiel, au cours duquel des unités indépendantes se succèdent. Malheureusement, les spécialistes de phonétique eux-mêmes ont parfois des difficultés à identifier individuellement ces unités discrètes dans le signal, même si quelques événements acoustiques particuliers peuvent être détectés. La parole est en réalité un flux continu, et il n'existe pas de pause entre les mots qui pourrait faciliter leur localisation automatique par les systèmes de reconnaissance.

De plus, les contraintes introduites par les mécanismes de production créent des phénomènes de coarticulation. La production d'un son est fortement influencée par les sons qui le précèdent mais aussi qui le suivent en raison de l'anticipation du geste articulatoire. Ces effets s'étendent sur la durée d'une syllabe, voire même au-delà, et sont amplifiés par une élocution rapide. L'identification correcte d'un segment de parole isolé de son contexte est parfois impossible. La prise en compte des phénomènes de coarticulation ne suffit pourtant pas à prédire la réalisation acoustique d'une phrase en raison de la grande variabilité de la parole.

1.1.1.3. Variabilité

On distingue généralement deux sources de variabilité qui peuvent rendre deux prononciations d'un même énoncé très différentes: la variabilité inter-locuteurs et la variabilité intra-locuteur. La variabilité inter-locuteurs est *a priori* la plus importante. Les différences physiologiques entre locuteurs, qu'il s'agisse de la longueur du conduit vocal ou du volume des cavités résonantes, modifient la production acoustique; à cet égard, les voix d'hommes, de femmes et d'enfants sont les plus caractéristiques. A cela s'ajoutent les habitudes acquises en fonction du milieu social et géographique, comme les accents régionaux. La variabilité intra-locuteur est plus réduite, mais n'est pas négligeable. L'état physique, par exemple la fatigue ou le rhume, les conditions psychologiques, comme le stress, et même le bruit de fond pendant l'élocution, influent sur la production et sur la prosodie, entraînant des variations complexes du rythme d'élocution, de la mélodie et de l'intensité du discours. Ces phénomènes sont encore mal modélisés et compliquent la réalisation des systèmes de reconnaissance. J.-S. Liénard défend l'idée d'une prise en compte de toutes les caractéristiques du locuteur pour la représentation de la parole [Liénard, 1989].

1.1.2. Tâche de reconnaissance

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

- le nombre d'utilisateurs du système: celui-ci peut être mono-locuteur, multi-locuteurs, ou indépendant du locuteur;
- la taille du vocabulaire: petit vocabulaire (moins de mille mots), grand vocabulaire (moins de cent mille mots) ou très grand vocabulaire (plus de cent mille mots);
- la complexité du langage utilisé: langage contraint par une syntaxe artificielle ou langage naturel;
- le mode d'élocution: mots isolés ou parole continue;
- la robustesse aux conditions d'enregistrement: système nécessitant de la parole de bonne qualité ou fonctionnant en milieu bruité.

Les systèmes les plus simples et les plus anciens concernent l'identification de mots isolés, en petit vocabulaire, éventuellement en mode multi-locuteurs. Un tel problème est assez bien résolu par des techniques d'alignement temporel dynamique [Rabiner *et al.*, 1979], et il existe des processeurs spécialisés réalisant cette tâche.

Avec des méthodes statistiques à base de modèles de Markov cachés, il est concevable de réaliser une reconnaissance de la parole continue indépendamment du locuteur, en grand vocabulaire, pour un enregistrement de bonne qualité et un langage artificiel [Lee, 1988]. En qualité téléphonique, les performances ne permettent pour le moment que des applications avec de petits vocabulaires [Gagnoulet & Jouvet, 1990].

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

1.1.3. Méthodes de reconnaissance

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

1.1.3.1. Approche analytique

L'approche analytique cherche à résoudre le problème de la parole continue en isolant des unités acoustiques courtes comme les phonèmes, les diphonèmes ou les syllabes. Un exemple classique de cette approche est l'analyse par traits: des indices acoustiques sont calculés à partir du signal de parole; ils permettent de faire des hypothèses locales sur certains traits phonétiques, comme le voisement, la nasalisation, le lieu d'articulation ou le degré d'ouverture du conduit vocal. En fonction de ces traits, le signal acoustique est segmenté et une identification phonétique des segments est réalisée. Le décodage acoustico-phonétique ainsi obtenu est exploité par des modules d'ordre linguistique. Les niveaux lexical, syntaxique ou sémantique utilisent des sources de connaissances spécialisées et sont organisés avec le module acoustique dans des architectures montantes ou descendantes (cf. [Haton *et al.*, 1991], chapitre 2). Les systèmes analytiques, conçus avec des objectifs ambitieux, sont restés au stade expérimental. Leur faiblesse provient d'un processus de décision trop précoce, à savoir une segmentation préalable à l'identification ou une identification phonétique sans prise en compte des niveaux linguistiques. Les méthodes globales, développées pour la reconnaissance de mots isolés, ne font pas d'hypothèse sur la structure phonétique des mots, ce qui évite une erreur pénalisante au début du traitement.

1.1.3.2. Approche globale

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

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

$$\tilde{m} = \arg\min_{m \in E_m} D(O, R_m) \tag{1.1}$$

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

L'alignement temporel dynamique (*Dynamic Time Warping* ou DTW en anglais) résout efficacement ce problème, en exploitant le principe d'optimalité de Bellman [Bellman, 1957]. La construction de l'alignement optimal entre les formes de référence et de test est réalisée par récursivité sur l'indice du temps, en exploitant le fait que le chemin optimal est l'extension d'un sous-chemin lui-même optimal. La complexité de l'alignement est de ce fait considérablement réduite puisqu'elle passe d'un ordre exponentiel à un ordre polynomial. Les premières applications de la programmation dynamique en parole sont développées en URSS en 1968 [Vintsiuk, 1968; Velichko, 1970], puis au Japon à partir de 1970 [Sakoe & Chiba, 1978]. La méthode est efficace pour la reconnaissance mono-locuteur à petit vocabulaire et en mots isolés. Des extensions de l'alignement temporel dynamique ont été proposées pour la reconnaissance indépendante du locuteur [Rabiner *et al.*, 1979] ou la reconnaissance de mots enchaînés [Sakoe, 1979; Myers & Rabiner, 1981, Bridle *et al.*, 1982]. Cependant, l'approche statistique propose un formalisme plus général et permet la reconnaissance de grands vocabulaires en parole continue de manière plus efficace que par DTW en intégrant la modélisation des niveaux linguistiques.

1.1.3.3. Approche statistique

F. Jelinek a proposé une formalisation statistique simple issue de la théorie de l'information et qui est aujourd'hui classique pour décomposer le problème de la reconnaissance de la parole continue [Jelinek, 1976]. Soit O une suite d'observations acoustiques, et M une suite de mots prononcés. Connaissant les observations O, on cherche la suite de mots $\tilde{M}$ la plus probable parmi toutes les suites possibles $E_M=E_m{}^*$, soit:

$$\tilde{M} = \underset{M \in E_M}{\arg\max}\, P(M/O) \tag{1.2}$$

En utilisant la règle de Bayes, il est possible d'écrire la probabilité qu'une suite de mots corresponde aux observations acoustiques comme:

$$P(M/O) = \frac{P(O/M).P(M)}{P(O)} \tag{1.3}$$

Puisque $P(O)$ ne dépend pas de M, l'équation (1.2) est équivalente à:

$$\tilde{M} = \underset{M \in E_M}{\arg\max}\, P(O/M).P(M) \tag{1.4}$$

où

- $P(O/M)$, la probabilité des observations O, étant donnée la suite de mots M, est estimée par une modélisation acoustique;
- $P(M)$, la probabilité *a priori* de la suite de mots M, est estimée par un modèle linguistique.

L'approche statistique permet ainsi d'intégrer les niveaux acoustiques et linguistiques dans un seul processus de décision. Ces niveaux sont classiquement représentés par des modèles de Markov cachés (*Hidden Markov Models* ou HMM). Les unités acoustiques modélisées peuvent être des mots comme dans l'approche globale, ou des unités plus courtes telles que le phonème comme dans l'approche analytique. La modélisation markovienne est plus générale que l'alignement temporel dynamique et tient compte non seulement de la non linéarité temporelle du processus mais aussi de la variabilité acoustique de la production de la parole. Son application à la reconnaissance de la parole continue a été rendue possible par l'augmentation continue de la puissance des ordinateurs et de la taille des bases de données disponibles.

Les chercheurs de CMU (Carnegie Mellon University) et d'IBM sont les premiers à avoir introduit le formalisme des modèles de Markov cachés en reconnaissance de la parole [Baker, 1975; Jelinek, 1976]. Au cours des dix dernières années, les systèmes des plus grands laboratoires internationaux travaillant en RAP, comme les systèmes SPHINX de CMU [Lee, 1988], BYBLOS de BBN (Bolt Beranek and Newman Inc.) [Chow *et al.*, 1987], TANGORA d'IBM [Averbuch *et al.*, 1987], ou ceux développés à AT&T [Wilpon *et al.*, 1993], ont été conçus avec une approche statistique markovienne. Cette approche a aussi été appliquée avec succès en France au CNET [Jouvet, 1994] ou au LIMSI qui obtient des performances équivalentes à celles des meilleurs systèmes actuels [Gauvain *et al.*, 1994]. Nous présentons tout d'abord les espaces de représentation adaptés au signal de parole, dont le choix est un préalable à la conception de tout système de RAP markovien; puis nous exposons la théorie des modèles de Markov cachés et leur application en reconnaissance de la parole continue.

1.2. Espaces de représentation

Nous avons constaté que le signal de parole présente de la redondance et contient des informations jugées superflues pour la reconnaissance, ce qui justifie la recherche d'une représentation plus compacte (cf. § 1.1.1). La représentation du signal est généralement déduite d'une analyse à court terme. Les trames successives du signal sont analysées par Transformée de Fourier Discrète, par un modèle de prédiction linéaire ou par transformation cepstrale [Rabiner & Schafer, 1978]. Enfin, il est nécessaire d'associer une distance adaptée à cet espace de représentation. Ainsi, les coefficients cepstraux associés à une distance euclidienne pondérée sont largement utilisés en reconnaissance de la parole depuis quinze ans [Davis & Mermelstein, 1980].

1.2.1. Représentations non paramétriques

Le signal de parole peut être analysé dans le domaine temporel ou dans le domaine spectral par des méthodes non paramétriques, sans faire l'hypothèse d'un modèle pour rendre compte du signal observé. Les représentations le plus souvent retenues sont l'énergie du signal et les sorties d'un banc de filtres numériques.

1.2.1.1. Chaîne de pré-traitement

Le calcul de la représentation du signal est réalisé par une chaîne d'analyse presque complètement numérique en suivant les étapes suivantes: échantillonnage, pré-accentuation, puis analyse à court terme sur des trames successives du signal.

La fréquence d'échantillonnage est typiquement de 8 kHz pour de la parole de qualité téléphonique et de 16 à 20 kHz pour la parole de bonne qualité. Le signal échantillonné est ensuite pré-accentué pour relever les hautes fréquences qui sont moins énergétiques que les basses fréquences; la pré-accentuation s'_n de l'échantillon s_n à l'instant n est calculée pour une valeur de α comprise entre 0,9 et 1 comme:

$$s'_n = s_n - \alpha . s_{n-1} \tag{1.5}$$

La plupart des représentations utilisées font l'hypothèse d'une stationnarité du signal, ce qui n'est pas valide dans le cas de la parole. Une analyse à court terme est donc réalisée sur une fenêtre glissante, chaque trame couvrant une durée de 20 à 50 ms sur laquelle le signal est supposé quasi-stationnaire, pour un pas d'analyse entre deux trames successives de l'ordre de la centi-seconde. Le découpage du signal en trames produit des discontinuités aux frontières des trames, qui se manifestent par des lobes secondaires dans le spectre; ces effets parasites sont réduits en appliquant aux échantillons $\{s'_n\}_{n=0\ldots N-1}$ de la trame une fenêtre de pondération $\{w_n\}_{n=0\ldots N-1}$ comme par exemple la fenêtre de Hamming [Harris, 1978]:

$$s''_n = w_n s'_n$$

avec

$$w_n = 0,54 - 0,46.\cos(2\pi \frac{n}{N-1}),\ 0 \leq n \leq N\text{-}1 \tag{1.6}$$

Par la suite, les échantillons pré-traités s''_n seront notés simplement s_n. L'analyse de chaque trame peut être réalisée directement dans le domaine temporel, par exemple pour calculer l'énergie, ou plus couramment dans le domaine spectral.

1.2.1.2. Analyse temporelle

L'énergie du signal est un indice qui peut par exemple contribuer à la détection du voisement d'un segment de parole. L'énergie totale E_0 est calculée directement dans le domaine temporel sur une trame de signal $\{s_n\}_{0 \leq n \leq N\text{-}1}$ comme:

$$E_0 = \sum_{n=0}^{N-1} {s_n}^2 \tag{1.7}$$

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

D'autres paramètres peuvent être calculés dans le domaine temporel, comme les coefficients d'auto-corrélation, le taux de passage par zéro, ou encore la fréquence fondamentale. L'estimation des coefficients d'auto-corrélation $\{\hat{r}_k\}$ est calculée par:

$$\hat{r}_k = \sum_{n=k}^{N-1} s_n s_{n-k}, \quad 0 \leq k \leq N-1 \tag{1.8}$$

Ces coefficients sont utilisés dans le cadre de la modélisation auto-régressive (cf. § 1.2.2). Cependant, la production de la parole rend souhaitable une analyse du signal dans le domaine spectral pour la reconnaissance.

1.2.1.3. Analyse spectrale

Une analyse spectrale du signal permet de mettre en évidence certains phénomènes caractéristiques de la production de la parole qui peuvent contribuer à l'identification phonétique. En effet, la génération d'une énergie ventilatoire est à l'origine de plusieurs sons: un son quasi-périodique produit par les vibrations des cordes vocales, un bruit de friction dû à des constrictions du conduit vocal ou un bruit d'explosion par le relâchement d'une occlusion. Le signal de parole, provenant de la combinaison de ces différents sons, entre en résonance lors de la traversée des cavités orales et nasales. L'articulation des phonèmes a une influence directe sur la forme du conduit vocal et des cavités, et donc sur les résonances qui apparaissent dans l'enveloppe du spectre.

Des spectrogrammes ont été utilisés pour représenter la parole dès les années 40 [Koenig *et al.*, 1946], en utilisant des bancs de filtres analogiques. Actuellement, les spectres sont obtenus numériquement par Transformée de Fourier Discrète, en particulier grâce à l'algorithme de la Transformée de Fourier Rapide (FFT) [Cooley & Tukey, 1965]. Le spectre à court terme $\{S_k\}_{k=0\ldots N-1}$ est calculé à partir des N échantillons $\{s_n\}_{n=0\ldots N-1}$ comme:

$$S_k = \sum_{n=0}^{N-1} s_n e^{-j2\pi nk/N}, \ 0 \leq k \leq N\text{-}1 \tag{1.9}$$

L'intensité en décibels du spectre est directement visualisable sous la forme d'un spectrogramme pour une évaluation qualitative du signal.

Le nombre de paramètres spectraux calculés sur une trame par FFT reste trop élevé pour un traitement automatique ultérieur. L'énergie du spectre est calculée à travers un banc de filtres numériques couvrant la bande passante, ce qui permet de ne conserver qu'une vingtaine de valeurs d'énergie par exemple sur une bande passante de 8 kHz. Des filtres triangulaires sont préférés pour leur simplicité et leur effet de lissage sur le spectre. Il n'est pas nécessaire que ces filtres soient répartis sur une échelle linéaire; l'échelle Mel, par exemple, est linéaire en basses fréquences et logarithmique en hautes fréquences. L'intérêt de l'échelle Mel est d'être assez proche d'échelles issues d'études sur la perception sonore et sur les bandes passantes critiques de l'oreille, comme l'échelle Bark [Zwicker, 1961], et elle est très largement utilisée par les chercheurs. Les énergies en sortie du banc de filtre peuvent être utilisées directement comme paramètres d'un système de reconnaissance, comme dans le système d'IBM [Bahl *et al.*, 1988]. Dans les systèmes les plus récents, une transformation cesptrale est systématiquement appliquée à ces énergies.

1.2.2. Représentations paramétriques

Les méthodes précédentes ne font pas d'hypothèse sur la nature du signal analysé. Si l'on dispose d'un modèle adapté au signal de parole, l'estimation des paramètres du modèle doit permettre une meilleure caractérisation du signal. L'analyse par prédiction linéaire et l'analyse cepstrale utilisent un modèle de production. Des modèles issus d'études sur la perception sont aussi développés dans les laboratoires de recherche.

1.2.2.1. Analyse par prédiction linéaire

L'analyse par prédiction linéaire est une alternative intéressante au calcul du spectre à court terme pour plusieurs raisons: elle est basée sur un modèle de production simple mais cohérent, et ce modèle est facilement inversible, c'est-à-dire qu'il est possible de calculer les paramètres du modèle à partir du signal avec des algorithmes rapides. Le cadre théorique est celui d'une modélisation source-filtre. Le conduit vocal est modélisé comme une succession de tubes de sections cylindriques qui sont des résonateurs. En tenant compte des autres facteurs participant à la production comme les cavités nasales résonantes et le rayonnement des lèvres, le filtre est modélisé par un modèle auto-régressif à moyenne mobile (ARMA), dont la transmittance dans l'espace z est une fraction rationnelle. L'estimation des paramètres d'un modèle ARMA est délicate. Il est préférable de simplifier encore ce modèle et de modéliser le filtre comme un filtre linéaire tout-pôles qui produit un signal auto-régressif (AR); la théorie des développements limités montre que tout filtre ARMA peut être approximé par un filtre AR d'ordre suffisamment élevé. Soit p l'ordre du modèle, chaque échantillon s_n est prédit comme une combinaison linéaire des p échantillons précédents, à laquelle s'ajoute un bruit blanc gaussien $\{e_n\}$ de variance σ^2:

$$s_n = \sum_{i=1}^{p} a_i s_{n-i} + e_n \tag{1.10}$$

La minimisation de l'erreur de prédiction par la méthode des moindres carrés conduit aux équations de Yule-Walker qui expriment le vecteurs des coefficients $A = (1, a_1 \ldots a_p)^t$ comme:

$$R.A = (\sigma^2, 0 \ldots 0)^t \tag{1.11}$$

où la matrice d'autocorrélation du signal R est une matrice de Toëplitz constituée des p+1 premiers coefficients d'autocorrélation:

$$R = \left[R_{ij}\right]_{1 \le i,j \le p+1} \quad \text{avec} \quad R_{ij} = \hat{r}_{|i-j|} \tag{1.12}$$

Un algorithme rapide pour résoudre l'équation (1.11) et calculer les coefficients auto-régressifs $\{a_i\}_{i=1\ldots p}$ a été développé par Levinson en 1947 et modifié par Durbin en 1960 (voir par exemple [Makhoul, 1975] pour une présentation théorique). Les applications de la prédiction linéaire en analyse de la parole sont nombreuses et anciennes [Atal & Hanauer, 1971]. Une transformation directe des coefficients auto-régressifs vers les coefficients cepstraux est possible.

1.2.2.2. Transformation cepstrale

La représentation cepstrale présente l'intérêt de dissocier, dans le cadre de la théorie d'un modèle source-filtre, l'excitation glottique et les résonances du conduit vocal. Cette séparation est réalisée par un homomorphisme qui transforme la convolution des signaux dans le domaine temporel en une addition dans le domaine cepstral. Il est possible ensuite de ne conserver que les paramètres caractéristiques de l'enveloppe formantique.

Le signal observé $\{s_n\}$ est la convolution dans le domaine temporel du signal d'excitation $\{g_n\}$ et de la réponse impulsionnelle du conduit vocal $\{h_n\}$:

$$s_n = g_n * h_n \tag{1.13}$$

Dans le domaine spectral, la convolution devient un produit qui est transformé en addition par le logarithme:

$$\log|S_k| = \log|G_k| + \log|H_k| \tag{1.14}$$

où $\{S_k\}$, $\{G_k\}$ et $\{H_k\}$ sont les spectres respectifs de $\{s_n\}$, $\{g_n\}$ et $\{h_n\}$. La transformée de Fourier inverse de cette quantité donne le cepstre, qui est la somme des contributions de la source et du conduit. Les coefficients d'indice faible contiennent la quasi-totalité de l'énergie correspondant au conduit, alors que la fréquence fondamentale du locuteur fait apparaître un pic à une "quéfrence" d'indice plus élevé. Un filtrage cepstral passe-bas permet de ne conserver que la contribution du conduit. Une séparation plus fine de la source et du filtre, basée sur une estimation de la fréquence fondamentale, est possible. Une méthode itérative est détaillée dans [Montacié, 1991], et un calcul direct a été proposée dans [Galas & Rodet, 1990]. Cette technique semble surtout utile pour les voix féminines, dont le pic correspondant à la fréquence fondamentale est décalé vers les indices plus faibles.

1.2.2.3. Coefficients cepstraux

Les coefficients cepstraux peuvent être calculés à partir de la sortie d'un banc de filtres (cf. Figure 1.1) ou des coefficients de la prédiction linéaire. Ainsi, les paramètres LPCC (*Linear Prediction Cepstral Coefficients*) sont calculés à partir d'une modélisation auto-régressive du signal. Si un modèle auto-régressif $A = (1, a_1 \ldots a_p)^t$ d'ordre p a été estimé sur une trame du signal, les d premiers coefficients cepstraux C_k sont obtenus par:

$$C_k = a_k - \sum_{i=1}^{k-1} \frac{i}{k} C_i a_{k-i}, \quad 1 \le k \le d \tag{1.15}$$

Ces coefficients sont utilisés à AT&T [Rabiner, 1989] et à CMU [Lee & Hon, 1989].

Les paramètres MFCC (*Mel Frequency Cepstral Coefficients*) sont des coefficients cepstraux obtenus à partir des énergies d'un banc de filtres en échelle de fréquence Mel [Davis & Mermelstein, 1980]. Les d premiers coefficients cepstraux C_k peuvent être calculés directement à partir du logarithme des énergies E_i sortant d'un banc de F filtres par la transformée en cosinus discrète:

$$C_k = \sum_{i=1}^{F} \log E_i . \cos\left[\frac{\pi k}{F}\left(i - \frac{1}{2}\right)\right], \quad 1 \le k \le d \tag{1.16}$$

Il s'agit d'une représentation compacte; à partir d'une vingtaine de filtres en échelle Mel, une dizaine de coefficients cepstraux sont généralement considérés comme suffisants pour les expériences de reconnaissance de la parole et du locuteur. De plus, la transformée en cosinus discrète présente des similitudes avec la transformation de Karhunen-Loève et permet d'obtenir des coefficients peu corrélés [Soong & Rosenberg, 1988]. Le coefficient C_0 qui est la somme des énergies n'est pas utilisé; il est éventuellement remplacé par le logarithme de l'énergie totale E_0 calculée dans le domaine temporel et normalisée. Certains chercheurs calculent des paramètres BACC (*Bark Auditory Cepstral Coefficients*) en échelle Bark; mais les différences entre les deux échelles sont peu importantes.

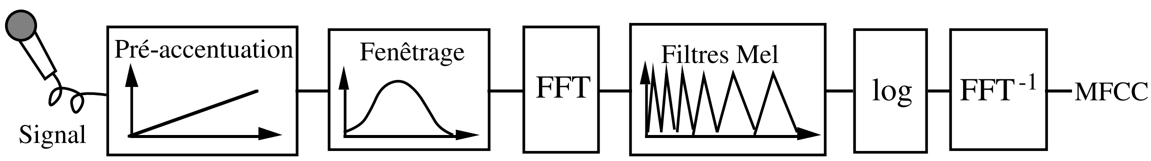

Figure 1.1. Chaîne d'analyse du signal produisant les coefficients MFCC.

1.2.2.4. Autres représentations

Les représentations spectrales, auto-régressives ou cepstrales sont les plus usuelles, mais d'autres représentations sont utilisées en traitement automatique de la parole, comme les représentations temps-fréquence [d'Alessandro & Demars, 1992]. Les représentations du signal basées sur la perception suscitent beaucoup d'espoirs. En effet, l'étude de la production des sons et de leur perception est complémentaire. Le système auditif est une référence naturelle à toute analyse acoustique automatique devant servir à la reconnaissance de la parole. Les échelles fréquentielles Mel ou Bark issues d'études psycho-acoustiques ont déjà été intégrées avec succès dans les chaînes d'analyse. Des modèles complexes d'audition et de perception existent, et peuvent éventuellement être exploités en reconnaissance automatique.

S. Seneff décrit une modélisation basée sur les propriétés du système auditif humain. Un banc de filtres en bandes critiques répartis sur l'échelle Bark constitue l'entrée d'un modèle non-linéaire de la cochlée. Une mesure de l'énergie des décharges neuronales et une mesure du synchronisme sont produites en sortie [Seneff, 1988]. H. Hermansky propose un pré-traitement par Prédiction Linéaire Perceptive ou PLP. Des filtres répartis sur une échelle Bark ont une enveloppe similaire aux courbes de masquage, et leur sortie est pondérée en fonction de la réponse auditive. Un modèle auto-régressif est ensuite estimé sur la racine cubique des énergies de sortie [Hermansky, 1990; Junqua, 1990]. L'intégration de représentations issues de modèles perceptifs dans les systèmes de reconnaissance est cependant compliquée par la nécessité de leur associer des distances pertinentes.

1.2.3. Distances et mesures de distance

Dans toute approche de reconnaissance, le choix de l'espace de représentation du signal est étroitement lié à celui de la distance définie dans cet espace. Les spectres, les coefficients de la prédiction linéaire ou les coefficients cepstraux peuvent être assimilés à des vecteurs numériques de dimension d. Il est possible d'utiliser toutes les distances classiques, en particulier les distances de Minkovski, parmi lesquelles la classique distance euclidienne, et la distance de Mahalanobis qui normalise les coefficients par leur matrice de covariance. Des distances spécifiques aux espaces de représentation de parole existent aussi, comme les distances cepstrales pondérées [Juang & Rabiner, 1987], la mesure d'Itakura pour comparer les modèles auto-régressifs [Itakura, 1975], ou encore la distance par Appariement de Pics Spectraux [Caraty, 1987].

1.2.3.1. Distances usuelles

Les distances de Minkovski L_r entre deux vecteurs $X = (x_1 \ldots x_d)^t$ et $Y = (y_1 \ldots y_d)^t$ de dimension d s'expriment:

$$L_r(X,Y) = \left(\sum_{k=1}^{d} \left| x_k - y_k \right|^r \right)^{1/r} \tag{1.17}$$

Les distances les plus courantes sont la distance de Manhattan pour r=1, la distance du "max" pour r infini, et la distance euclidienne d_E pour r=2. La distance euclidienne, en particulier, peut aussi être définie vectoriellement comme:

$$d_E(X,Y)^2 = (X-Y)^t(X-Y) \tag{1.18}$$

La distance euclidienne donne *a priori* la même importance à chacun des coefficients. Si l'on dispose d'un ensemble de n vecteurs de paramètres $\{X_i\}_{1\le i\le n}$, il est possible d'estimer leur vecteur moyen $\hat{\mu}$ comme:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.19}$$

et la matrice de covariance des coefficients $\hat{\Sigma} = [\sigma_{kl}]_{1\le k,l\le d}$ se calcule par:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^t - \hat{\mu}\hat{\mu}^t \tag{1.20}$$

La distance de Mahalanobis entre deux vecteurs de paramètres tient compte de cette statistique en normalisant le produit scalaire par l'inverse de la matrice de covariance, ce qui a pour effet de décorréler linéairement les coefficients:

$$d_M(X,Y)^2 = (X-Y)^t\hat{\Sigma}^{-1}(X-Y) \tag{1.21}$$

Dans le cas particulier où les coefficients ne sont pas corrélés, la matrice de covariance est diagonale, et cette diagonale est constituée des variances des coefficients $\sigma_i{}^2 = \sigma_{ii}$. La distance de Mahalanobis devient une distance euclidienne pondérée par l'inverse des variances des coefficients:

$$d_P(X,Y)^2 = \sum_{k=1}^{d}\left[w_k(x_k - y_k)\right]^2 \quad \text{avec} \quad w_k = \frac{1}{\sigma_k} \tag{1.22}$$

1.2.3.2. Mesures de distance adaptées à une représentation

Les distances usuelles ne sont pas nécessairement adaptées aux représentations choisies. Dans le cas de représentations paramétriques, le modèle utilisé permet parfois de déduire une mesure adaptée, comme par exemple certaines distances cepstrales pondérées, la mesure d'Itakura pour les coefficients auto-régressifs.

Une distance euclidienne pondérée par l'indice du coefficient a été proposée comme distance cepstrale [Hanson & Wakita, 1986]. Ces pondérations corrigent l'effet de la décroissance rapide des coefficients d'indice élevé; de plus, cette distance présente des similitudes avec une mesure pondérée sur la pente du spectre proposée par D. Klatt [Klatt, 1982]:

$$d_P(X,Y)^2 = \sum_{k=1}^{d} w_k{}^2(x_k - y_k)^2 \quad \text{avec} \quad w_k = k \tag{1.23}$$

Les chercheurs d'AT&T préconisent une pondération sinusoïdale des coefficients cepstraux [Juang & Rabiner, 1987]:

$$\hat{C}_k = w_k C_k \quad \text{avec} \quad w_k = \left[1 + \frac{d}{2}\sin\left(\frac{\pi k}{d}\right)\right], 1 \leq k \leq d \tag{1.24}$$

et l'utilisation d'une distance euclidienne, ce qui est équivalent à l'utilisation d'une distance euclidienne pondérée par une fenêtre sinusoïdale.

Les distances classiques ne sont pas adaptées à la comparaison de modèles auto-régressifs. Si $A = (1, a_1 \ldots a_p)^t$ et $B = (1, b_1 \ldots b_p)^t$ sont des vecteurs contenant les coefficients de prédiction de deux modèles auto-régressifs d'ordre p, la mesure d'Itakura est définie [Itakura, 1975] comme:

$$d_I(A, B) = \log\left[\frac{A^t R_b A}{B^t R_b B}\right] \tag{1.25}$$

où R_b est la matrice d'autocorrélation du signal ayant servi à l'estimation du modèle B. La mesure fait le rapport de l'erreur de prédiction du signal par un modèle de référence A et de l'erreur de prédiction par un modèle optimal calculé sur le signal lui-même. Ce rapport étant par définition plus grand que un, la mesure est positive, mais elle n'est pas symétrique.

Enfin, M.-J. Caraty a défini une distance par Appariement de Pics Spectraux (APS) qui est une mesure de la distorsion inter-spectres fondée sur des critères de perception sonore [Caraty, 1987]; cette mesure utilise une représentation du signal obtenue par une détection des pics du spectre. Par analogie avec les formants, un pic spectral P est caractérisé par sa fréquence F, sa bande passante B mesurée à 3 dB et son intensité I en dB. La mesure de distorsion δ_{APS} entre deux pics P_1 et P_2 de deux spectres différents est définie comme:

$$\delta_{APS}(P_1, P_2) = \omega_F(F_1)\frac{|F_1 - F_2|}{F_1 + F_2} + \omega_B(F_1)\frac{|B_1 - B_2|}{B_1 + B_2} + \omega_I(F_1)|I_1 - I_2| \tag{1.26}$$

où ω_F, ω_B et ω_I sont des fonctions de pondération. La mesure de distorsion globale inter-spectres est alors calculée à partir de cette mesure inter-pics en suivant un appariement des pics spectraux spécifique à APS. Des expériences utilisant des stimuli synthétiques ont permis d'évaluer le pouvoir discriminant de 17 distances pour l'identification des voyelles [Yé *et al.*, 1989]; elles ont montré le bon comportement de la mesure APS par rapport à la distance d'Itakura, et à des distances cepstrales pondérées.

1.2.4. Choix d'une représentation du signal

Le choix d'une représentation du signal parmi celles présentées est en définitive guidé par les résultats expérimentaux dans la tâche considérée. Nous présentons quelques études comparatives entre différents coefficients associés à des distances afin de justifier d'un choix acceptable dans notre système. Les coefficients LPCC ou MFCC ainsi que l'énergie sont les coefficients les plus utilisés dans les systèmes de reconnaissance. L'usage complémentaire des coefficients différentiels du premier et du second ordre s'est généralisé depuis une dizaine d'années dans les systèmes de reconnaissance à base de HMM. Enfin, des techniques d'analyse des données sont parfois proposées pour réduire la complexité des données manipulées.

1.2.4.1. Étude comparative de représentations

L'étude classique de S. Davis et P. Mermelstein [Davis & Mermelstein, 1980] compare plusieurs représentations du signal: cepstre en sortie d'un banc de filtres en échelle Mel (MFCC) ou en échelle linéaire (LFCC), coefficients de prédiction linéaire (LPC) ou de réflexion (RC), cepstre calculé à partir des coefficients auto-régressifs (LPCC). Ces représentations sont associées à la distance euclidienne, ou à la mesure d'Itakura en ce qui concerne les coefficients LPC. Les tests consistent en un alignement temporel dynamique entre des paires de mots minimales. Les MFCC donnent les meilleurs résultats. Cette étude montre plus généralement l'intérêt d'un pré-traitement par banc de filtres, d'une échelle fréquentielle non linéaire, et de la représentation cepstrale.

B. Hanson et T. Applebaum s'intéressent aux conséquences de l'effet Lombard et du bruit sur la reconnaissance de mots isolés [Hanson & Applebaum, 1990]. Ils remarquent la nette supériorité d'une distance cepstrale pondérée sur une distance euclidienne. De plus, le pré-traitement perceptif des coefficients PLP améliore les résultats de l'analyse par prédiction linéaire. J.-C. Junqua, H. Wakita et H. Hermansky aboutissent aux mêmes conclusions [Junqua *et al.*, 1993]. B. Chigier et H. Leung ont utilisé des représentations spectrales et cepstrales classiques ou dérivées du modèles d'audition de S. Seneff [Chigier & Leung, 1992]. Leurs expériences sont réalisées en enregistrement normal avec la base TIMIT et sur du signal de qualité téléphonique avec la base NTIMIT. Elles consistent en une identification phonétique par un classifieur gaussien ou un perceptron multi-couches. Le meilleur résultat est obtenu pour le perceptron avec des coefficients PLP, et pour le classifieur gaussien avec des coefficients en échelle Bark. Les critères de classification testés ne semblent pas adaptés au modèle d'audition de S. Seneff.

Très récemment, C. Jankovski, H. Vo et R. Lippmann ont comparé plusieurs pré-traitements et représentations lors d'expériences de reconnaissance de mots isolés par modèles de Markov pour diverses conditions d'enregistrement [Jankovski *et al.*, 1995].

Des coefficients MFCC servent de référence, et sont comparés avec des coefficients LPCC et des pseudo-coefficients cepstraux obtenus à partir du modèle d'audition de S. Seneff et d'un modèle auditif proposé par O. Ghitza [Ghitza, 1994]. Le critère de comparaison acoustique des modèles est similaire à une distance de Mahalanobis diagonale. En milieu très bruité, les coefficients basés sur des modèles d'audition sont plus performants que les MFCC, mais l'écart devient très faible pour des enregistrements en condition normale. De plus, les coefficients LPCC sont moins performants que les coefficients MFCC, même avec une analyse d'ordre élevée et une transformation bilinéaire permettant de se rapprocher de l'échelle Mel [Lee *et al.*, 1990], et l'écart devient considérable en milieu bruité.

1.2.4.2. Coefficients différentiels

En raison de la coarticulation, les zones transitoires sont essentielles pour la reconnaissance de la parole. La prise en compte des variations spectrales dans le système ESOPE a ainsi permis une amélioration du taux de reconnaissance phonétique évaluée à 15 % [Mariani, 1982]. S. Furui a réalisé des expériences perceptives sur l'identification de syllabes consonne-voyelle (CV) dont le début ou la fin sont tronquées par pas de 10 ms [Furui, 1986b]. Ces expériences montrent que les parties du signal de parole concentrant les plus fortes transitions spectrales correspondent à des "points perceptifs critiques" nécessaires à l'identification des syllabes, tandis que les zones stationnaires des silences et des voyelles semblent moins cruciales.

Les systèmes de reconnaissance classiques ne modélisent pas l'évolution du signal. Il est possible d'introduire dans ces systèmes une information sur la dynamique temporelle du signal en utilisant, en plus des paramètres initiaux, des coefficients différentiels du premier ordre issus des coefficients cepstraux ou de l'énergie [Furui, 1986a]. Soit $C_k(t)$ le coefficient cepstral d'indice k de la trame t, alors le coefficient différentiel $\Delta C_k(t)$ correspondant est calculé sur $2n_\Delta+1$ trames d'analyse par:

$$\Delta C_k(t) = \frac{\sum_{i=-n_\Delta}^{n_\Delta} i \cdot C_k(t+i)}{\sum_{i=-n_\Delta}^{n_\Delta} i^2} \tag{1.27}$$

Ces coefficients, appelés aussi coefficients delta, estiment la pente de la régression linéaire d'un coefficient à un instant donné. La dérivée de l'énergie ΔE_0 est calculée de la même façon. Leur utilisation améliore sensiblement les performances des systèmes markoviens de reconnaissance et s'est généralisée [Soong & Rosenberg, 1988; Lee *et al.*, 1990; Schwartz *et al.*, 1989; Lee *et al.*, 1992; Junqua *et al.*, 1993]. Pour des expériences de reconnaissance de mots isolés par DTW, S. Furui trouve un optimum pour 7 à 9 trames de lissage, soit une durée de 56 à 72 ms. J.-C. Junqua obtient un résultat très similaire avec une durée de 50 à 70 ms [Junqua *et al.*, 1993]. En reconnaissance du locuteur, F. Soong et A. Rosenberg retiennent une durée élevée de l'ordre de 100 ms [Soong & Rosenberg, 1988]. Un lissage sur 5 trames soit 50 ms est utilisé dans les systèmes de reconnaissance d'AT&T ou de BBN [Lee *et al.*, 1992; Schwartz *et al.*, 1989]. Le système de CMU réalise une estimation des coefficients delta par simple différence de deux vecteurs; la durée de lissage correspond là encore à 50 ms [Lee *et al.*, 1990]; l'usage de ces coefficients delta améliore de 6 % le taux de reconnaissance phonétique du système SPHINX sur la base TIMIT.

Des coefficients différentiels du second ordre peuvent aussi contribuer à l'amélioration des systèmes d'identification phonétique [Hanson & Applebaum, 1990; Wilpon *et al.*, 1991; Lee *et al.*, 1992; Lamel & Gauvain, 1993]. B. Hanson et T Applebaum ont montré leur intérêt pour de la parole bruitée et soumise à l'effet Lombard. C.-H. Lee constate une amélioration faible mais significative du taux de reconnaissance par mots sur les bases DARPA. Ces coefficients $\Delta\Delta C_k$ et $\Delta\Delta E_0$ sont calculés par régression linéaire des coefficients delta sur $n_{\Delta\Delta}$ trames (typiquement $n_{\Delta\Delta}$=1 ou 2).

1.2.4.3. Réduction du nombre de coefficients

L'utilisation des coefficients différentiels dans les systèmes de reconnaissance entraîne un doublement ou même un triplement de la taille des vecteurs de paramètres gérés par les systèmes de reconnaissance. Cette augmentation est coûteuse en ressources de calcul et de mémoire et introduit une redondance de l'information représentée. Il est possible de ne choisir que certains coefficients parmi tous ceux disponibles, en tronquant les coefficients différentiels d'ordre élevé [Ljolje, 1994b] ou en classant les coefficients sur un critère particulier [Bocchieri & Wilpon, 1993]. Des analyses sont proposées pour réduire la dimension de l'espace des paramètres, comme une Analyse en Composantes Principales (ACP) pour décorréler les coefficients [Wang *et al.*, 1993; Chigier & Leung, 1992; Jankovski *et al.*, 1995] ou une Analyse Linéaire Discriminante (ALD) [Tridgell *et al.*, 1992].

A. Tridgell évalue le F-ratio de chaque coefficient comme le rapport de la variance intra-classe sur la variance inter-classes phonétiques; un nombre réduit de coefficients optimal au sens du F-ratio est ensuite obtenu au moyen d'une analyse discriminante, et les gains en reconnaissance sont sensibles [Tridgell *et al.*, 1992]. X. Wang, L. ten Bosch et L. Pols utilisent l'ACP pour décorréler les coefficients issus d'un banc de filtres, ce qui permet de représenter ensuite la dispersion des coefficients avec des matrices de covariance diagonales; par contre, la réduction de la dimension des vecteurs de paramètres entraîne toujours une dégradation des résultats [Wang *et al.*, 1993]. Il y a peu à attendre *a priori* d'une ACP sur des coefficients cepstraux, puisque la transformation cepstrale produit des coefficients faiblement corrélés; B. Chigier et H. Leung montrent que l'application d'une ACP à des coefficients spectraux se compare à l'utilisation d'un nombre identique de coefficients cepstraux [Chigier & Leung, 1992]. C. Jankovski, H. Vo et R. Lippmann aboutissent aux mêmes résultats, avec un léger avantage à l'ACP sur le cepstre en milieu non bruité [Jankovski *et al.*, 1995].

1.2.4.4. Conclusions

Nous tirons des études précédentes les enseignements suivants:

- Les analyses par banc de filtres sont comparables aux analyses par prédiction linéaire en milieu normal mais semblent plus robustes au bruit.
- Les échelles fréquentielles Mel ou Bark issues d'études sur la perception sont préférables à une échelle linéaire; l'intégration de modèles d'audition plus complets dans les systèmes de reconnaissance reste expérimental.
- L'analyse cepstrale permet d'obtenir un nombre réduit de coefficients peu corrélés à partir des sorties d'un banc de filtres; l'analyse en composantes principales est une alternative possible à cette réduction qui ne lui est pas sensiblement supérieure.
- Il est souhaitable d'associer au cepstre une distance euclidienne pondérée, comme la distance de Mahalanobis, plutôt qu'une simple distance euclidienne.
- Les coefficients différentiels permettent de prendre explicitement en compte la dynamique du signal dans les systèmes de reconnaissance à base de HMM.

Les coefficients MFCC restent ainsi un choix raisonnable pour représenter le signal de parole dans une tâche de reconnaissance. Le tableau 1.1 présente les pré-traitements et les coefficients de quelques systèmes de reconnaissance décrits dans des publications.

Référence	Laboratoire	BP	α	F	P	Analyse	Coefficients	TΔ	Catégorie
[Davis 80]		5		26	6	20 Mel	10 MFCC		DTW+d_E
[Furui 86]	NTT	4		32	8	AR	10 LPCC+10Δ+ΔE	72	DTW+d_E
[Soong 88]	ATT	3	0,95	45	15	9 AR	8 LPCC+8Δ	105	DTW+d_M
[Robinson 90]	CUED - REPN	8		32	16	20 Bark	20 BASC		NN
[El-Bèze 90]	IBM - Tangora	10		26	10	20 Mel	20 MFSC		HMM d
[Chow 87]	BBN - Byblos	10		20	10	TFD	14 MFCC		HMM d
[Lee 89]	CMU - Sphinx	8	0,97	20	10	14 AR; BL	12 LPCC+E+13Δ	40	HMM d
[Huang 93a]	CMU - Sphinx II	8	0,97	20	10	14 AR	12 LPCC+E+13Δ+13ΔΔ	40/80	HMM sc
[Lee 92]	ATT	3,8	0,95	30	10	10 AR; L	12 LPCC+12Δ+ΔE+13ΔΔ	50/70	HMM c
[Wilpon 93]	ATT	3	oui	45	15	8 AR; L	12 LPCC+12Δ+ΔE+13ΔΔ	75	HMM c
[Zhao 93]	Panasonic	5		20	10	8 PLP	8 PLP+E+9Δ	50	HMM c
[Young 92]	CUED - HTK	8	oui	16	10	Mel	12 MFCC+E+13Δ		HMM c
[Woodland 95]	CUED - HTK95	8				Mel	12 MFCC+E+13Δ+13ΔΔ		HMM c
[Gauvain 94]	LIMSI	8		30	10	15 Bark	15 BACC+E+16Δ+16ΔΔ		HMM c
[Jouvet 94]	CNET - Phil90	4	oui	32	16	24 Mel	8 MFCC+E+9Δ+9ΔΔ		HMM c
[Ljolje 94b]	ATT	4		30	10	AR	9 LPCC+9Δ+6ΔΔ+ΔE+ΔΔE		HMM c

Légende

BP: Largeur de bande d'analyse (kHz)

α: Facteur de pré-accentuation $(1 - \alpha z^{-1})$

F: Taille de la fenêtre d'analyse (ms)

P: Pas d'analyse (ms)

Analyse: 9 AR = analyse auto-régressive d'ordre 9
20 Mel/Bark = banc de 20 filtres en échelle Mel ou Bark
BL = transformation bilinéaire
L = filtrage cepstral par fenêtre sinusoïdale

Coefficients: LP-/MF-/BA-(Linear Prediction/Mel Frequency/Bark Auditory)
-CC/-SC (Cepstrum/Spectrum Coefficients)
PLP (Perceptual Linear Prediction)
E (logarithme de l'énergie)
Δ et ΔΔ (coef. différentiels du 1er et 2ème ordre)

TΔ: Durée de la régression pour le calcul des coefficients différentiels du 1er ordre (ms)

Catégorie: DTW + d_E/d_M (alignement dynamique + distance euclidienne ou de Mahalanobis)
NN (réseaux de neurones)
HMM d/sc/c (modèles de Markov cachés discrets/semi-continus/continus)

Tableau 1.1. Caractéristiques des modules de pré-traitement pour quelques systèmes de reconnaissance (taille et pas de la fenêtre d'analyse, type d'analyse, coefficients calculés).

1.3. Modèles de Markov cachés

Depuis leur introduction en traitement de la parole [Baker, 1975; Jelinek, 1976], les modèles de Markov cachés (*Hidden Markov Models* ou HMM) ont pris une importance considérable, au point que la quasi-totalité des systèmes de RAP utilisent cette modélisation. Les modèles de Markov cachés supposent que le phénomène modélisé est un processus aléatoire et inobservable qui se manifeste par des émissions elles-mêmes aléatoires. Ces deux niveaux donnent à l'approche markovienne une flexibilité qui est séduisante pour modéliser un phénomène aussi complexe que la production de la parole. De nombreuses présentations théoriques des HMM existent dans la littérature; nous reprenons en partie les notations de L. Rabiner [Rabiner, 1989].

1.3.1. Définitions

Un modèle de Markov λ est un automate probabiliste d'états finis constitué de N états. Un processus aléatoire se déplace d'état en état à chaque instant, et on note q_t le numéro de l'état atteint par le processus à l'instant t. L'état réel q_t du processus n'est pas directement observable — on dit qu'il est “caché” — mais le processus émet après chaque changement d'état un symbole discret o_t qui appartient à un alphabet fini de n_V symboles $V = \{v_k\}_{1 \le k \le n_V}$. Dans le cas d'un processus markovien du premier ordre, la probabilité de passer de l'état i à l'état j à l'instant t et d'émettre le symbole v_k ne dépend ni du temps, ni des états aux instants précédents. Un modèle de Markov caché ou HMM est alors défini par:

- $S = \{s_i\}_{1 \le i \le N}$ l'ensemble des N états, en sachant que le processus part de l'état initial s_1 à l'instant t=0 et arrive à l'état final s_N à l'instant t=T
- $A = \{a_{ij}\}_{1 \le i,j \le N}$ l'ensemble des probabilités de transition entre les états i et j:

$$a_{ij} = P(q_t = j / q_{t-1} = i) \tag{1.28}$$

- $V = \{v_k\}_{1 \le k \le n_V}$ l'ensemble des n_V symboles observables,
- $B = \{b_j(k)\}_{1 \le j \le N,\ 1 \le k \le n_V}$ l'ensemble des probabilités d'émission du symbole v_k lors de l'arrivée dans l'état j, avec:

$$b_j(k) = P(o_t = v_k / q_t = j) \tag{1.29}$$

Des variantes existent cependant. La probabilité d'émission est parfois notée $b_{ij}(k)$ dans le cas où l'on associe l'émission du symbole à la transition plutôt qu'à l'état d'arrivée:

$$b_{ij}(k) = P(o_t = v_k / q_{t-1} = i, q_t = j) \tag{1.30}$$

Les deux approches sont équivalentes, et le choix est guidé par des critères d'implémentation. D'autre part, le principe de l'émission de symboles discrets peut être généralisé à un espace continu.

La figure 1.2 donne l'exemple d'un modèle autorisant le bouclage sur le deuxième état et des oscillations entre le deuxième et le troisième état; les transitions non représentées par un arc sont supposées de probabilité nulle. Enfin, l'émission d'une observation o par le troisième état suit la probabilité $b_3(o)$.

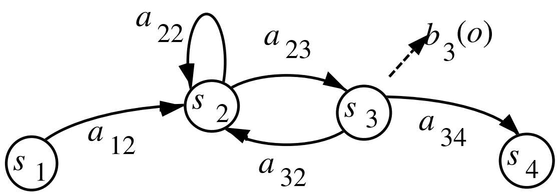

Figure 1.2. Exemple de modèle de Markov.

La réalisation par la machine λ d'un processus markovien de durée T est décrite par:

- $Q=(q_1...q_T)$ un chemin *a priori* caché parmi les N états; on pose de plus par convention $q_0=1$ puisque tous les processus partent de l'état initial à l'instant $t=0$, et on impose l'arrivée à l'état final par $q_T=N$;
- $O=(o_1...o_T)$ une suite d'observations appartenant à l'alphabet de n_V symboles.

A partir d'une suite d'observations O supposées émises par un modèle, différents problèmes peuvent être posés:

- L'évaluation de la probabilité que la suite des observations ait été émise par un modèle. Lorsque plusieurs modèles existent, cette évaluation permet le choix du modèle le plus probable.
- La recherche de la séquence d'états d'un modèle ayant produit les observations. La séquence cachée de plus forte probabilité est déterminée par l'algorithme de Viterbi.
- L'apprentissage des paramètres d'un modèle. A partir d'un modèle donné *a priori* et d'observations supposées émises par ce modèle, on cherche les probabilités de transition et d'émission maximisant la vraisemblance des observations.

Nous allons présenter la résolution de ces problèmes dans le cadre simplifié de la reconnaissance de mots isolés. Les modèles de mots émettent des observations du signal de parole. L'évaluation de la probabilité d'émission d'un mot par différents modèles permet alors l'identification d'un mot inconnu. Le chemin optimal parmi les états cachés sert pour l'initialisation des modèles avant leur apprentissage, mais la procédure est surtout utile pour le décodage de la parole continue.

1.3.2. Modélisation du signal de parole

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. 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, et 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.

1.3.2.1. Modèles de mots

Les modèles de Markov utilisés pour représenter la parole sont des modèles "gauche-droit", 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:

$$i > j \Rightarrow a_{ij} = 0 \tag{1.31}$$

Des modèles gauche-droit particuliers autorisant le bouclage à l'état courant, le passage à l'état suivant ou le saut d'un état ont été introduits par R. Bakis [Bakis, 1976] pour représenter des mots (cf. Figure 1.3). Ces transitions sont similaires aux contraintes sur la pente du chemin imposées lors de l'alignement par DTW. 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.

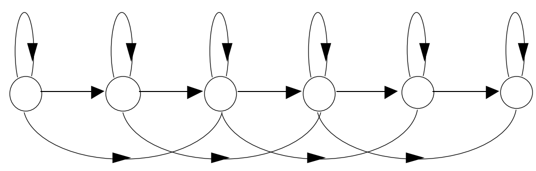

Figure 1.3. Exemple de modèle de Bakis.

1.3.2.2. Observations discrètes

L'espace de représentation du signal de parole est généralement un espace multi-dimensionnel continu $\mathcal{E}$, et les vecteurs de coefficients calculés sur une trame de signal sont des points de cet espace. Il est possible de modéliser ces observations par des modèles de Markov à émissions discrètes au moyen de la quantification vectorielle (QV). La QV permet le passage de l'espace $\mathcal{E}$ vers un espace discret, en partitionnant cet espace et en choisissant un représentant pour chaque classe. L'ensemble des représentants constitue un dictionnaire de n_V prototypes noté $V = \{v_k\}_{1 \le k \le n_V}$. Chaque vecteur $X \in \mathcal{E}$ est quantifié par le prototype du dictionnaire dont il est le plus proche au sens d'une distance $d(X,Y)$ définie dans l'espace $\mathcal{E}$:

$$X \xrightarrow{QV} \tilde{X} = v_{\tilde{r}} \quad \text{avec} \quad \tilde{r} = \underset{1 \le r \le n_V}{\arg\min}\, d(X, v_r) \tag{1.32}$$

Il suffit ensuite de manipuler l'indice r du prototype à la place du vecteur X. Divers algorithmes ont été proposés pour la réalisation du dictionnaire des prototypes, basés le plus souvent sur une classification hiérarchique descendante ou sur les nuées dynamiques [Gray, 1984]. Le dictionnaire est l'alphabet des symboles émis par un modèle de Markov. Cependant, la QV introduit des distorsions, et il est préférable lorsque cela est possible de travailler directement dans l'espace continu $\mathcal{E}$.

1.3.2.3. 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. Cette solution évite les distorsions introduites par la QV, mais pose le problème du choix des densités de probabilité et de la robustesse de leur estimation. L'utilisation d'une combinaison linéaire de gaussiennes dans l'espace $\mathbb{R}^d$ est fréquente:

$$b_j(o) = \sum_{k=1}^{G} g_k \mathcal{N}(o, \mu_k, \Sigma_k) \tag{1.33}$$

où μ_k et Σ_k sont respectivement la moyenne et la matrice de covariance de la gaussienne, et g_k la pondération qui lui est affectée. Nous rappelons que la densité de probabilité d'une loi normale de moyenne μ et de matrice de covariance Σ en dimension d est:

$$\mathcal{N}(o, \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp^{-\frac{1}{2}(o-\mu)^t \Sigma^{-1} (o-\mu)} \tag{1.34}$$

On peut remarquer les similitudes de cette densité avec une distance de Mahalanobis. L'hypothèse qui est souvent faite 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. D'autres types de densités de probabilité sont possibles, qu'il s'agisse d'une représentation paramétrique différente comme le Laplacien, de la densité de probabilité de l'erreur de prédiction par un modèle auto-régressif [Juang & Rabiner, 1985] ou encore de la sortie d'un réseau de neurones dans le cas de systèmes de reconnaissance hybrides [Bourlard & Wellekens, 1990].

1.3.2.4. 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 $o_t = (o_t^1 \ldots o_t^f)$ composée de *f* flux est le produit de la probabilité d'émission pour chacun des flux:

$$b_j(o_t) = \prod_{k=1}^{f} b_j^k(o_t^k) \tag{1.35}$$

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 [Lee, 1988]. 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.

1.3.3. 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.

1.3.3.1. Décision bayésienne

Le formalisme des HMM peut être exploité pour la reconnaissance de mots isolés. Supposons que les observations $O=(o_1 \ldots o_T)$ sont les trames acoustiques d'un mot inconnu. On cherche à trouver le mot $\tilde{m}$ qui a été prononcé parmi l'ensemble E_m des mots du vocabulaire, soit d'après la règle de décision bayésienne:

$$\tilde{m} = \arg\max_{m \in E_m} P(m/O) = \arg\max_{m \in E_m} P(O/m)P(m) \tag{1.36}$$

Chaque mot m est modélisé par une machine λ_m, d'où l'hypothèse suivante:

$$P(O/m) = P(O/\lambda_m) \tag{1.37}$$

L'équation (1.36) peut alors être reformulée:

$$\tilde{m} = \arg\max_{m \in E_m} P(O/\lambda_m)\ P(m) \tag{1.38}$$

Sa résolution nécessite l'estimation de la probabilité d'émission des observations $P(O/\lambda_m)$ pour chacune des machines modélisant un mot.

1.3.3.2. Probabilité d'émission des observations

La probabilité que la suite d'observations $O=(o_1...o_T)$ soit émise par la machine λ 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 E_Q de tous les chemins possibles de longueur T:

$$P(O/\lambda) = \sum_{Q \in E_Q} P(O,Q/\lambda) \tag{1.39}$$

D'après les définitions du modèle, la probabilité de partir à l'instant t=0 de l'état d'indice q_0=1 et de suivre le chemin $Q=(q_1...q_T)$ est le produit des probabilités de transition sur le chemin:

$$P(Q/\lambda) = \prod_{t=1}^{T} a_{q_{t-1}q_t} \tag{1.40}$$

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

$$P(O/Q,\lambda) = \prod_{t=1}^{T} b_{q_t}(o_t) \tag{1.41}$$

d'où la probabilité conjointe du chemin et des observations:

$$P(O,Q/\lambda) = P(Q/\lambda).P(O/Q,\lambda) = \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \tag{1.42}$$

La probabilité d'émission des observations est finalement:

$$P(O/\lambda) = \sum_{Q \in E_Q} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \tag{1.43}$$

Avec une machine à N états, le nombre de chemins possibles est de l'ordre de N^T, et l'ensemble des chemins E_Q devient très rapidement impossible à décrire. Il existe heureusement un algorithme rapide, dit "Forward-Pass" (estimation directe), qui permet de calculer récursivement cette quantité en maîtrisant l'explosion combinatoire [Baum, 1972].

1.3.3.3. Estimation directe

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

$$\alpha_t(i) = P(o_1...o_t, q_t = i/\lambda) \tag{1.44}$$

Alors une récurrence sur le temps en parallèle pour tous les états permet d'obtenir $P(O/\lambda)$, en suivant les étapes suivantes:

- Départ du processus dans l'état initial:

$$\alpha_0(i) = \begin{cases} 1, & i = 1 \\ 0, & 1 < i \le N \end{cases} \tag{1.45}$$

- Récurrence sur le temps et sur les états pour calculer les valeurs de $\alpha_t(j)$:

    Pour t allant de 1 à T,

    Pour j allant de 1 à N,

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(o_t) \tag{1.46}$$

- Arrivée du processus dans l'état final:

$$P(O/\lambda) = \alpha_T(N) \tag{1.47}$$

Cette méthode exploite une structure en treillis combinant les états et les instants afin de partager des valeurs intermédiaires entre plusieurs chemins.

1.3.3.4. Estimation rétrograde

L'estimation directe est suffisante pour obtenir la probabilité d'émission recherchée. Cependant, une estimation rétrograde dans le temps est aussi possible, avec la probabilité $\beta_t(i)$ que les observations après l'instant t soient émises en partant de l'état d'indice i:

$$\beta_t(i) = P(o_{t+1} \ldots o_T / q_t = i, \lambda) \tag{1.48}$$

et la récurrence est très similaire à la précédente:

- Initialisation:

$$\begin{cases} \beta_T(i) = 0, & 1 \le i < N \\ \beta_T(N) = 1 \end{cases} \tag{1.49}$$

- Récurrence pour t allant de T à 1:

$$\beta_{t-1}(i) = \sum_{j=1}^{N} a_{ij} b_j(o_t) \beta_t(j), \quad 1 \le i \le N \tag{1.50}$$

- Terminaison:

$$P(O/\lambda) = \beta_0(1) \tag{1.51}$$

Les variables directes et rétrogrades sont utilisées conjointement lors de l'apprentissage des modèles qui est détaillé au §1.3.5.

1.3.4. Recherche des états cachés

La procédure d'estimation directe ou rétrograde fournit la probabilité d'émission des observations cumulée 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 directe ou rétrograde. 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.

1.3.4.1. Chemin optimal

Au vu des observations O émises par le modèle λ, la séquence d'états Q la plus probable ayant pu émettre ces observations est donnée par:

$$\tilde{Q} = \underset{Q \in E_Q}{\arg\max}\ P(Q/O,\lambda) \tag{1.52}$$

La règle de Bayes nous permet d'écrire, après simplification du terme constant:

$$\tilde{Q} = \underset{Q \in E_Q}{\arg\max}\ P(O,Q/\lambda) \tag{1.53}$$

et donc, en reprenant l'équation (1.42):

$$P(O,\tilde{Q}/\lambda) = \max_{Q \in E_Q} \prod_{t=1}^{T} a_{q_{t-1}q_t} b_{q_t}(o_t) \tag{1.54}$$

La recherche du maximum remplace ici la somme de l'équation (1.43). L'algorithme de Viterbi [Viterbi, 1967; Forney, 1973], qui est comme l'algorithme de DTW une application de la programmation dynamique, résout ce problème de manière récursive.

1.3.4.2. Algorithme de Viterbi

La variable $\Phi(t,i)$ est définie comme la probabilité maximale que les observations observées jusqu'à l'instant t aient été émises par le modèle λ en suivant un chemin qui arrive à l'état d'indice i:

$$\Phi(t,i) = \max_{q_1 \ldots q_{t-1}} P(o_1 \ldots o_t, q_1 \ldots q_{t-1} q_t = i/\lambda) \tag{1.55}$$

Alors une récurrence similaire à celle suivie pour le calcul de la probabilité d'émission s'applique, à laquelle s'ajoute la mémorisation du meilleur chemin:

- Le processus est initialement dans l'état d'indice 1:

$$\Phi(0,i) = \begin{cases} 1, & i = 1 \\ 0, & 1 < i \le N \end{cases} \tag{1.56}$$

- Récurrence sur les valeurs de $\Phi(t,i)$:

  Pour t allant de 1 à T,

  Pour j allant de 1 à N,

$$\Phi(t,j) = \max_{1 \le i \le N} \Phi(t-1,i)\, a_{ij}\, b_j(o_t) \tag{1.57}$$

et l'on conserve la mémoire du meilleur état précédent:

$$m(t,j) = \arg\max_{1 \le i \le N} \Phi(t-1,i)\, a_{ij}\, b_j(o_t) \tag{1.58}$$

- Arrivée du processus dans l'état final:

$$P(O,\tilde{Q}/\lambda) = \Phi(T,N) \tag{1.59}$$

et construction du meilleur chemin $\tilde{Q} = (\tilde{q}_1 \ldots \tilde{q}_T)$ avec:

$$\tilde{q}_T = m(T,N) \tag{1.60}$$

puis pour t allant de T à 2:

$$\tilde{q}_{t-1} = m(t,\tilde{q}_t) \tag{1.61}$$

La probabilité d'émission sur le meilleur chemin peut être utilisée pour la reconnaissance comme une approximation de la probabilité d'émission par le modèle; mais cette méthode de résolution est sous-optimale puisqu'elle néglige les chemins de plus faible probabilité:

$$P(O,\tilde{Q}/\lambda) < P(O/\lambda) \tag{1.62}$$

La segmentation du signal fournie par l'algorithme de Viterbi sert principalement à l'initialisation des modèles à l'apprentissage et à la reconnaissance de la parole continue.

1.3.5. 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, est supposé fixée *a 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 œuvre est généralement plus difficile.

1.3.5.1. Maximum de vraisemblance

L'estimation par maximum de vraisemblance (*Maximum Likelihood Estimation* ou MLE) consiste à choisir les paramètres du modèle λ afin de rendre maximale la probabilité d'émission des observations O par le modèle:

$$\tilde{\lambda} = \arg\max_{\lambda} P(O/\lambda) \tag{1.63}$$

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 $\tilde{\lambda}_n$, le nouveau modèle $\tilde{\lambda}_{n+1}$ vérifie:

$$P(O/\tilde{\lambda}_{n+1}) \geq P(O/\tilde{\lambda}_n) \tag{1.64}$$

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.

1.3.5.2. Ré-estimation du modèle

La ré-estimation des paramètres du modèle λ est basée sur le comptage du nombre moyen de transitions observées entre les états *i* et *j*. La probabilité $\omega_t(i,j)$ de suivre cette transition à l'instant *t* peut s'exprimer au moyen des variables directes et rétrogrades introduites au paragraphe 1.3.3:

$$\omega_t(i,j) = P(q_{t-1} = i, q_t = j/O,\lambda) = \frac{\alpha_{t-1}(i)a_{ij}b_j(o_t)\beta_t(j)}{P(O/\lambda)} \tag{1.65}$$

Le nombre moyen de transitions entre *i* et *j* est donc:

$$\gamma_{ij} = \sum_{t=1}^{T} \omega_t(i,j) \tag{1.66}$$

et la probabilité de transition est ré-estimée par:

$$\tilde{a}_{ij} = \frac{\gamma_{ij}}{\sum_{k=1}^{N} \gamma_{ik}} \tag{1.67}$$

Il est important de remarquer que la structure du modèle est conservée au cours de la ré-estimation; les transitions initialement interdites entre deux états le restent:

$$a_{ij} = 0 \Rightarrow \tilde{a}_{ij} = 0 \tag{1.68}$$

L'estimation de la probabilité d'émission associée à un état nécessite le décompte des observations correspondant à chaque catégorie de symbole:

$$\tilde{b}_j(k) = \frac{1}{\gamma_j} \sum_{\substack{t=1 \\ o_t = v_k}}^{T} \omega_t(j), \quad 1 \leq k \leq n_V \tag{1.69}$$

avec:

$$\omega_t(j) = P(q_t = j/O,\lambda) = \frac{\alpha_t(j)\beta_t(j)}{P(O/\lambda)} \quad \text{et} \quad \gamma_j = \sum_{t=1}^{T} \omega_t(j) \tag{1.70}$$

La ré-estimation des probabilités d'émission est différente pour des modèles continus. Nous détaillons le cas de densités de probabilité continues représentées par une gaussienne multi-dimensionnelle, mais ces formules peuvent être généralisées au cas de multi-gaussiennes. Le vecteur de moyenne et la matrice de covariance de la densité de probabilité associée à l'état *i* sont recalculés comme:

$$\tilde{\mu}_i = \frac{1}{\gamma_i} \sum_{t=1}^{T} \omega_t(i) o_t \tag{1.71}$$

et:

$$\tilde{\Sigma}_i = \frac{1}{\gamma_i} \sum_{t=1}^{T} \left\{ \omega_t(i)(o_t - \mu_i)(o_t - \mu_i)^t \right\} \tag{1.72}$$

1.3.5.3. Autres critères d'apprentissage

Le critère du maximum de vraisemblance n'est pas nécessairement le meilleur critère d'apprentissage des modèles, car sa convergence nécessite une bonne initialisation des paramètres et un nombre élevé de données d'apprentissage. D'autres critères ont été proposés, comme les critères MAP et MMI.

Le critère bayésien ou MAP (*Maximum A Posteriori*), maximise la probabilité *a posteriori*:

$$\tilde{\lambda} = \arg\max_{\lambda} P(\lambda/O) = \arg\max_{\lambda} P(O/\lambda)P(\lambda) \tag{1.73}$$

Le modèle λ est considéré comme une variable aléatoire, ce qui permet une estimation de la probabilité P(λ) [Lee *et al.*, 1991; Gauvain & Lee, 1994]. Les problèmes qui se posent sont alors le choix d'une densité de probabilité pour les paramètres du modèle λ, le calcul de cette densité par apprentissage, et l'estimation *a posteriori* du modèle. Par exemple, les pondérations des multi-gaussiennes sont modélisées par une densité de Dirichlet et les moyennes et les matrices de covariance de chaque gaussienne par une densité normale de Wishart. Des formules de ré-estimation itératives analogues aux formules de Baum-Welch sont ensuite appliquées; les moyennes des gaussiennes sont ainsi obtenues par une combinaison linéaire des valeurs *a priori* des paramètres et des valeurs estimées sur les données. Lorsque les modèles sont *a priori* équiprobables, le critère MAP devient identique au critère MLE. L'apprentissage bayésien a été expérimenté en adaptation au locuteur, en considérant que les paramètres d'un système indépendant du locuteur sont les valeurs a priori de ces paramètres; cet apprentissage permet une convergence plus rapide, en durée d'adaptation, qu'un apprentissage par MLE sur les seules phrases d'adaptation [Lee & Gauvain, 1993]. Cependant, les modèles sont appris de manière indépendante avec les critères MAP et MLE, ce qui ne permet pas un apprentissage discriminant.

Le critère MMI (*Maximum Mutual Information*) tient compte de tous les modèles lors de la ré-estimation, ce qui autorise à décomposer la probabilité *a priori* des observations $P(O)$ comme une somme des probabilités conjointes sur tous les modèles:

$$\tilde{\lambda} = \arg\max_{\lambda} P(\lambda/O) = \arg\max_{\lambda} \frac{P(O/\lambda)P(\lambda)}{\sum_{m\in E_m} P(O/\lambda_m)P(\lambda_m)} \tag{1.74}$$

Tous les modèles sont modifiés simultanément de manière à maximiser la probabilité d'émission par le bon modèle tout en la diminuant pour les autres. Ce critère discriminant peut être appliqué par des techniques de descente du gradient ou par des ré-estimations itératives proches des équations de Baum-Welch [Bahl *et al.*, 1986; Kapadia *et al.*, 1993; Normandin *et al.*, 1994]. Les expériences décrites par Y. Normandin, R. Cardin et R. De Mori montrent que des ré-estimations par MMI améliorent significativement des modèles initialement appris par MLE; cependant ces techniques sont plus coûteuses en temps de calcul à l'apprentissage, et leur convergence n'est pas toujours maîtrisée.

1.3.6. Implémentation

Seuls les principes généraux des algorithmes ont été présentés. Leur mise en œuvre nécessite quelques précisions. Ainsi, les probabilités calculées sont souvent trop faibles pour être manipulées directement, et il est préférable de transposer les calculs dans le domaine logarithmique. De plus, afin de permettre une estimation plus robuste des paramètres, les modèles sont généralement appris à partir de plusieurs séquences d'observation plutôt qu'une seule, et les probabilités d'émission de plusieurs états peuvent partager les mêmes données d'apprentissage par un mécanisme de "liens".

1.3.6.1. Changement d'échelle

La valeur des variables directes et rétrogrades calculées pour l'évaluation et l'apprentissage décroît rapidement pour des durées d'observations élevées, au point de sortir de la dynamique numérique habituelle des calculateurs. Une solution consiste à les normaliser à l'instant t par la somme des $\alpha_t(i)$ [Rabiner, 1989] ce qui leur redonne une dynamique acceptable. Il est possible aussi de transposer tous les calculs dans le domaine logarithmique. La multiplication est transformée en addition, et l'addition de deux valeurs a et b à partir de leur logarithme est réalisée par:

$$\log(a+b) = \log a + \log(1+\exp^{\log b - \log a}), \quad \text{si } \log a \geq \log b \tag{1.75}$$

ce qui limite considérablement les erreurs numériques. Cette dernière solution est celle retenue à CMU et à l'Université de Cambridge [Lee, 1988; Young, 1992a].

1.3.6.2. Initialisation des modèles

La convergence des modèles vers un maximum le plus proche possible du maximum global au cours de leur ré-estimation par les équations de Baum-Welch nécessite une bonne initialisation des probabilités d'émission. Cette initialisation peut être obtenue par la procédure itérative des "k-moyennes segmentales", qui utilise l'algorithme de Viterbi pour aligner les trames d'observation au modèle et recalculer en conséquence les probabilités [Rabiner, 1989]. A la première itération, les probabilités d'émission sont initialisées par une segmentation *a priori* du signal, par exemple au moyen d'un alignement uniforme entre les observations et le modèle. Lors des itérations suivantes, le décodage des observations d'apprentissage est réalisé avec l'algorithme de Viterbi pour trouver la meilleure suite d'états dans le modèle et associer chaque trame émise à un état particulier. Les formules de ré-estimation déjà présentées s'appliquent en remplaçant la probabilité de l'équation (1.65) par l'alternative:

$$\omega_t(i,j) = \begin{cases} 1 & \text{si } q_{t-1} = i \text{ et } q_t = j \\ 0 & \text{sinon} \end{cases} \tag{1.76}$$

Selon ce principe, la moyenne d'une densité de probabilité gaussienne associée à l'état i est estimée à partir des trames émises par cet état comme:

$$\tilde{\mu}_i = \frac{1}{\operatorname{card} T_i} \sum_{t \in T_i} o_t \quad \text{où} \quad T_i = \{t / q_t = i\} \tag{1.77}$$

Lorsque la densité de probabilité d'un état est représentée par une multi-gaussienne, une procédure de *k*-moyennes appliquée à ces mêmes trames fournit les nouvelles moyennes des *k* gaussiennes de la loi.

1.3.6.3. Séquences multiples

Nous avons supposé jusqu'à présent que les paramètres du modèle étaient ré-estimés au vu d'une seule suite d'observations O. Pour une estimation robuste, il est nécessaire qu'un nombre important d'observations soit disponible, et donc que la durée de l'émission soit très élevée. Dans le cas de modèles de mot ou de phonème, la faible durée d'une seule séquence d'observations peut être compensée par l'utilisation de plusieurs séquences. Les formules de ré-estimation sont légèrement modifiées, sans que cela change les propriétés de convergence de l'algorithme.

Considérons un ensemble de R suites d'observation $\{O_r\}_{1 \le r \le R}$, chacune de durée T_r. Les variables avant et arrière doivent être calculées indépendamment pour chaque séquence d'observations, ce qui permet d'exprimer le compteur de l'équation (1.65) pour chaque séquence comme:

$$\omega_t^r(i,j) = \frac{\alpha_{t-1}^r(i) a_{ij} b_{ij}(o_t^r) \beta_t^r(j)}{P(O_r / \lambda)} \tag{1.78}$$

Il faut utiliser dans le décompte toutes les observations de toutes les séquences. Ainsi, le nombre moyen de transitions entre *i* et *j* pour toutes les séquences devient:

$$\gamma_{ij} = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \omega_t^r(i,j) \tag{1.79}$$

La suite de la procédure s'en déduit logiquement. Rien n'impose que toutes les séquences utilisées proviennent du même locuteur, ce qui donne la possibilité de réaliser avec la même méthode l'apprentissage de modèles dépendants ou non du locuteur.

1.3.6.4. Données partagées

Il n'est pas toujours possible de disposer d'un nombre suffisant de données d'apprentissage. Le partage des données d'apprentissage entre plusieurs probabilités d'émission est une façon d'améliorer la robustesse de leur estimation. Nous nous plaçons à titre d'exemple dans le cas d'une densité de probabilité gaussienne associée aux états. Supposons que l'on note $I=\{i\}_{1\leq i\leq N}$ l'ensemble des états "liés" partageant la même distribution statistique. Cette propriété est conservée par la procédure de ré-estimation si l'on cumule les compteurs pour tous les états de la classe I:

$$\omega_t(I) = \sum_{i\in I} \omega_t(i) \quad \text{et} \quad \gamma_I = \sum_{t=1}^{T} \omega_t(I) \tag{1.80}$$

d'où une probabilité d'émission identique pour tous les états de la classe; on dit que les états sont "liés". Cette procédure est en particulier très utile pour l'apprentissage de modèles phonétiques dépendants du contexte qui est présenté au §1.4. Le principe des liens peut être généralisé à tous les paramètres d'un modèle [Young, 1992b]. Un mode particulier de partage des données, où toutes les densités de probabilité d'émission sont exprimées comme une combinaison linéaire d'un nombre fini de gaussiennes partagées par tous les états, conduit aux modèles de Markov semi-continus, qui réunissent les caractéristiques des modèles discrets et des modèles continus [Huang *et al.*, 1993b; Bellegarda & Nahamoo, 1990].

1.4. 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.

Les espaces de représentation ont déjà été présentés (cf. § 1.2). Les autres points doivent être considérés avec attention. En particulier, pour les systèmes de reconnaissance de parole continue à grand vocabulaire, il n'est pas réaliste d'apprendre un modèle pour chaque mot, et 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, car 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.

1.4.1. 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, car 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.

1.4.1.1. Unités sub-lexicales

La modélisation par mots est très efficace pour des applications de reconnaissance de mots isolés ou en vocabulaire limité [Rabiner *et al.*, 1989]. Pour la reconnaissance de parole continue à grand vocabulaire, il est impossible d'apprendre un modèle pour chacun des mots du vocabulaire. La modélisation d'unités acoustiques de taille plus courte que le mot devient nécessaire. Ces unités peuvent être des phonèmes indépendants du contexte [Jelinek, 1976], des modèles phonétiques dépendants du contexte [Schwartz *et al.*, 1985], des diphones, des syllabes pour le système PARSYFAL développé à IBM-France [El-Bèze, 1990], ou encore des unités acoustiques sans signification phonologique comme les fénones utilisés à IBM [Bahl *et al.*, 1993]. Les unités les plus utilisées sont les suivantes:

- 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; mais 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.

- Le multi-phonème regroupe plusieurs phonèmes. Un diphone, par exemple, désigne deux phonèmes successifs, et par extension la transition entre les parties stationnaires de deux phonèmes. La coarticulation est enfin modélisée, au détriment d'un nombre élevé de modèles, de l'ordre de plusieurs milliers.
- Un allophone est une réalisation acoustique particulière d'un phonème. Par exemple, un modèle phonétique dépendant simultanément du contexte phonétique gauche et droit est un modèle d'allophone particulier, parfois appelé triphone bien que son support temporel soit celui d'un phonème isolé [Schwartz *et al.*, 1985]. Cela crée un très grand nombre de combinaisons, mais en regroupant certains allophones ayant une réalisation acoustique similaire [Lee, 1990], il est possible d'établir un compromis entre le nombre de modèles et la finesse de la modélisation acoustique.
- Le fénone est un modèle acoustique très simple, puisqu'une machine fénonique est typiquement caractérisée par un seul état et une seule densité de probabilité d'émission [Bahl *et al.*, 1993]. La description du niveau phonétique devient superflue, et l'on passe directement des fénones aux mots.

Les modèles de mots alors sont construits par concaténation de modèles sub-lexicaux, en autorisant éventuellement plusieurs variantes correspondant à des prononciations différentes. Nous nous limitons par la suite à la description de modèles de phonèmes ou d'allophones.

1.4.1.2. 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, mais 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-droit à trois états émetteurs [Schwartz *et al.*, 1985; Lamel & Gauvain, 1993; Ljolje, 1994a; Woodland *et al.*, 1995]; mais rien n'oblige *a priori* tous les phonèmes à être modélisés de la même façon.

Les chercheurs d'IBM [Derouault, 1987] ont proposé un modèle phonétique de structure légèrement plus complexe, qui a été utilisé aussi à CMU [Lee, 1988] (cf. Figure 1.4). Le modèle comprend 7 états et 12 ou 13 transitions possibles entre ces états. Les probabilités d'émission sont associées aux transitions. Par une mise en commun des données entre certaines transitions lors de l'estimation des probabilités d'émission, le nombre de lois différentes est ramené à trois, notées b_1, b_2 et b_3, associées respectivement au début, au milieu et à la fin du phonème. Les branches hautes et basses du modèle correspondent schématiquement aux modes d'élocution lents et rapides.

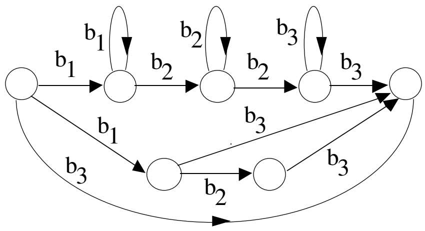

Figure 1.4. Modèle phonétique utilisé à IBM et à CMU.

Il peut paraître insatisfaisant d'apprendre de manière rigoureuse les paramètres des modèles, tout en choisissant empiriquement leur topologie. Les chercheurs du CNET ont utilisé des connaissance phonologiques *a priori* pour la conception de modèles d'allophones [Jouvet *et al.*, 1991a]. Des techniques d'inférence grammaticale ont été adaptées pour automatiser la conception des modèles [Casacuberta *et al.*, 1990; Lockwood & Blanchet, 1993]. Enfin, R. De Mori, M. Galler et F. Brugnara proposent une méthode de recherche de la topologie par recuit simulé, guidée par la diminution du taux d'erreur [De Mori *et al.*, 1995].

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 [Huang *et al.*, 1993b], et 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. Des procédures automatiques ont là encore été proposées, et seront discutées dans la partie expérimentale de cette étude.

1.4.1.3. Modèles phonétiques dépendants du contexte

Les modèles phonétiques indépendants du contexte ne prennent pas en compte les phénomènes de coarticulation, qui sont essentiels en production de la parole. Il est possible de modéliser des allophones dépendants du contexte phonétique gauche ou droit, c'est-à-dire du phonème précédant ou suivant. Pour n_λ phonèmes différents, n_λ^2 allophones en contexte gauche et autant en contexte droit doivent être représentés par des modèles. La modélisation simultanée des contextes gauche et droit nécessite l'apprentissage de n_λ^3 modèles; cela représente plusieurs dizaines de milliers de modèles, même si en pratique certaines combinaisons de sons ne sont jamais prononcées dans la langue étudiée. L'apprentissage robuste d'un nombre aussi élevé de modèles est pratiquement impossible.

Des techniques ont été proposées pour faciliter l'apprentissage d'un système de RAP utilisant des modèles phonétiques contextuels. En particulier, plusieurs allophones, supposés avoir une réalisation acoustique similaire, sont regroupés dans une classe et sont modélisés par une seule machine; on parle alors d'allophones généralisés. Le regroupement des allophones peut être réalisé par une classification automatique basée sur une mesure de similarité entre modèles [Lee *et al.*, 1990], en construisant un arbre de décision [Bahl *et al.*, 1991] ou en utilisant des connaissances de nature phonologique [Ljolje, 1994a]. De plus, il est possible de modéliser l'influence du contexte gauche et du contexte droit de manière indépendante, et d'interpoler les allophones gauche et droit à partir de modèles dépendants d'un seul contexte; ces modèles sont appelés quasi-triphones. A. Ljolje combine des modèles phonétiques indépendants du contexte, des allophones gauches ou droits généralisés, des quasi-triphones et de véritables allophones dépendants simultanément des contextes gauche et droit dans un même système de reconnaissance [Ljolje, 1994a].

1.4.2. Modèles de langage

La modélisation acoustique permet à elle seule de réaliser la transcription phonétique d'une phrase. En l'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 [Mariani, 1990]... Il est donc indispensable, autant que possible, d'introduire dans les systèmes de RAP des connaissances sur les niveaux supérieurs du langage.

1.4.2.1. 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:

$$P(m_1 \ldots m_n) = P(m_1) \prod_{i=2}^{n} P(m_i / m_1 \ldots m_{i-1}) \qquad (1.81)$$

En pratique, la manipulation des probabilités correspondant à toutes les suites de mots de longueur quelconque est impossible, et 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:

$$P(m_i / m_1 \ldots m_{i-1}) \approx P(m_i / m_{i-1} \ldots m_{i-k+1}) \tag{1.82}$$

Les modèles *k*-classes sont une généralisation des modèles *k*-grammes, pour laquelle la probabilité d'un mot dépend de la classe des mots précédents, les classes étant définies comme une partition de l'ensemble des mots. Les modèles bigrammes (*k*=2) ou trigrammes (*k*=3) sont les plus utilisés.

1.4.2.2. Bigrammes et trigrammes

Si l'on note $f(m_1 \ldots m_n)$ la fréquence de la suite de mots $m_1 \ldots m_n$ dans un corpus d'apprentissage, un modèle bigramme estime la probabilité d'un mot conditionné par son prédécesseur comme:

$$\tilde{P}(m_j / m_i) = \frac{f(m_i, m_j)}{f(m_i)}, \text{ si } f(m_i) > 0 \tag{1.83}$$

Les modèles trigrammes doivent être estimés sur de grandes bases de données, mais les conditions réelles n'autorisent pas une mesure fiable des fréquences de toutes les suites de trois mots. La combinaison linéaire de modèles de trigramme, bigramme et unigramme permet alors de lisser cette estimation:

$$\tilde{P}(m_3 / m_1, m_2) = \delta_1 \frac{f(m_1, m_2, m_3)}{f(m_1, m_2)} + \delta_2 \frac{f(m_2, m_3)}{f(m_2)} + \delta_3 f(m_3), \text{ si } f(m_1, m_2) > 0 \tag{1.84}$$

Les trigrammes de mots sont utilisés dans les meilleurs systèmes de reconnaissance [Jelinek, 1985; Ljolje, 1994a; Gauvain *et al.*, 1994; Woodland *et al.*, 1995]. L'intégration de modèles de langages plus complexes reste néanmoins nécessaire pour les systèmes de compréhension de la parole [Corazza *et al.*, 1993].

1.4.3. 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 les 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.

1.4.3.1. Modèles connectés

Connaissant la transcription phonétique d'une phrase d'apprentissage, une machine composite est fabriquée en mettant les modèles correspondants en séquence. Par exemple, si la phrase est transcrite par la suite phonétique $\Phi=(\varphi_1\ldots\varphi_T)$, alors un modèle de la phrase $\Lambda=(\lambda_1\ldots\lambda_T)$ est fabriqué, où λ_t modélise le phonème φ_t. Ce modèle composite est constitué en reliant le dernier état du modèle λ_t à l'état initial du modèle λ_{t+1}. Le modèle Λ peut être ré-estimé sur la phrase complète avec la procédure standard déjà décrite, et il n'est donc pas nécessaire de disposer d'une segmentation de la phrase. Tous les modèles sont de cette manière ré-estimés simultanément, ce qui permet l'exploitation de grandes bases de données, car 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 (cf. Figure 1.5).

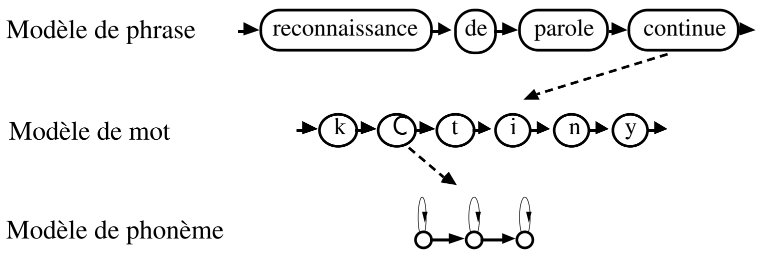

Figure 1.5. Modélisation d'une phrase à partir de modèles phonétiques.

1.4.3.2. Procédure d'apprentissage

La procédure de ré-estimation des modèles connectés ne garantit pas que la segmentation implicitement réalisée soit correcte, et risque de diverger si les modèles sont mal initialisés; elle ne doit donc être appliquée qu'avec des modèles déjà estimés sur des segments isolés. La procédure complète d'apprentissage se compose finalement des étapes suivantes:

- Segmentation manuelle de quelques phrases, afin de disposer de plusieurs représentants pour chaque modèle;
- Pour chacune des unités acoustiques:
  - Extraction des segments acoustiques correspondants à cette unité;
  - Initialisation du modèle par alignement uniforme;
  - Ré-estimation(s) du modèle par l'algorithme de Viterbi;
  - Ré-estimation(s) du modèle par l'algorithme de Baum-Welch;
- Transcription phonétique du reste de la base de données;

- Ré-estimation(s) en modèles connectés:
  - Pour chaque phrase:
    - 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.

1.4.4. Décodage de parole continue

En reconnaissance de la parole continue, la suite de mots recherchée est celle qui maximise l'équation 1.4 rappelée ici:

$$\tilde{M} = \arg\max_{M \in E_M} P(O/M).P(M) \tag{1.85}$$

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.

1.4.4.1. Réseau de modèles

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 *a priori* équiprobables. Dans le cas le plus simple, n'importe quelle suite de mots est autorisée, et le modèle composite est constitué des 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 (cf. Figure 1.6 pour l'exemple d'un réseau modélisant une suite de chiffres); 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 (cf. par exemple l'état initial et l'état final du réseau de la figure 1.6).

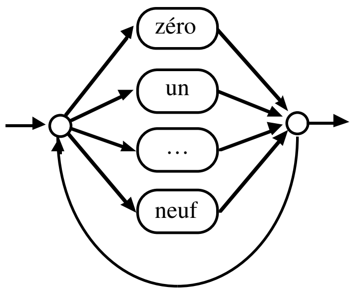

Figure 1.6. Exemple de réseau de modèles autorisant
l'émission d'une suite quelconque de chiffres.

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. Cette reconnaissance est donc sous-optimale par rapport à l'équation (1.85).

1.4.4.2. Algorithme du passage de jeton

Des variantes de l'algorithme de Viterbi ont été proposées pour le décodage de la parole continue. Les algorithmes "One Pass" [Vintsiuk, 1971; Bridle *et al.*, 1982; Ney, 1984] ou "Level Building" [Myers & Rabiner., 1981], initialement développés pour la reconnaissance de mots enchaînés par DTW, peuvent être utilisés. L'algorithme du "Token Passing" ou "passage de jeton" est plus général [Young, 1992a]. Il gère sans difficultés un réseau lexical de modèles de mots, constitués eux-mêmes de modèles sub-lexicaux. Le jeton est un objet placé dans un état qui contient la probabilité du meilleur chemin arrivant dans cet état à l'instant courant; de plus, il conserve la mémoire de ses déplacements:

- Un jeton de valeur nulle est placé dans tous les états initiaux autorisés;
- A chaque instant:
    - Le jeton de chaque état est propagé vers tous les états connectés et il est augmenté du coût de la transition et de l'émission;
    - Dans chaque état, le meilleur jeton arrivant est conservé;
- A la fin, le meilleur jeton parmi tous les états est récupéré.

Le décodage de la phrase en mots ou en phonèmes se déduit de la suite des modèles ayant émis les observations sur le chemin optimal.

1.4.4.3. Décodage acoustico-phonétique

Nous modélisons la parole par des unités sub-lexicales, typiquement par des modèles phonétiques dépendants ou non du contexte. En l'absence d'informations sur le lexique et la syntaxe du langage utilisé dans l'application, il est possible de réaliser avec ces modèles un décodage acoustico-phonétique (DAP), qui fournit une transcription phonétique des phrases prononcées. Le réseau de modèles doit autoriser le décodage par un enchaînement quelconque de phonèmes.

On peut s'interroger sur l'utilité du DAP, alors que les recherches sur les systèmes analytiques indiquent que celui-ci ne doit pas être une étape préalable à l'utilisation des connaissances linguistiques. Cependant, la qualité de la reconnaissance en mots dépend étroitement de la qualité de la modélisation acoustique, et une amélioration de la reconnaissance phonétique se traduit par une augmentation des performances du système de reconnaissance [Gauvain *et al.*, 1994]. L'évaluation des modèles acoustiques par DAP permet alors une mise au point rapide du système de reconnaissance, et une meilleure détection des faiblesses du système dans les étapes initiales de la reconnaissance. Cependant le module acoustique ainsi obtenu ne prend son sens qu'avec l'intégration d'un modèle de langage adapté à l'application développée.

1.4.4.4. Intégration d'un 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 DAP, si λ_i et λ_j sont respectivement les modèles des phonèmes φ_i et φ_j, la probabilité de passer à l'intérieur du modèle composite de λ_i à λ_j est celle estimée par un modèle de bigramme phonétique:

$$\tilde{P}(\lambda_j / \lambda_i) = \frac{f(\varphi_i, \varphi_j)}{f(\varphi_i)} \tag{1.86}$$

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. L'algorithme "Stack Decoding" est une alternative au "Token Passing" qui simplifie l'intégration d'un modèle de langage [Jelinek, 1976]. Une autre stratégie est possible: les *N* phrases ou suites de symboles les plus probables (*N-best*) sont obtenues avec les seuls modèles acoustiques; ces phrases sont ensuite reclassées en tenant compte de la probabilité *a priori* de chacune des phrases [Schwartz & Kubala, 1992]. Cette dernière stratégie a par exemple été utilisée pour intégrer un modèle de trigramme lexical [Gauvain *et al.*, 1994; Woodland *et al.*, 1995; Lefèvre, 1995].

1.4.5. Evaluation des résultats

L'évaluation d'un système de reconnaissance est réalisée en comparant le décodage obtenu sur un certain nombre de phrases de test avec l'étiquetage de référence de ces phrases. La précision de cette estimation dépend du nombre de tests réalisés.

1.4.5.1. Calcul des taux de décodage

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 & 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. Les taux d'identification T_{ide}, d'omission T_{omi}, de substitutions T_{sub} et d'insertion T_{ins} sont calculés en rapportant respectivement le nombre d'unités identifiées N_{ide}, omises N_{omi}, substituées N_{sub} et insérées N_{ins} au nombre d'unités de référence N_{ref}, en particulier:

$$T_{ide} = N_{ide}/N_{ref} = 100\% - (T_{sub} + T_{omi}) \tag{1.87}$$

Cependant, le taux d'identification ne tient pas compte des insertions; un taux de reconnaissance T_{rec} est alors arbitrairement défini comme la différence du taux d'identification et du taux d'insertion:

$$T_{rec} = T_{ide} - T_{ins} = 100\% - (T_{sub} + T_{omi} + T_{ins}) \tag{1.88}$$

Le taux de reconnaissance fournit une indication globale mais peu détaillée sur la qualité du système de RAP.

1.4.5.2. Intervalle de confiance

Le taux de reconnaissance n'est qu'une estimation des performances d'un système de RAP. Sa fiabilité dépend du nombre de test réalisés, et donc du nombre d'unités acoustiques à reconnaître dans le cas de la parole continue. Afin de juger si une variation du taux de reconnaissance est significative ou non au cours de nos expériences, il est indispensable de connaître la précision de cette mesure. C. Montacié et G. Chollet proposent de modéliser les réussites par une distribution binomiale [Montacié & Chollet, 1987]. Soit N le nombre de tests réalisés, et P le taux de reconnaissance mesuré, alors l'intervalle de confiance à x % du taux de reconnaissance $[P_-, P_+]$ est donné par:

$$P_\pm = \frac{P + \frac{{z_x}^2}{2N} \pm z_x \sqrt{\frac{P(1-P)}{N} + \frac{{z_x}^2}{4N^2}}}{1 + \frac{{z_x}^2}{N}}, \quad \text{avec} \begin{cases} z_{95\%} = 1{,}96 \\ z_{99\%} = 2{,}48 \end{cases} \tag{1.89}$$

Il y a donc $x\%$ de chances que le taux exact $\tilde{P}$ soit dans cet intervalle. Dans le cas où le nombre d'échantillons N est grand, l'expression précédente se simplifie, et l'erreur de mesure du taux de reconnaissance est par exemple bornée dans 95% des cas par:

$$\left|P - \tilde{P}\right| \leq_{95\%} 1{,}96\sqrt{\frac{P(1-P)}{N}} \tag{1.90}$$

Il est évident que la modélisation des réussites ou des échecs d'identification par une loi binomiale est discutable, mais elle fournit un ordre de grandeur souvent suffisant. Il existe cependant des tests plus spécialisés qui permettent de juger si les différences entre deux algorithmes appliqués à une même base de test sont significatives [Gillick & Cox, 1989]; le test de McNemar peut ainsi être utilisé en reconnaissance de mots isolés, et le test des paires associées pour la reconnaissance de parole continue.

1.5. Limites des systèmes de reconnaissance

Les systèmes actuels de reconnaissance automatique de la parole continue sont construits à partir d'une modélisation probabiliste avec des modèles de Markov cachés [Lee *et al.*, 1990; Gauvain *et al.*, 1994; Woodland *et al.*, 1995], ou avec une architecture hybride neuronale et markovienne [Robinson *et al.*, 1994]. L'approche probabiliste a montré son efficacité, et de nombreux travaux continuent à améliorer les performances des systèmes existants [Huang *et al.*, 1993a; Jouvet *et al.*, 1994; Ljolje, 1994a; Zhao, 1994; De Mori *et al.*, 1995]. Cependant, ces systèmes doivent être perfectionnés sur de nombreux points. Ainsi, localisation et identification phonétique sont très étroitement associées durant la phase de décodage par des modèles de Markov classiques, et la segmentation devient un résultat annexe du décodage; pourtant, la modélisation de la durée phonétique reste insuffisante. De plus, la grande variabilité inter-locuteurs diminue l'efficacité de modèles phonétiques indépendants du locuteur, qui deviennent faiblement discriminants. Enfin, la représentation du signal choisie est généralement appliquée de manière identique à des zones stationnaires comme le centre des voyelles et à des zones fortement transitoires comme les explosions. Les analyses du signal utilisées font généralement l'hypothèse d'une stationnarité à court terme du signal de parole qui n'est pas pertinente. Nous pensons qu'il est nécessaire de ré-introduire dans ces systèmes des traitements spécifiques, adaptés aux multiples aspects de la parole.

Nous avons choisi d'étudier plus particulièrement deux aspects essentiels pour la qualité d'un système de reconnaissance de la parole continue à base de modèles de Markov cachés: l'adaptation au locuteur des modèles acoustiques et le contrôle temporel de l'enchaînement des modèles. Nous proposons des solutions originales à ces problèmes. Afin d'évaluer la pertinence de nos traitements, nous avons développé un système de RAP de référence à base de modèles phonétiques contextuels. Ce système se compare honorablement aux systèmes de reconnaissance développés par les équipes de recherche les plus performantes.

2. Construction d'un système de référence

Dans le cadre de la reconnaissance de la parole continue indépendamment du locuteur, notre objectif est de développer un système de référence fondé sur les modèles de Markov cachés; il doit nous permettre d'évaluer des techniques nouvelles visant à l'amélioration des systèmes usuels de RAP. Le système de référence est mis au point à partir de la plate-forme HTK de l'Université de Cambridge et sur la base de données TIMIT. Sa réalisation est décomposée en deux étapes. Un système de base est tout d'abord construit avec des modèles phonétiques indépendants du contexte; un certain nombre de choix sont faits sur ce système, comme le nombre d'états des modèles, le type de densités de probabilité d'émission associées aux états et la représentation du signal par des coefficients cepstraux et des coefficients différentiels. Un système de référence plus performant est ensuite développé à partir du système de base; il est construit avec des modèles phonétiques dépendants du contexte et avec une modélisation plus fine des densités de probabilité. En particulier, une procédure automatique basée sur le test du χ^2 est proposée; elle permet de décider pour chaque état du nombre de gaussiennes adapté à cet état. Les résultats en reconnaissance de notre système de référence le situent à un niveau de performance comparable à celui atteint par des systèmes markoviens de RAP nécessitant l'estimation d'un nombre de paramètres beaucoup plus élevé. Les traitements qui nous semblent pertinents dans le domaine de l'adaptation au locuteur et du contrôle temporel seront intégrés au système de référence et évalués au moyen d'expériences de décodage acoustico-phonétique.

2.1. Principe d'un 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 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.

A la suite des efforts d'organismes internationaux (NIST et DARPA aux USA, GRECO-CNRS en France), un grand nombre de bases de données de parole sont disponibles sur CD-ROM (TIMIT, TI-DIGITS, *Resource Management*, *Wall Street Journal*, SWITCHBOARD, BDSONS, BREF) et sont diffusées par des structures comme le LDC (*Linguistic Data Consortium*). Des protocoles pour évaluer les systèmes de reconnaissance sur ces bases ont été mis en place (projet ESPRIT-SAM, par exemple). Cependant, les meilleurs systèmes sont d'une grande complexité et sont difficilement reproductibles, car ils intègrent des expertises acquises au fil des années par les meilleures équipes, comme celles de CMU, d'AT&T, de BBN, ou d'IBM. Il n'est donc pas facile d'atteindre des performances comparables à ces systèmes. Il serait souhaitable d'obtenir un consensus parmi les chercheurs sur les caractéristiques d'un "bon" système, qui puisse servir de référence aux équipes cherchant à construire un système de RAP.

G. Chollet et C. Gagnoulet définissent un système de référence comme étant un système facile à implémenter, robuste aux conditions d'utilisation, qui ne soit pas contraint à un fonctionnement en temps réel, et représentatif de l'état de l'art en reconnaissance de la parole [Chollet & Gagnoulet, 1982]. Ils ont proposé en 1982 le code en langage FORTRAN d'un système d'identification de mots isolés par DTW, à partir d'une représentation du signal sous forme de vecteurs MFCC. Leur proposition a suscité un grand intérêt, et J. Baker a présenté la même année au NBS (*National Bureau of Standards*, USA) une extension de leur système de référence à la reconnaissance de mots enchaînés [Baker, 1982]. Actuellement, les meilleurs systèmes de RAP sont basés sur des modèles de Markov. Les techniques utilisées pour implémenter un système représentatif de l'état de l'art sont plus complexes à mettre en œuvre qu'en 1982. La plate-forme logicielle HTK, développée par l'Université de Cambridge [Young, 1994], fournit tous les outils nécessaires à la réalisation d'un tel système; G. Chollet considère qu'HTK est un bon candidat pour la construction d'un système de référence [Chollet, 1995].

2.2. Plate-forme HTK

La plate-forme HTK (*Hidden Markov Model Toolkit*, ou "boîte à outils de modèles de Markov cachés") a été développée à l'Université de Cambridge par S.J. Young et son équipe. Elle est constituée d'un ensemble d'outils logiciels qui permettent de construire des systèmes de reconnaissance de la parole continue à base de modèles de Markov cachés [Young, 1992a]. HTK est remarquable par la très grande liberté de choix laissée tout au long de la construction du système de reconnaissance. Les modèles peuvent représenter des mots ou tout type d'unité sub-lexicale, et leur topologie est librement configurable. Les densités de probabilité d'émission, qui sont associées aux états, sont décrites par des multi-gaussiennes. Les modèles sont initialisés avec l'algorithme de Viterbi, puis ré-estimés par l'algorithme optimal de Baum-Welch. Le décodage est réalisé par l'algorithme de Viterbi, sous la contrainte d'un réseau syntaxique défini par l'utilisateur, et le résultat est enfin évalué par alignement dynamique avec la chaîne phonétique ou lexicale de référence.

L'ensemble de ces outils est écrit en langage C, et la documentation détaille leur utilisation et les principes de leur implémentation, ce qui permet d'intégrer de manière efficace les modifications souhaitées dans le système de reconnaissance. De plus, HTK est un système largement répandu dans le monde de la recherche; en 1992, ses concepteurs revendiquaient déjà plus d'une centaine d'utilisateurs. Nous avons donc décidé de construire notre système de décodage acoustico-phonétique de référence avec HTK. Ce système nous servira ensuite à évaluer de manière quantitative l'amélioration apportée par nos traitements spécifiques.

2.2.1. Présentation d'HTK

HTK dans sa version 1.4 est structuré en 9 librairies utilisées par 12 outils de base (cf. Tableau 2.1). Les outils manipulent des fichiers de différents types: signaux, étiquettes, paramètres, description des modèles, définition de réseaux. Les formats des fichiers de signaux et d'étiquettes des bases de données les plus répandues sont reconnus. Les autres fichiers sont dans un format particulier à HTK, décrit dans le manuel de référence. En particulier, les modèles et les réseaux sont définis dans des fichiers texte, ce qui facilite leur création et leur modification par l'utilisateur. Les options d'utilisation des outils sont transmises en argument sur la ligne de commande. Il est donc facile d'automatiser les processus d'apprentissage et de décodage avec des scripts écrits dans le langage de commande du système d'exploitation (par exemple en C-shell si l'on travaille sous Unix).

Librairies		*Outils de base*	
HShell	Interface avec le système d'exploitation	HLEd	Édition des fichiers d'étiquettes
HMath	Procédures mathématiques	HHEd	Édition des modèles
HSigP	Procédures de traitement du signal	HCode	Calcul des paramètres du signal
HDBase	Stockage en mémoire des paramètres	HInit	Initialisation d'un modèle
HSpIO	Gestion des fichiers de données	HRest	Ré-estimation d'un modèle
HLabel	Gestion des fichiers d'étiquettes	HERest	Ré-estimation des modèles enchaînés
HModel	Gestion des modèles	HVite	Décodage en parole continue
HParse	Lecture du réseau syntaxique	HResults	Résultats du décodage
HGraf	Affichage graphique	HList	Affichage des fichiers de données
		HLStats	Calcul de statistiques sur les étiquettes
		HSLab	Affichage du signal et des étiquettes
		HLab2Net	Génération automatique d'un réseau

Tableau 2.1. Librairies et outils de base de HTK.

2.2.2. Utilisation d'HTK

Les principaux outils de base de HTK s'enchaînent naturellement pour réaliser les différentes étapes d'un système de reconnaissance: calcul des paramètres du signal, apprentissage des modèles, et expériences de reconnaissance (cf. Figure 2.1).

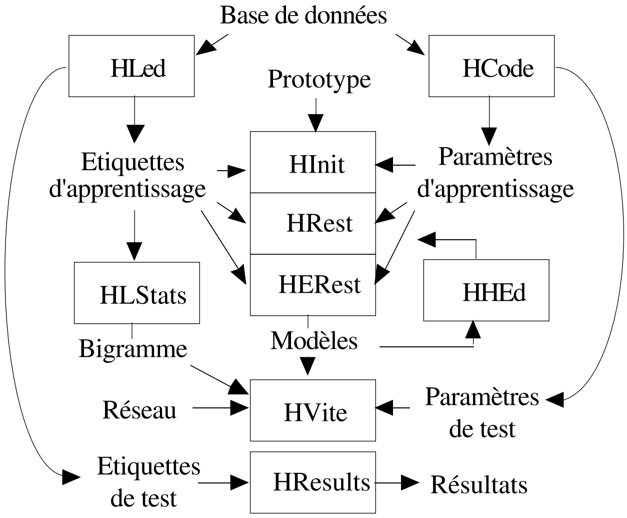

Figure 2.1. Structure d'un système de reconnaissance avec HTK
(d'après HTK Reference Manual p. 5).

2.2.2.1. Pré-traitement des données

Avant l'apprentissage des modèles, il est nécessaire de préparer les données d'apprentissage en calculant les paramètres du signal et en étiquetant les phrases d'apprentissage. La représentation du signal est obtenue avec l'outil HCode, qui produit en particulier des coefficients LPCC ou MFCC ainsi que l'énergie. Les coefficients différentiels du premier et du second ordre peuvent être calculés ultérieurement lors de la lecture des fichiers de paramètres, ce qui économise leur stockage en mémoire de masse. Les phrases d'apprentissage doivent être toutes étiquetées en fonction des unités acoustiques modélisées. Les bases de données de parole sont parfois fournies avec un étiquetage phonétique qui ne correspond pas exactement aux unités acoustiques modélisées. L'éditeur HLEd permet alors de modifier les étiquettes, par exemple pour regrouper plusieurs phonèmes différents dans une seule classe ou pour fusionner deux segments adjacents.

2.2.2.2. Topologie des modèles

Pour chaque unité acoustique, il faut définir un modèle prototype contenant la topologie choisie, à savoir le nombre d'états du modèle, les états entre lesquels les transitions sont possibles, le type de loi de probabilité associée à chaque état. L'état initial et l'état final ont la particularité de ne pas émettre d'observation, mais de servir uniquement à la connexion des modèles en parole continue. Il n'est pas nécessaire que tous les modèles utilisent le même prototype. Les probabilités d'émission sont associées aux états et sont décrites par une combinaison linéaire de gaussiennes multi-variables, caractérisées par leur moyenne et leur matrice de covariance dans l'espace des paramètres. La matrice de covariance est théoriquement symétrique, mais peut être choisie diagonale si l'on suppose l'indépendance entre les composantes des vecteurs de paramètres, et les vecteurs de paramètres peuvent être séparés en flux de données indépendants.

2.2.2.3. Apprentissage

Pour chacune des machines modélisant une unité acoustique, l'outil HInit initialise les probabilités d'émission des états du modèle au moyen de la procédure itérative des "k-moyennes segmentales" basée sur l'algorithme de Viterbi. Cette phase nécessite l'étiquetage fin des phrases d'apprentissage utilisées, car il faut extraire tous les segments correspondant à l'unité modélisée. Cette fastidieuse segmentation manuelle peut cependant être limitée à une fraction de l'ensemble d'apprentissage, de manière à disposer de quelques représentants pour l'initialisation de chaque modèle acoustique. L'estimation des paramètres d'un modèle est affinée avec HRest, qui applique l'algorithme optimal de Baum-Welch jusqu'à la convergence et ré-estime les probabilités d'émission et de transition. Dans une phase suivante, il est possible d'appliquer plusieurs itérations de l'outil HERest, qui ré-estime simultanément l'ensemble des modèles sur de la parole continue non segmentée.

Les modèles obtenus peuvent être améliorés, en augmentant par exemple le nombre de gaussiennes servant à estimer la probabilité d'émission d'une observation dans un état. Le choix du nombre optimal de gaussiennes est un problème délicat, généralement guidé par des heuristiques. Une commande de l'éditeur de modèles HHEd réalise l'augmentation du nombre de gaussiennes modélisant une densité de probabilité. Les modèles doivent être ensuite ré-estimés par HRest ou HERest. HTK offre aussi des facilités pour travailler avec des modèles contextuels, dépendants des contextes phonétiques gauches ou droits.

2.2.2.4. Reconnaissance

Le module de décodage de la parole continue, HVite, utilise l'algorithme de Viterbi pour trouver la séquence d'états la plus probable correspondant aux paramètres observés dans un modèle composite, et en déduire les unités acoustiques correspondantes. Le modèle composite autorise la succession des modèles acoustiques en fonction d'une syntaxe choisie par le concepteur du système. Il est possible d'écrire la syntaxe aux niveaux phonétique ou lexical, les mots du lexique pouvant être définis eux-mêmes par la concaténation d'unités sub-lexicales. Parallèlement à cette syntaxe, HVite tient compte d'un modèle de langage de type bigramme, estimé sur les étiquettes des phrases d'apprentissage par l'outil HLStats. Le résultat du décodage est comparé aux étiquettes de référence par un alignement dynamique réalisé par HResults, afin de compter les étiquettes identifiées, omises, substituées par une autre, et insérées, et de calculer le taux de reconnaissance.

2.3. Base de données TIMIT

Nous avons choisi de développer notre système de référence avec la base de données acoustiques TIMIT pour plusieurs raisons. Tout d'abord, cette base a été constituée pour illustrer au mieux la variabilité acoustique de l'anglais américain, et elle est fournie avec une segmentation phonétique de référence (étiquetage fin) qui simplifie l'apprentissage initial des modèles phonétiques. De plus, TIMIT peut être considérée comme une base de données de référence. Sa large diffusion dans la communauté internationale permet une évaluation objective des performances des systèmes développés.

2.3.1. Description de la base TIMIT

La base TIMIT est une base de données acoustiques et phonétiques dédiée à la reconnaissance de la parole indépendamment du locuteur [NIST, 1990]. Elle contient les enregistrements de 630 locuteurs américains, répartis en 8 "dialectes régionaux" ("dr1" à "dr8") et prononçant chacun 10 phrases. Ces phrases proviennent de 3 corpus:

- 2 phrases de calibration, prononcées par tous les locuteurs, servent à illustrer les variations régionales (identifiées "sa1" et "sa2");
- 5 phrases sont tirées au sort parmi 450 phrases phonétiquement équilibrées (identifiées "sx3" à "sx452").
- 3 phrases sont choisies pour maximiser les contextes acoustiques; chaque phrase n'est prononcée qu'une seule fois, soit un total de 1890 phrases différentes pour les 630 locuteurs (identifiées "si453" à "si2342");

Le vocabulaire total de la base est de 6100 mots. Le texte est lu, et les conditions d'enregistrement sont bonnes. Les 630 locuteurs de la base (438 hommes et 192 femmes) sont répartis entre l'ensemble d'apprentissage (462 locuteurs dont 326 hommes et 136 femmes) et l'ensemble de test (168 locuteurs dont 112 hommes et 56 femmes). Chaque locuteur est identifié par une lettre indiquant son genre ("m" pour les hommes et "f" pour les femmes), ses 3 initiales et un chiffre. L'ensemble d'apprentissage et l'ensemble de test sont respectivement notés EA et ET. Un sous-ensemble de l'ensemble de test, appelé noyau de test, ne contient que 24 locuteurs de test: deux hommes et une femme pour chacun des 8 "dialectes". Si l'on exclut les phrases de calibration, ces 24 locuteurs prononcent 192 phrases toutes différentes, et absentes de l'ensemble d'apprentissage. Le noyau de test est noté ET_{192}.

Pour chaque phrase, nous disposons du texte en anglais, du signal échantillonné à 16 kHz sur 16 bits, de la segmentation phonétique en 61 classes, et de la segmentation en mots. La segmentation phonétique est très fine; en particulier, l'occlusion précédant l'explosion des occlusives et des affriquées est étiquetée individuellement. Le tableau 2.2

donne la liste des étiquettes de la base TIMIT, le phonème correspondant dans l'Alphabet Phonétique International (API) d'après les indications de S. Seneff et V. Zue [Seneff & Zue, 1988] et une mise en contexte de ce phonème dans un mot anglais.

TIMIT	API	Exemple	TIMIT	API	Exemple	TIMIT	API	Exemple
Occlusives:			*Nasales*:			*Voyelles*:		
pcl p	p	pea	m	m	mom	iy	i^{y}	beet
tcl t	t	tea	em	m¡	bottom	ih	î	bit
kcl k	k	key	n	n	noon	ix	I	debit
bcl b	b	bee	nx	\$	winner	eh	E	bet
dcl d	d	day	en	n¡	button	æ	æ	bat
gcl g	g	gay	ng	N	sing	aa	A	bott
dx	\	muddy	eng	N¡	washington	ao	O	bought
q	?	bat	*Liquides:*			uh	Ü	book
Affriquées:			l	l	lay	uw	u	boot
dcl jh	j&	joke	el	l¡	bottle	ux	ü	toot
tcl ch	c&	choke	r	r	ray	ax	´	about
Fricatives:			*Semi-voyelles*:			ax-h	´∞	suspect
f	f	fin	w	w	way	ah	U	but
th	Q	thin	y	y	yacht	er	‰	bird
s	s	sea	*Fricative glottale*:			axr	ÿ	butter
sh	s&	she	hh	h	hay	*Diphtongues*:		
v	v	van	hv	H	ahead	ey	e^{y}	bait
dh	D	then	*Silences*:			ay	A^{y}	bite
z	z	zone	h#			oy	O^{y}	boy
zh	z&	azure	pau			aw	A^{w}	bout
			epi			ow	o^{w}	boat

Tableau 2.2. Étiquetage de TIMIT, code API correspondant et exemple de mot anglais contenant le phonème.

2.3.2. Regroupement d'allophones

L'étiquetage d'origine en 61 classes est généralement jugé trop détaillé pour l'apprentissage de modèles phonétiques, et une réduction du nombre de classes phonétiques par regroupement d'allophones est réalisée. K.F. Lee et H.W. Hon ont proposé un regroupement en 39 classes phonétiques [Lee & Hon, 1989] qui a été ensuite réutilisé par d'autres chercheurs à des fins de comparaison [Robinson *et al.*, 1990; Niles, 1991; Chigier & Leung, 1992]. Ce regroupement s'effectue en deux étapes:

- avant l'apprentissage des modèles, regroupement en 48 classes par fusion d'allophones (m/em, n/nx, ng/eng, ax/ax-h, er/axr, ux/uw, hh/hv), regroupement des silences et des occlusions (nouvelle étiquette sil pour regrouper les silences h#/pau, cl pour les occlusions sourdes pcl/tcl/kcl, vcl pour les occlusions voisées bcl/dcl/gcl) et suppression de l'étiquette q (qui ne correspond pas toujours à une occlusive);

- des confusions sont autorisées entre certaines classes lors du calcul des taux de décodage (sh/zh, n/en, l/el, ih/ix, aa/ao, ax/ah, sil/epi/cl/vcl), conduisant finalement à des résultats sur 39 classes.

Nous avons choisi les mêmes regroupements pour notre système de reconnaissance, afin de comparer nos résultats à ceux déjà publiés, et en l'absence de critère objectif permettant d'en choisir un autre *a priori* plus pertinent. Nous présentons dans le tableau 2.3 des statistiques sur les 48 classes phonétiques d'apprentissage, en rappelant les regroupements réalisés, et en donnant pour chaque classe le nombre de représentants dans l'ensemble d'apprentissage ainsi que la durée moyenne des segments en milli-secondes.

Étiquette(s)	Nombre	Durée (ms)	Étiquette(s)	Nombre	Durée (ms)
Occlusives:			*Semi-voyelles*:		
b	2181	17	w	2216	60
d	2432	24	y	995	54
g	1191	27	*Fricative glottale*:		
p	2588	44	hh,hv	1660	67
t	3948	49	*Voyelles*:		
k	3794	52	iy	4626	95
dx	1864	29	{ ih	4248	78
Affriquées:			ix	7370	51
jh	1013	61	eh	3277	93
ch	820	86	ae	2292	136
Fricatives:			{ aa	2256	123
f	2215	103	ao	1865	123
th	745	92	uh	500	76
s	6176	113	uw,ux	1952	100
{ sh	1317	118	{ ax,ax-h	3892	47
zh	149	81	ah	2266	89
v	1994	60	er,axr	4138	95
dh	2376	36	*Diphtongues*:		
z	3682	84	ey	2271	127
Nasales:			ay	1934	155
m,em	3566	65	oy	304	168
{ n,nx	6896	52	aw	728	161
en	630	78	ow	1653	128
ng,eng	1220	61	*Silences*:		
Liquides:			{ sil=(h#,pau)	8283	191
{ l	4425	61	cl=(pcl,tcl,kcl)	12518	58
el	951	90	vcl=(bcl,dcl,gcl)	7219	54
r	4681	56	epi	908	42

Tableau 2.3. Statistiques sur le nombre de représentants et la durée moyenne des 48 classes phonétiques (les confusions autorisées au décodage sont désignées par une accolade).

2.4. Modèles indépendants du contexte

Pour le développement du système de référence, nous avons choisi de construire tout d'abord un système de base avec des modèles acoustiques indépendants du contexte. Ce système est simple, et tous les modèles ont la même topologie. Les choix faits pour ce système sont validés par des expériences de décodage acoustico-phonétique sur le noyau de test de la base TIMIT.

2.4.1. Caractéristiques du système de base

Les choix qui ont été faits pour le système de base sont les suivants:

- Représentation centi-seconde du signal par 12 MFCC, l'énergie, et 13 coefficients différentiels du premier ordre, soit 26 coefficients;
- Unités acoustiques: 48 phonèmes indépendants du contexte;
- Topologie des modèles: modèle de Bakis à 3 états émetteurs;
- Probabilité d'émission modélisée par une combinaison linéaire de gaussiennes à matrice de covariance diagonale;
- Modèle de langage: bigramme phonétique.

Tous les modèles ont la même topologie, et les probabilités d'émission de tous les états sont représentées par un nombre identique de G gaussiennes. Ce système de base est noté SB_G. Les modèles acoustiques et le modèle de langage sont appris sur les 8 phrases "si" et "sx" des 462 locuteurs d'apprentissage de la base TIMIT, soit 3696 phrases contenant 140225 segments acoustiques. L'apprentissage est réalisé en modèles isolés avec les outils HInit pour l'initialisation par les k-moyennes segmentales et HRest pour la ré-estimation par la procédure de Baum-Welch.

2.4.1.1. Nombre de paramètres du système

Une comparaison équitable entre les performances de différents systèmes de reconnaissance doit tenir compte du nombre de paramètres de chaque système, c'est-à-dire du nombre de valeurs numériques modifiées par l'apprentissage. Une augmentation du nombre de paramètres du système permet une modélisation acoustique plus fine, à condition que le modèle soit pertinent et qu'une quantité suffisante de données d'apprentissage soit disponible. Dans le cas du système de base SB_G, nous avons:

- n_λ machines (n_λ=48);
- N états par machine (N=3);
- G gaussiennes par état ($G \geq 1$);
- d moyennes et d variances par gaussienne (d=26).

Alors, pour le système SB_1 (G=1), les probabilités d'émission de tous nos modèles nécessitent l'apprentissage de $n_\lambda.N.G.2d$ = 48×3×1×52 = 7488 paramètres. Les probabilités de transition d'un modèle de Bakis à 3 états ne contiennent que 4 paramètres indépendants (trois pour le bouclage sur un état et un pour le saut de l'état central), soit 192 paramètres pour tous les modèles. Le nombre de paramètres du modèle acoustique provient donc essentiellement des probabilités d'émission. Enfin, le bigramme phonétique est décrit par $n_\lambda{}^2$ probabilités, soit 2304 paramètres pour le modèle de langage. Par la suite, la complexité du système de reconnaissance est définie comme le nombre de paramètres pour les probabilités d'émission.

2.4.1.2. Validation des choix du système

La robustesse du système de base relativement à certains des choix réalisés a été étudiée. En particulier, nous testons l'apport des coefficients différentiels du premier ordre par rapport aux coefficients statiques, l'indépendance des coefficients pour justifier l'usage de gaussiennes à matrice de covariance diagonale, l'effet du changement du nombre d'états des modèles phonétiques, ou l'utilité du bigramme phonétique. Chaque expérience étudie l'impact d'un de ces choix, en fixant les autres paramètres du système de base. Le nombre de gaussiennes par densité de probabilité a cependant été ajusté de manière à pouvoir comparer des systèmes de reconnaissance de complexité identique.

Les expériences de validation sont réalisées par décodage acoustico-phonétique sur le noyau de test ET_{192}. Les 192 phrases du noyau de test contiennent 7215 étiquettes phonétiques, et l'intervalle de confiance à 95% est de l'ordre de ±1,1 % pour des taux de reconnaissance compris entre 50% et 70%. L'utilisation d'un ensemble de test réduit permet une expérimentation plus rapide des différents choix, et elle évite une trop forte influence de l'ensemble de test sur la conception du système. Les taux de décodage du système de base sur l'ensemble de test complet sont présentés à la fin de ce chapitre.

2.4.2. Apport des coefficients différentiels

Le système de base utilise des coefficients MFCC et l'énergie, ainsi que les coefficients différentiels de ces paramètres. Le signal échantillonné à 16 kHz subit une pré-accentuation de 0,97. Les paramètres sont calculés toutes les centi-secondes sur une fenêtre d'analyse de 25 ms pondérée par une fenêtre de Hamming. Les 12 premiers coefficients cepstraux, obtenus à partir d'un banc de 24 filtres en échelle fréquentielle Mel, sont conservés. Le logarithme de l'énergie de la trame, normalisé sur toute la phrase, est ajouté aux 12 coefficients cepstraux pour former un vecteur de 13 coefficients. Des coefficients différentiels du premier et du second ordre, calculés automatiquement par les outils de HTK, sont éventuellement utilisés avec les coefficients dits "statiques". Ces choix *a priori* sont cohérents avec les pré-traitements des systèmes existants (cf. Tableau 1.1).

Nous avons testé l'apport de coefficients différentiels du premier puis du deuxième ordre aux 13 coefficients initiaux, en travaillant avec des vecteurs de dimension d=13 (12 MFCC; E), d=26 (12 MFCC; E; 12 ΔMFCC; ΔE) et d=39 (12 MFCC; E; 12 ΔMFCC; ΔE; 12 ΔΔMFCC; $\Delta\Delta E$). Les coefficients différentiels sont calculés sur 5 trames successives ($n_{\Delta}=n_{\Delta\Delta}=2$), ce qui correspond à un lissage sur 50 ms pour les coefficients du premier ordre et sur 90 ms pour les coefficients du deuxième ordre. La probabilité d'émission de chacun des états des modèles est représentée par une combinaison linéaire de G gaussiennes à matrice de covariance diagonale. Toutes les autres conditions expérimentales sont celles du système de base déjà décrites (i.e., 48 unités phonétiques modélisées par une machine de Bakis à 3 états, et un décodage sur le noyau de test en utilisant un bigramme phonétique). Des modèles avec une seule gaussienne par état sont initialisés par l'algorithme de Viterbi avec l'outil HInit et ré-estimés par les équations de Baum-Welch avec l'outil HRest. Le nombre de gaussiennes est ensuite augmenté de manière itérative: des modèles à G gaussiennes par état servent à l'initialisation de modèles à $2G$ gaussiennes par état, en suivant une procédure standard de HTK décrite plus loin (§2.5.4); les modèles sont enfin ré-estimés avec l'outil HRest.

Le Tableau 2.4 présente les taux de décodage pour les trois séries d'expériences et en fonction du nombre de gaussiennes par densité de probabilité. Les meilleurs résultats sont obtenus pour d=26 coefficients et G=8 gaussiennes par état. Au delà de 8 gaussiennes, les performances du système ne sont pas significativement supérieures, tandis que la convergence de l'apprentissage est plus longue à atteindre. L'apport des coefficients différentiels du premier ordre est majeur, de 6 à 10 % sur le taux de reconnaissance suivant les expériences. Pour un nombre identique de gaussiennes, la dérivée seconde n'apporte pas d'amélioration significative dans le cadre de ces expériences; son intérêt a été mis en évidence lors d'expériences de reconnaissance de mots en milieu bruité (cf. §1.2.4), et on peut signaler la bonne qualité d'enregistrement de la base TIMIT. Pour un même nombre de paramètres d'apprentissage du système de reconnaissance, les 26 coefficients statiques et delta sont les plus performants (cf. Figure 2.2). Ils ont donc été choisis pour le système de base.

Une expérience a été réalisée en supprimant certains coefficients différentiels d'indice élevé de l'ensemble de 39 coefficients pour n'en conserver que 26 (12 MFCC; E; 8 ΔMFCC, ΔE; 3 ΔΔMFCC; $\Delta\Delta E$); le taux de reconnaissance est légèrement supérieur, de 0,5%, à celui obtenu avec les 26 coefficients (12 MFCC; E; 12 ΔMFCC, ΔE) pour un nombre identique de paramètres du système. Cette expérience a été réalisée trop tardivement pour obtenir des résultats robustes et utiliser ces coefficients dans notre système de reconnaissance.

Nombre de gaussiennes *G*	Paramètres du système (milliers)	Phonèmes identifiés (%)	Phonèmes omis (%)	Phonèmes substitués (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
d=13 (12 MFCC + E)						
1	4	51,9	19,3	28,8	4,7	**47,2**
2	7,5	53,5	18,4	28,1	4,3	**49,2**
4	15	54,7	18,1	27,2	4,5	**50,2**
8	30	55,8	16,9	27,3	4,0	**51,8**
d=26 (12 MFCC + E + 12 ΔMFCC + ΔE)						
1	7,5	58,9	13,6	27,5	5,7	**53,2**
2	15	62,5	12,2	25,3	5,7	**56,8**
4	30	65,0	11,3	23,7	5,8	**59,2**
8	60	66,5	10,9	22,6	5,3	**61,2**
d=39 (12 MFCC + E + 12 ΔMFCC + ΔE + 12 ΔΔMFCC + ΔΔE)						
1	11	62,0	10,0	28,0	9,3	**52,7**
2	22,5	64,8	8,8	26,4	8,8	**56,0**
4	45	66,5	8,5	25,0	8,2	**58,3**
8	90	68,9	8,0	23,1	7,8	**61,1**

Tableau 2.4. Apport des coefficients différentiels aux taux de décodage du noyau de test ET_{192} par le système de base.

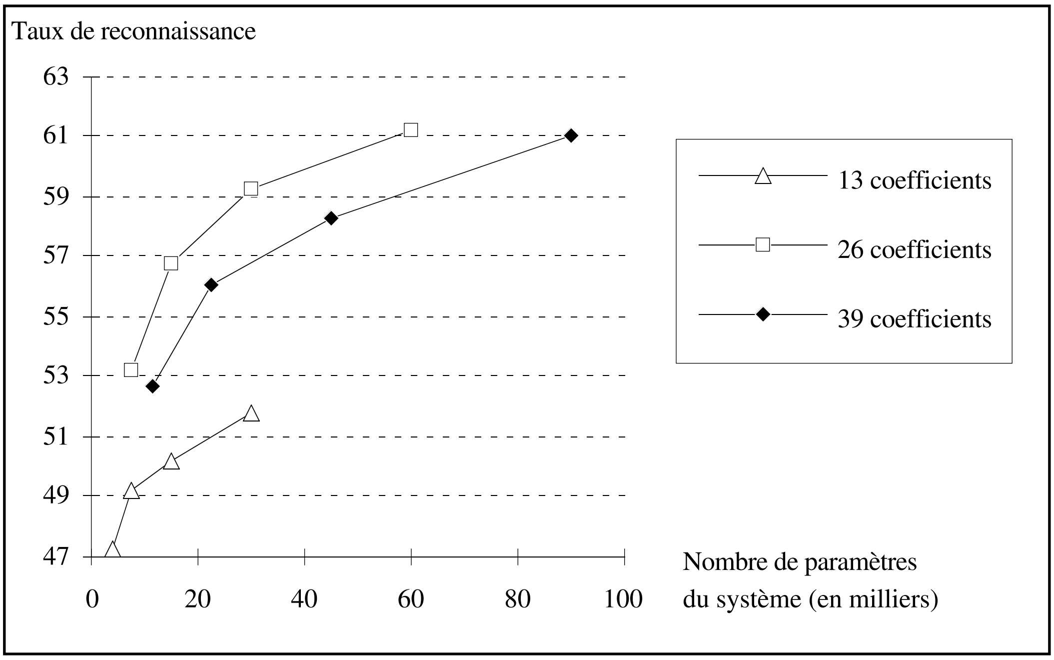

Figure 2.2. Apport des coefficients différentiels aux taux de décodage du noyau de test ET_{192} par le système de base en fonction du nombre de paramètres du système.

2.4.3. Topologie des modèles

Tous les phonèmes sont modélisés par une machine de Bakis à trois états (cf. Figure 2.3). Il s'agit d'un choix classique (cf. § 1.4.1), et on peut interpréter l'état initial et l'état final comme des états transitoires qui encadrent un état central plus stationnaire. L'état central peut être éventuellement sauté si la réalisation acoustique est très courte.

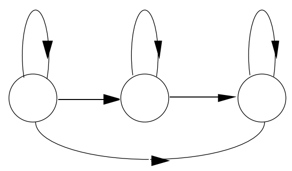

Figure 2.3. Topologie des modèles phonétiques utilisés.

Pour valider le choix du nombre d'états du modèle prototype, nous avons réalisé l'apprentissage de modèles à un, deux, trois et quatre états. Différents choix du nombre de gaussiennes *G* par densité de probabilité ont été testés afin d'obtenir des systèmes de complexité comparable (cf. Tableau 2.5). La comparaison avec un modèle à 3 états, pour un nombre identique de paramètres estimés (cf. Figure 2.4), donne l'avantage à ce dernier, le rayon de l'intervalle de confiance à 95% étant de 1,1%. Il s'agit cependant d'un choix global, concernant les 48 modèles simultanément. Il est possible, en fonction d'une connaissance *a priori* sur les phonèmes, ou *a posteriori* en observant l'adéquation de différentes topologies avec un phonème, de choisir des topologies plus adaptées.

Nombre de gaussiennes *G*	Paramètres du système (milliers)	Phonèmes identifiés (%)	Phonèmes omis (%)	Phonèmes substitués (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
N=1 (modèles à 1 état)						
4	10	65,2	6,9	27,9	19,1	**46,1**
N=2 (modèle gauche-droit à 2 états)						
1	5	58,6	11,6	29,8	10,7	**47,9**
2	10	62,8	10,3	26,9	9,4	**53,4**
6	**30**	66,5	8,9	24,6	9,0	**57,5**
N=3 (modèle de Bakis à 3 états)						
1	7,5	58,9	13,6	27,5	5,7	**53,2**
2	15	62,5	12,2	25,3	5,7	**56,8**
4	**30**	65,0	11,3	23,7	5,8	**59,2**
N=4 (modèle de Bakis à 4 états)						
1	10	57,7	15,7	26,7	4,3	**53,4**
2	20	59,9	16,3	23,8	3,3	**56,6**
3	**30**	60,5	16,8	22,7	3,1	**57,4**

Tableau 2.5. Influence du nombre d'états des modèles phonétiques sur les taux de décodage du noyau de test ET_{192} par le système de base.

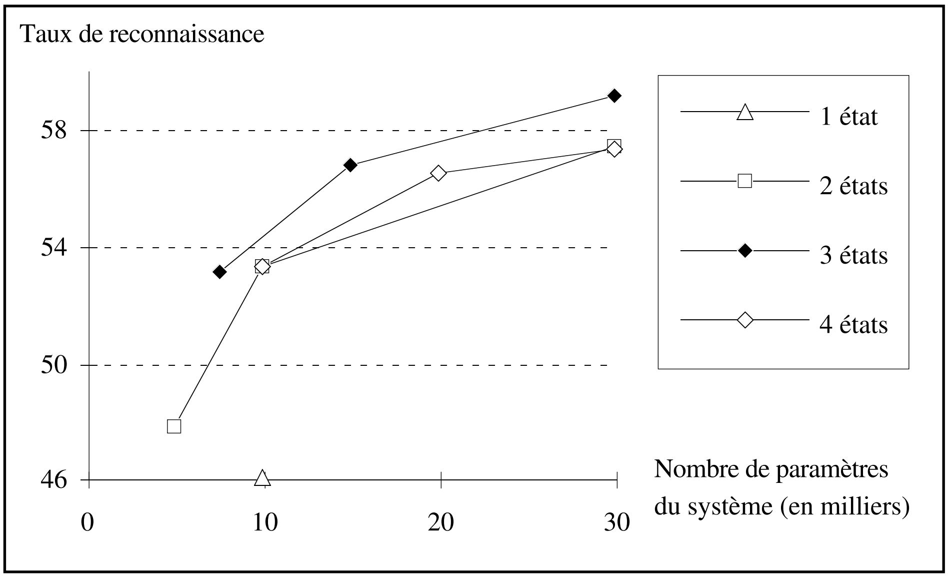

Figure 2.4. Influence du nombre d'états des modèles phonétiques sur les taux de décodage du noyau de test ET_{192} en fonction du nombre de paramètres du système.

Une expérience complémentaire, consistant à choisir pour chacun des phonèmes la topologie maximisant la probabilité d'émission sur l'ensemble d'apprentissage, a été réalisée. Les phonèmes courts (occlusives, semi-voyelles, liquides) sont plutôt mieux représentés par un modèle à deux états, et les phonèmes longs (fricatives ou affriquées sourdes et diphtongues) par un modèle à quatre états. Cette expérience n'a pas permis d'obtenir des résultats de décodage significativement meilleurs, et le modèle à trois états a été conservé pour tous les phonèmes.

2.4.4. Bigramme phonétique

Un bigramme phonétique ou lexical peut être utilisé lors du décodage par HVite. Les expériences de DAP réalisées pour la mise au point du système de reconnaissance utilisent un bigramme phonétique. Ce bigramme est estimé par l'outil HLStats sur la totalité de l'ensemble d'apprentissage. L'apport du bigramme au décodage est évalué par comparaison avec une expérience de décodage sans bigramme, pour laquelle tous les modèles phonétiques sont équiprobables à chaque transition. Le tableau 2.6 donne les taux de décodage avec et sans bigramme phonétique avec le système de base à 8 gaussiennes par état. Le gain du taux de reconnaissance apporté par le bigramme est important, de l'ordre de 4 %. Le système de reconnaissance de base utilise ce bigramme.

Bigramme	Taux d'identification	Taux d'omission	Taux de substitution	Taux d'insertion	Taux de reconnaissance
non	62,8	13,6	23,6	5,4	**57,4**
oui	66,5	10,9	22,6	5,3	**61,2**

Tableau 2.6. Influence d'un bigramme phonétique sur les taux de décodage du noyau de test ET_{192} par le système de base SB_8.

2.5. Etude des densités de probabilité

Les densités de probabilité gaussiennes sont les représentations les plus utilisées pour modéliser les densités de probabilité dans les HMM continus; il s'agit en effet d'une loi plausible en l'absence d'hypothèse sur les données. Dans cette partie, nous étudions les avantages et les limitations de cette représentation, qui ont conduit les chercheurs à adopter une modélisation par combinaison linéaire de gaussiennes multi-variables (i.e., multi-gaussiennes) plutôt qu'une seule loi normale; cependant le choix du nombre optimal de gaussiennes reste généralement guidé par des heuristiques. Nous proposons un critère objectif et automatique pour ce choix, fondé sur le test du χ^2.

2.5.1. Estimation de densités multi-dimensionnelles

Il a été démontré qu'il est possible d'estimer n'importe quelle densité de probabilité en tout point de continuité dans un espace multi-dimensionnel par une combinaison linéaire de fonctions, dites "fonctions noyaux", vérifiant certaines contraintes [Parzen, 1962; Cacoullos, 1966]:

$$\frac{1}{n{h_n}^d}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h_n}\right) \xrightarrow{n\to\infty} f(x) \tag{2.1}$$

où $\begin{cases} X_1 \ldots X_n \text{ sont } n \text{ vecteurs de dimension } d \text{ et de distribution } f(x) \\ \text{la suite } \{h_n\} \text{ converge vers } 0 \\ \text{le noyau } K \text{ est borné, d' intégrale unitaire, et vérifie } \lim\limits_{|y|\to\infty} |y|^d K(y) = 0 \end{cases}$

Les gaussiennes représentent des noyaux acceptables [Dubuisson, 1990], et le choix de représenter une densité de probabilité par une combinaison linéaire de gaussiennes multi-dimensionnelles est donc validé. La convergence est même démontrée pour une fonction noyau multi-dimensionnelle définie comme le produit de noyaux mono-dimensionnels:

$$\frac{1}{n}\sum_{j=1}^{n}\left\{\prod_{i=1}^{d}\frac{1}{h_n}K\left(\frac{x^i - X_j^i}{h_n}\right)\right\} \xrightarrow{n\to\infty} f(x) \tag{2.2}$$

où $x = (x^1 \ldots x^d)$, ce qui nous assure qu'une densité de probabilité quelconque peut être représentée par un nombre suffisant de gaussiennes à matrices de covariances diagonales. Il s'agit cependant là d'une propriété asymptotique. Dans la réalité, on ne dispose que d'un nombre fini de mesures pour estimer la densité de probabilité, et la convergence n'est jamais atteinte.

Le problème qui se pose alors est de trouver la représentation qui est la mieux adaptée aux données disponibles, c'est-à-dire celle qui nécessite l'estimation du plus petit nombre de paramètres pour modéliser les données et dont la distribution est la plus fidèle au processus sous-jacent de génération des observations. Il existe de nombreuses possibilités pour l'estimation des densités de probabilité, parmi lesquelles les estimateurs de Parzen [Soudoplatoff, 1986] ou les réseaux de neurones. Nous allons dans un premier temps chercher à évaluer la qualité du choix gaussien, puis proposer des solutions pour améliorer la représentation des données.

2.5.2. Indépendance des coefficients

Dans les modèles testés lors de la mise au point du système de base, les densités de probabilité d'émission sont représentées exclusivement par des lois normales à matrice de covariance diagonale. Il est généralement considéré que les coefficients cepstraux calculés par la transformée en cosinus discrète sont faiblement corrélés entre eux (cf. §1.2.2), ce qui autorise en première approximation l'utilisation d'une matrice de covariance diagonale. Bien que la plupart des systèmes de reconnaissance de la parole existants fassent cette hypothèse, A. Ljolje montre qu'une gaussienne à matrice de covariance complète produit une vraisemblance plus élevée sur l'ensemble d'apprentissage qu'une combinaison de gaussiennes diagonales [Ljolje, 1994b]. L'indépendance des coefficients doit donc être vérifiée par l'étude de la matrice de corrélation. Des lois de probabilité à matrices de covariance non diagonales sont ensuite expérimentées sur le système de base.

2.5.2.1. Etude de la matrice de corrélation

La matrice de corrélation R est définie par normalisation de la matrice de covariance Σ comme:

$$R = \left[r_{ij}\right]_{1 \le i,j \le d}, \text{ avec } r_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{2.3}$$

La matrice R est symétrique, et ses termes diagonaux valent 1 par définition. Lorsque les coefficients sont non corrélés linéairement, les termes non diagonaux sont nuls, et R devient la matrice Identité.

Nous avons calculé la valeur des termes de la matrice de corrélation sur la totalité de l'ensemble d'apprentissage de la base TIMIT, pour nos coefficients acoustiques (soit plus d'un million de vecteurs de dimension d=26 pour trois heures de parole). Les résultats (cf. Figure 2.5) ne permettent pas de conclure à une complète indépendance des coefficients; ils montrent que les corrélations sont quasiment négligeables entre les coefficients cepstraux et les coefficients différentiels du premier ordre, mais elles existent à l'intérieur de ces deux groupes de coefficients; ainsi, la corrélation entre les coefficients C_1 et E est d'environ 0,5.

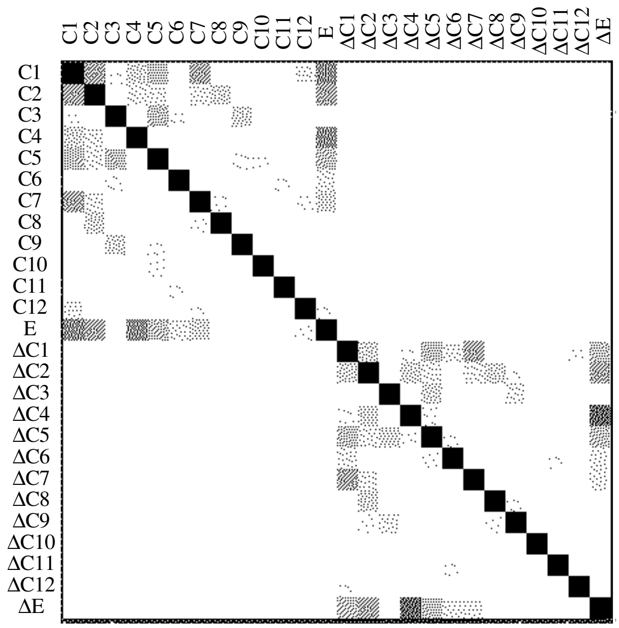

Figure 2.5. Matrice de corrélation des 26 coefficients sur l'ensemble d'apprentissage (les trames foncées indiquent une corrélation proche de 1 en valeur absolue, les trames claires des corrélations proches de zéro).

2.5.2.2. Matrice de covariance non diagonale

Le système de base a été comparé avec un système équivalent pour lequel les densités de probabilité sont modélisées par une combinaison linéaire de G gaussiennes à matrice de covariance non diagonale (cf. Tableau 2.7). Au delà de trois gaussiennes par état, les matrices de covariance deviennent singulières pour certains modèles peu représentés dans l'ensemble d'apprentissage, et il n'a pas été possible d'obtenir des résultats.

Pour un nombre identique de gaussiennes, les taux de décodage sont meilleurs pour les matrices de covariance non diagonale que pour les matrices de covariance diagonales. Cependant, le nombre de paramètres à estimer pour une gaussienne devient considérablement plus élevé: la moyenne est toujours représentée par un vecteur de dimension d, tandis que la matrice de covariance contient $d.(d+1)/2$ valeurs différentes, soit au total 377 paramètres par gaussienne et par état en dimension 26, au lieu de 52 paramètres dans le cas d'une gaussienne diagonale. De plus, l'inversion de toutes les matrices de covariance après chaque ré-estimation rend l'apprentissage des modèles considérablement plus long, à nombre de paramètres identique.

Pour un nombre réduit de paramètres du système de reconnaissance (cf. Figure 2.6), l'hypothèse diagonale reste acceptable, même si elle modélise moins finement la distribution des données. Des expériences menées sur les paramètres incluant la dérivée seconde conduisent aux mêmes conclusions. Les densités de probabilité du système de base sont donc construites avec des gaussiennes diagonales.

Nombre de gaussiennes *G*	Paramètres du système (milliers)	Phonèmes identifiés (%)	Phonèmes omis (%)	Phonèmes substitués (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
Matrice de covariance non diagonale						
1	54	65,2	12,0	22,9	5,7	**59,5**
2	108	67,9	11,3	20,8	5,1	**62,8**
3	163	69,2	10,6	20,2	5,1	**64,1**
Matrice de covariance diagonale						
1	7,5	58,9	13,6	27,5	5,7	**53,2**
2	15	62,5	12,2	25,3	5,7	**56,8**
4	30	65,0	11,3	23,7	5,8	**59,2**
8	60	66,5	10,9	22,6	5,3	**61,2**
12	90	67,5	10,3	22,2	5,5	**62,0**

Tableau 2.7. Influence de la structure des matrices de covariance sur les taux de décodage du noyau de test ET_{192} par le système de base.

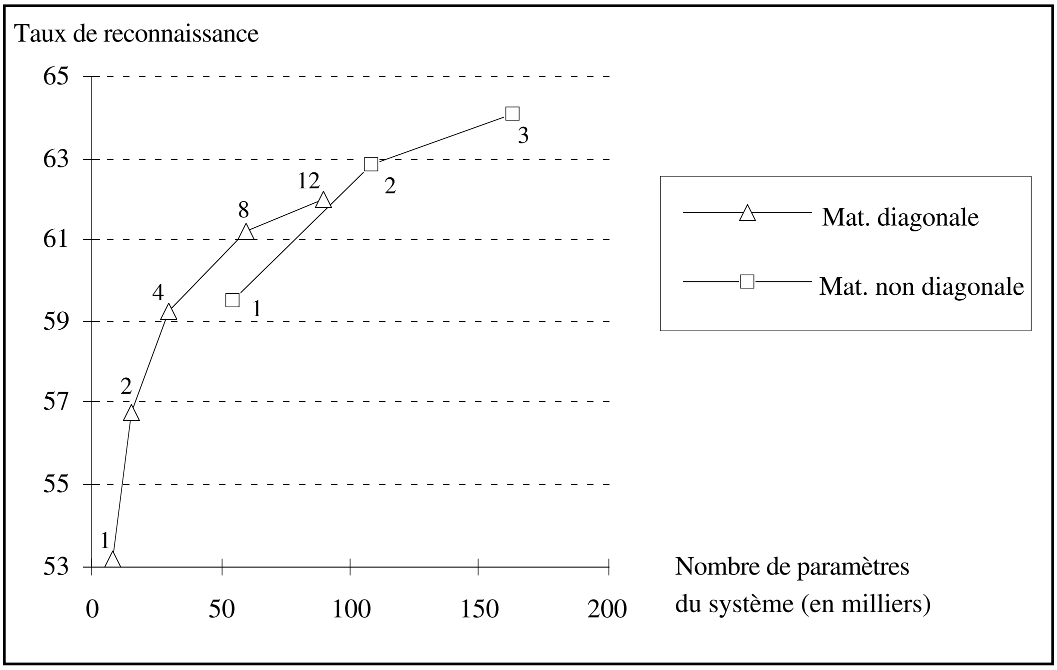

Figure 2.6. Influence de la structure des matrices de covariance sur les taux de reconnaissance de l'ensemble ET_{192}; pour chaque point, la valeur de *G* est précisée.

2.5.3. Limites de la représentation gaussienne

Lors de l'apprentissage du système SB_1, les coefficients sont représentés chacun indépendamment par une loi normale. Nous savons par l'expérience précédente que l'hypothèse d'une indépendance entre les coefficients est fausse, mais qu'elle peut être acceptable en première approximation. Afin de vérifier si les coefficients suivent une loi gaussienne, nous avons observé la distribution des coefficients et évalué leur répartition normale par un test statistique classique du χ^2.

2.5.3.1. Distribution des coefficients

Nous avons mesuré dans chaque état des modèles, pour chacun des 26 coefficients, la distribution des observations par un histogramme. Cet histogramme a été mis en correspondance avec la loi normale déduite de ces données. L'évaluation est réalisée sur les données d'apprentissage des modèles, ce qui devrait biaiser les résultats en faveur de l'hypothèse gaussienne. Nous présentons simplement des échantillons de ces mesures pour l'état final du modèle de 'iy' appris sur 4626 segments lors de l'apprentissage du système de base SB_1 (cf. Figure 2.7). Certains coefficients semblent assez bien modélisés, mais pour d'autres il est nécessaire d'affiner la représentation. On retrouve des résultats déjà observés sur des coefficients LPCC [Rabiner, 1989].

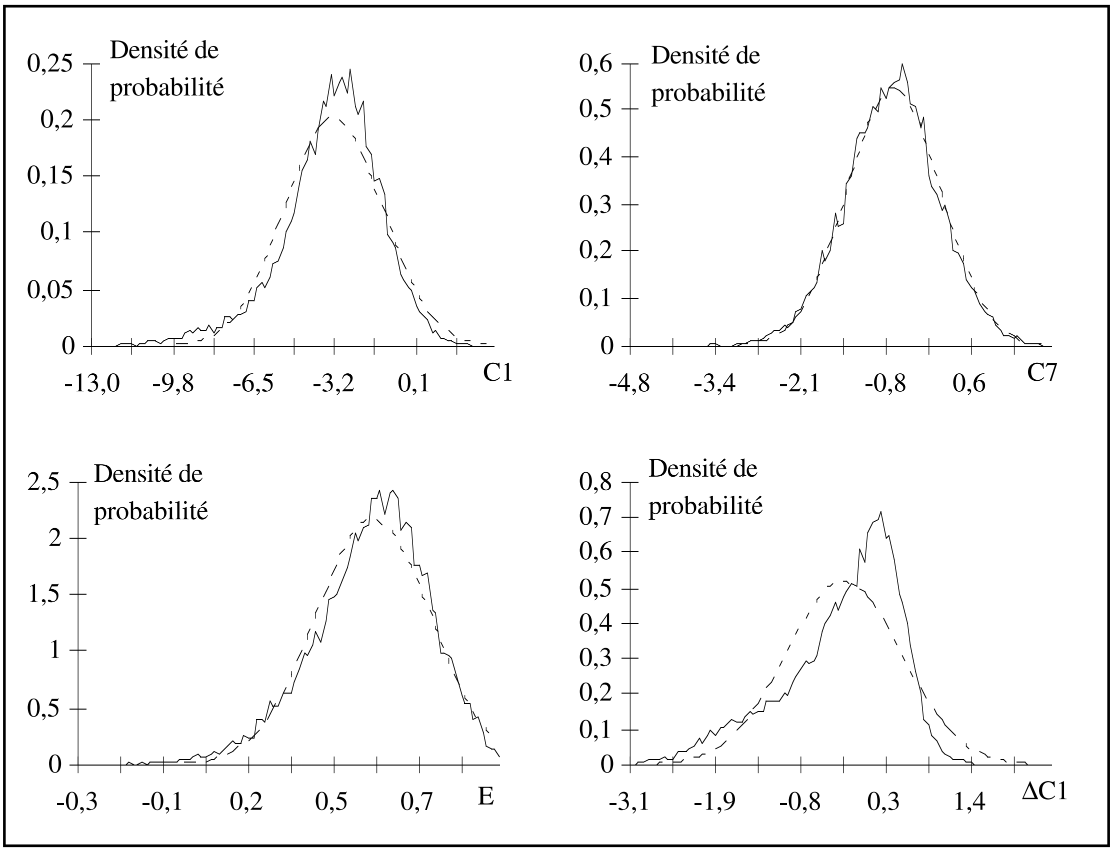

Figure 2.7. Densité de probabilité observée sur l'ensemble d'apprentissage (trait continu) et estimée par une gaussienne (trait pointillé) pour les coefficients C1, C7, E et ΔC1 de l'état final du modèle 'iy'.

2.5.3.2. Test du χ^2

L'observation des distributions des coefficients ne permet qu'un jugement subjectif de la bonne représentation des données par une gaussienne. Un critère plus objectif est donné par le test du χ^2. Il s'agit d'une statistique classique [Saporta, 1990]. Dans un espace à une dimension, pour vérifier que des données réelles $\{x_i\}_{1 \le i \le n}$ respectent une distribution gaussienne, l'espace de représentation $\mathbb{R}$ est découpé en C intervalles $I_1 \ldots I_C$, de probabilité *a priori* p_c; en pratique, on les choisit équiprobables, de probabilité p_c=1/C. Les n données, centrées et réduites, sont réparties dans ces intervalles, et le nombre d'échantillons est compté dans chaque intervalle. Alors, la quantité suivante suit un χ^2 à *C-3* degrés de liberté[1]:

$$Z_n = \sum_{c=1}^{C} \frac{(h_c - n.p_c)^2}{n.p_c} \text{ , avec } h_c = card\{x_i \,/\, x_i \in I_c\} \qquad (2.4)$$

Avec C=10 classes, nous mesurons un χ^2 à 7 degrés de liberté qui est, avec une probabilité d'au moins 95%, inférieur à 14:

$$P(Z_n \le 14) \ge 0,95 \qquad (2.5)$$

Si la valeur est supérieure à 14, on peut conclure que Z_n ne suit pas un χ^2 et que l'hypothèse d'une loi gaussienne est fausse.

Dans notre cas, l'algorithme de Baum-Welch calcule la contribution des observations acoustiques $O=(o_1,\ldots,o_T)$ sur tous les chemins possibles et donc la probabilité *a posteriori* que l'observation $o_t = (o_t^1,\ldots,o_t^d)$ ait été émise à l'instant t par l'état j. Cette probabilité a été notée $\omega_t(j)$ (cf. § 1.3.5). Un coefficient o_t^k de cette observation est donc affecté d'une pondération ω_t qui est la probabilité d'émission de sa trame par la gaussienne étudiée. La statistique sur le coefficient d'indice k est alors appliquée de la manière suivante: $h_c(k)$ est la somme des poids des trames dont le coefficient k est dans l'intervalle I_c, et n est remplacé par la somme totale des poids γ:

$$Z_\gamma(k) = \sum_{c=1}^{C} \frac{(h_c(k) - \gamma\, p_c)^2}{\gamma\, p_c} \quad \text{avec} \quad h_c(k) = \sum_{\{t/o_t^k \in I_c\}} \omega_t \quad \text{et} \quad \gamma = \sum_{t=1}^{T} \omega_t \qquad (2.6)$$

Pour décider de la validité de l'hypothèse gaussienne dans un espace multi-dimensionnel, il n'est pas concevable de découper l'espace $\mathbb{R}^d$ en hyper-cubes pour compter les échantillons; nous ne disposerions pas d'un nombre suffisant de données pour obtenir une statistique fiable. La statistique est donc calculée indépendamment pour chaque dimension du vecteur de paramètres. Si la matrice de covariance est diagonale et si le test du χ^2 est vérifié pour chacun des coefficients, alors on peut supposer que les coefficients sont indépendants entre eux et que chacun est bien représenté par une gaussienne. Si la matrice de covariance n'est pas diagonale, il devient nécessaire de décorréler les coefficients pour se ramener au cas précédent, ou bien d'utiliser des méthodes spécifiques permettant de tester la normalité des données dans un espace multi-dimensionnel [Smith & Jain, 1988].

---

[1] Si la moyenne et la variance de la loi normale sont connues a priori au lieu d'être estimées sur les échantillons, le χ^2 dispose alors de *C-1* degrés de liberté.

2.5.3.3. *Résultats des tests du χ^2*

Les expériences avec le test du χ^2 [Goût, 1993] confirment que le choix d'une mono-gaussienne diagonale n'est pas satisfaisant. Nous présentons les résultats pour les 26 coefficients des trois états d'un modèle de phonème 'iy' à une gaussienne par état, calculés sur l'ensemble d'apprentissage lors de l'apprentissage du système SB_1 (cf. Tableau 2.8). Les valeurs du test du χ^2 dans l'état final du modèle pour les coefficients C1, C7, E et ΔC1, qui sont respectivement de 466, 21, 280 et 2306, correspondent aux données présentées dans les histogrammes de la Figure 2.7. Nous observons que très peu de coefficients semblent suivre une loi normale: seules 8 valeurs sur les 78 sont inférieures à 14. Il n'y a aucun état pour lequel les 26 coefficients sont inférieurs au seuil.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	E
État 1	5	28	80	142	57	9	27	31	37	22	159	7	38
État 2	25	30	22	397	28	19	41	29	22	63	629	19	97
État 3	466	41	33	499	79	38	21	36	25	31	198	15	280
	ΔC1	ΔC2	ΔC3	ΔC4	ΔC5	ΔC6	ΔC7	ΔC8	ΔC9	ΔC10	ΔC11	ΔC12	ΔE
État 1	2790	159	62	44	30	100	29	15	69	14	106	56	712
État 2	83	105	75	144	105	74	77	100	37	61	67	75	277
État 3	2306	56	17	93	40	56	12	28	8	13	55	12	2100

Tableau 2.8. Valeur du test du χ^2 pour les 26 coefficients des 3 états du modèle 'iy'.

Il existe cependant une grande disparité entre les coefficients. Le pourcentage d'échecs des test du χ^2 sur tous les états des 48 modèles pour chacun des 26 coefficients (cf. Figure 2.8) montre que le premier coefficient cepstral, l'énergie, et leurs coefficients différentiels sont souvent très loin d'une loi de distribution gaussienne; mais les échecs sont fréquents pour tous les coefficients. Une modélisation plus fine est donc indispensable, par exemple la modélisation par multi-gaussiennes.

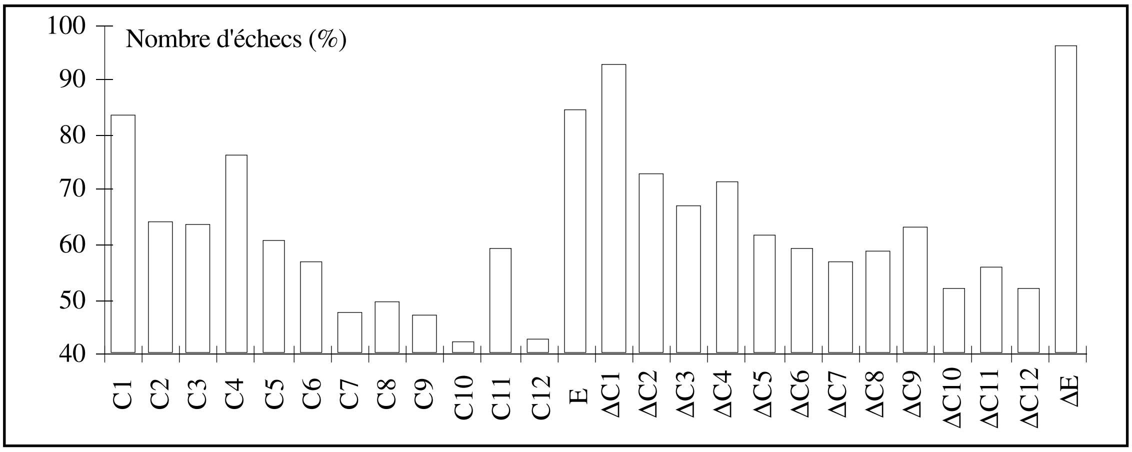

Figure 2.8. Pourcentage d'échecs du test du χ^2 pour les 26 coefficients calculé sur les 144 états du système SB_1.

2.5.4. Optimisation du nombre de gaussiennes

Sachant que des lois gaussiennes uni-modales ne suffisent pas à modéliser les données émises par un état d'une machine acoustique, on peut choisir de les remplacer par des lois multi-modales. Il est alors nécessaire de choisir le nombre de gaussiennes pour représenter chaque état, en faisant un compromis entre une bonne modélisation des données et le nombre limité de données d'apprentissage. Un nombre trop élevé de gaussiennes par rapport à la quantité de données disponibles conduit à un mauvais apprentissage, c'est-à-dire que les données d'apprentissage sont mémorisées et que les capacités de généralisation à de nouvelles données se dégradent. Différentes méthodes ont été proposées pour ce choix.

2.5.4.1. Méthodes classiques

La procédure préconisée par les concepteurs de HTK est la suivante: des modèles à une gaussienne par état sont appris, puis le nombre de gaussienne par état est augmenté progressivement, par clonage et perturbation des gaussiennes apportant la plus forte contribution au calcul de la densité de probabilité; enfin, les modèles sont ré-estimés. Plus précisément, si la probabilité d'émission dans un état est constituée de G gaussiennes (cf. équation 1.33) dont les paramètres g_k, μ_k et Σ_k ont été appris par l'algorithme de Baum-Welch:

$$b(o) = \sum_{k=1}^{G} g_k \mathcal{N}(o, \mu_k, \Sigma_k) \tag{2.7}$$

alors la gaussienne de poids maximum est choisie:

$$\tilde{k} = \arg\max_k g_k \tag{2.8}$$

et elle est divisée en deux gaussiennes dont les moyennes sont légèrement perturbées par le vecteur écart-type $\sigma = [\sigma_i]_{1 \le i \le d}$ déduit de la diagonale de la matrice de covariance Σ:

$$g_{\tilde{k}} \mathcal{N}(\mu_{\tilde{k}}, \Sigma_{\tilde{k}}) \xrightarrow[]{\text{clonage \& perturbation}} \begin{cases} \dfrac{g_{\tilde{k}}}{2} \mathcal{N}(\mu_{\tilde{k}} - 0{,}2\,\sigma_{\tilde{k}}, \Sigma_{\tilde{k}}) \\ \dfrac{g_{\tilde{k}}}{2} \mathcal{N}(\mu_{\tilde{k}} + 0{,}2\,\sigma_{\tilde{k}}, \Sigma_{\tilde{k}}) \end{cases} \tag{2.9}$$

La détection d'un maximum du taux de reconnaissance sur un ensemble de validation distinct de l'ensemble d'apprentissage permet de décider du nombre optimal de gaussiennes. Cette procédure présente deux inconvénients: la gaussienne qui est divisée en deux n'est pas nécessairement celle qui modélise le plus mal les données; de plus, le nombre de gaussiennes par densité de probabilité est *a priori* identique pour tous les états des modèles, car le critère de décision est global à l'ensemble des modèles.

D'autres procédures existent pour l'apprentissage de multi-gaussiennes. Ainsi, D. Jouvet, L. Mauuaury et J. Monné proposent une succession d'augmentations du nombre de gaussiennes suivies de fusions des gaussiennes les plus proches; au cours de cette dernière phase, celles qui sont estimées avec un nombre de trames trop faible, inférieur à quelques dizaines, sont supprimées [Jouvet *et al.*, 1991b]. Y. Zhao réalise une classification hiérarchique ascendante des trames acoustiques contribuant à une gaussienne, en limitant la variance intra-classe à un certain seuil [Zhao, 1993]. R. De Mori, M. Galler et F. Brugnara utilisent le recuit simulé pour optimiser le nombre de gaussiennes par état; leur recherche d'optimal est guidée par le taux de reconnaissance phonétique sur un ensemble de validation [De Mori *et al.*, 1995]. Enfin, si l'on sort du cadre d'une estimation par maximum de vraisemblance (MLE) pour laquelle une estimation correcte des probabilités est cruciale, la division des gaussiennes au cours d'un apprentissage optimisant le critère MMI semble la plus efficace, car les densités de probabilité ne sont alors affinées qu'aux frontières entre les classes [Normandin, 1995]. Nous proposons pour ce choix un critère de décision basé sur la statistique du χ^2.

2.5.4.2. Approche proposée

Lors de l'apprentissage d'un modèle, la contribution des observations acoustiques à chaque gaussienne composant une densité multi-gaussienne est calculée et permet de ré-estimer la moyenne et la variance de cette gaussienne (cf. équations 1.71 et 1.72):

$$\tilde{\mu} = \frac{1}{\gamma} \sum_{t=1}^{T} \omega_t o_t \tag{2.10}$$

et:

$$\tilde{\Sigma} = \frac{1}{\gamma} \sum_{t=1}^{T} \left\{ \omega_t (o_t - \mu)(o_t - \mu)^t \right\} \tag{2.11}$$

Le test du χ^2 nous semble un bon critère *a priori* pour choisir la ou les gaussiennes de la multi-gaussienne qui modélisent mal les données d'apprentissage. Ces gaussiennes sont ensuite divisées suivant la procédure (2.9) et les modèles sont ré-estimés.

Il nous a cependant paru irréaliste d'imposer la réussite du test sur toutes les dimensions simultanément; nous imposons donc que la valeur moyenne des statistiques calculées sur toutes les dimensions respecte le seuil, mais nous autorisons éventuellement l'échec du test pour un nombre limité $D_{\max}$ de coefficients, avec $D_{\max} \leq d$. Nous décidons donc que la représentation par une gaussienne est satisfaisante si:

$$\begin{cases} \operatorname{card}\left\{ k / Z_\gamma(k) > 14 \right\} \leq D_{\max} \\ \left(\bar{Z}_\gamma = \frac{1}{d} \sum_{k=1}^{d} Z_\gamma(k) \right) \leq 14 \end{cases} \tag{2.12}$$

De plus, un nombre minimal de trames acoustiques $Tr_{\min}$ doit contribuer au calcul de la gaussienne que l'on souhaite diviser. L'intérêt de ce seuil, que l'on peut fixer à quelques dizaines, voire quelques centaines de trames, est de permettre une estimation assez fiable des gaussiennes. En effet, si $\tilde{\mu}$ et $\tilde{\sigma}^2$ sont estimés d'après n échantillons distribués suivant la loi $\mathcal{N}(\mu,\sigma^2)$, l'intervalle de confiance à 95% est:

$$\begin{cases} P(\left|\tilde{\mu}-\mu\right| \leq 1{,}96\dfrac{\sigma}{\sqrt{n}}) \geq 0{,}95 \\ P(\left|\tilde{\sigma}^2-\sigma^2\right| \leq 1{,}96\dfrac{\sqrt{2}\sigma^2}{\sqrt{n-1}}) \geq 0{,}95 \end{cases} \tag{2.13}$$

Avec 50 échantillons, une erreur de 40% sur l'estimation de la variance σ^2 est encore possible; il est donc souhaitable de disposer d'un nombre élevé d'échantillons.

La procédure que nous proposons est automatique, et estime les densités de probabilité de chaque état en tenant compte simultanément de la dispersion des données ainsi que de leur quantité. Nous l'appliquons sur les données d'apprentissage, car le calcul de la statistique s'intègre avec un surcoût quasiment nul à la ré-estimation des modèles par l'algorithme de Baum-Welch (que ce soit en modèles isolés avec HRest ou en parole continue avec HERest). Il est tout à fait envisageable de pratiquer ce test sur un ensemble de validation distinct de l'ensemble d'apprentissage, pour éviter le sur-apprentissage. Ce risque est cependant réduit dans notre cas par l'utilisation d'un seuil élevé pour imposer un nombre minimal de trames par gaussienne.

2.5.4.3. Résultats expérimentaux

Nous avons comparé l'effet de la procédure classique et de la procédure que nous proposons pour augmenter le nombre de gaussiennes sur un signal synthétique. La distribution de ce signal est obtenue par la combinaison linéaire de trois gaussiennes mono-dimensionnelles dont les pondérations $\{g_k\}$, les moyennes $\{\mu_k\}$ et les variances $\{\sigma^2{}_k\}$ sont données dans le Tableau 2.9.

k	g_k	μ_k	$\sigma^2{}_k$
1	0,6	0,0	1,0
2	0,3	-2,0	3,0
3	0,1	3,5	0,5

Tableau 2.9. Paramètres de la multi-gaussienne ayant servi à générer les données synthétiques.

Un HMM à un seul état représenté par une mono-gaussienne a été appris sur 5167 échantillons, et le nombre de gaussiennes a été augmenté en suivant la procédure standard ainsi que la procédure basée sur la statistique du χ^2. Dans chaque cas, le logarithme de la probabilité d'émission des données par les modèles a été calculé. L'évolution de cette probabilité en fonction du nombre de gaussiennes a été comparée à la probabilité d'émission par le modèle idéal ayant généré les données (cf. Figure 2.9). La procédure que nous proposons converge à 7 gaussiennes, tandis que la procédure standard n'atteint une modélisation comparable que pour 16 gaussiennes.

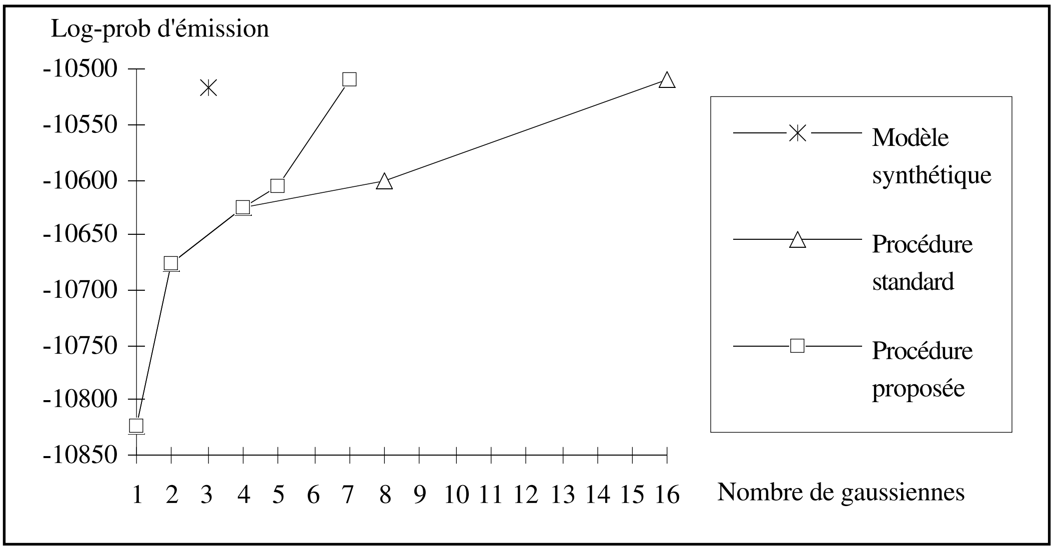

Figure 2.9. Evolution de la probabilité d'émission en fonction du nombre de gaussiennes pour les différentes procédures d'apprentissage.

L'observation confirme que la modélisation des données synthétiques par 8 gaussiennes obtenues par la procédure classique n'est pas aussi satisfaisante que celle avec 7 gaussiennes obtenues par la procédure automatique basée sur le test du χ^2 (cf. Figure 2.10). Ces résultats prometteurs nous ont conduit à appliquer cette nouvelle procédure lors de l'apprentissage du système de reconnaissance.

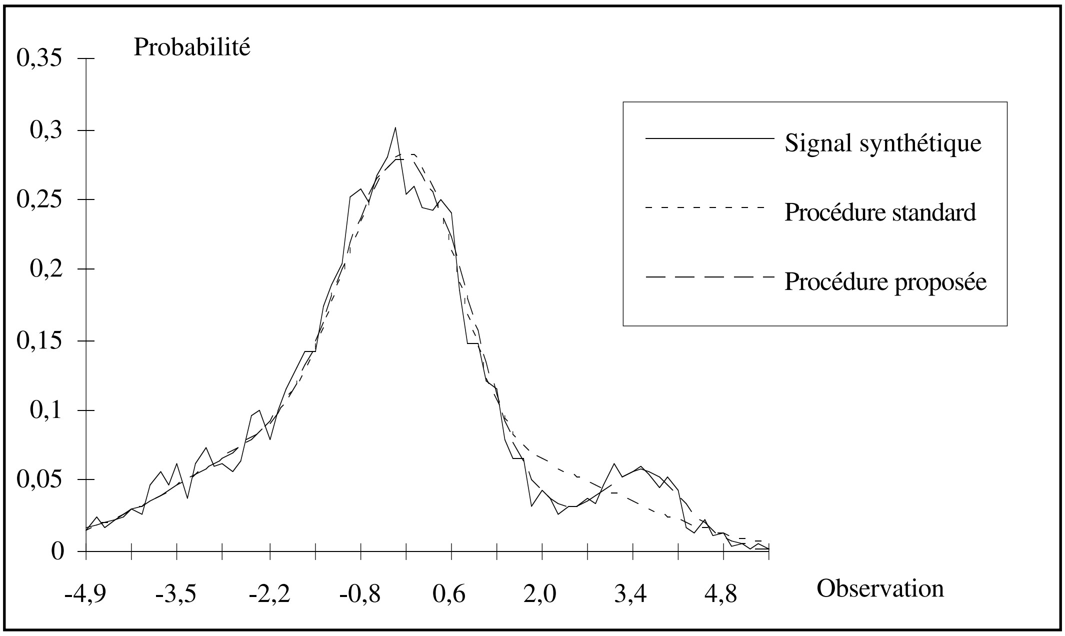

Figure 2.10. Distribution des données synthétiques et lois de probabilité estimées par la procédure standard (8 gaussiennes) et la procédure proposée (7 gaussiennes).

Nous avons réalisé l'apprentissage du système de base avec la nouvelle procédure, en choisissant des valeurs de seuils assez élevées ($Tr_{\min}$=500 et $D_{\max}$=26), ce qui permet de limiter le nombre de gaussiennes du système. Comparé avec l'apprentissage standard, la nouvelle procédure permet d'atteindre des taux de décodage comparables pour un nombre total de paramètres largement inférieur (cf. Tableau 2.10 et Figure 2.11). Nous espérons donc réaliser grâce à cette nouvelle approche l'apprentissage d'un système de reconnaissance avec un nombre plus réduit de paramètres.

Itération	Nombre de gaussiennes par état *G*	Nombre total de gaussiennes	Paramètres du système (milliers)	Phonèmes reconnus (%)
Procédure standard de division des gaussiennes				
1	1	144	7,5	53,2
2	2	288	15,0	56,8
3	4	576	30,0	59,2
4	8	1152	60,0	61,2
Procédure automatique basée sur le test du χ^2				
1	1	144	7,5	53,2
2	≤2	280	14,6	57,0
3	≤4	482	25,1	58,9
4	≤8	703	36,6	61,4
5	≤16	845	43,9	61,4

Tableau 2.10. Apprentissage du système de base par la procédure standard de division des gaussiennes et par la procédure automatique basée sur la statistique du χ^2; décodage du noyau de test ET_{192}.

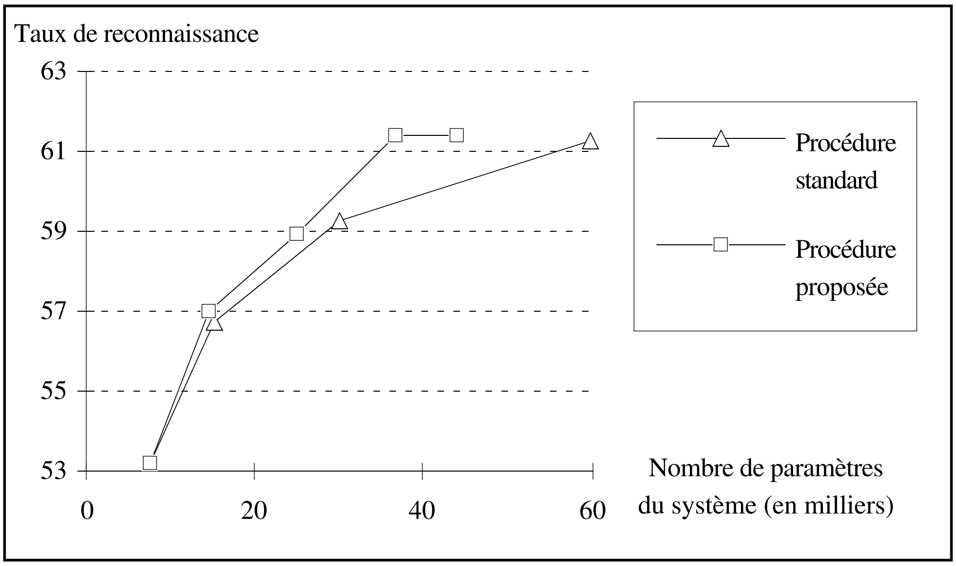

Figure 2.11. Apprentissage du système de base par la procédure standard et la procédure proposée: taux de reconnaissance en fonction du nombre total de paramètres

2.5.5. Représentation non paramétrique

La représentation des données par des multi-gaussiennes a permis d'améliorer de plus de 8 % en valeur absolue les taux de reconnaissance par rapport à une représentation mono-gaussienne (cf. Tableau 2.4), ce qui est important. Il faut noter que les coefficients MFCC semblent bien adaptés à une telle représentation, puisqu'ils sont faiblement corrélés, que leur distribution reste assez proche d'une distribution normale (cf. Figure 2.7), et qu'une distance euclidienne pondérée leur est naturellement associée (cf. §1.2.3). Pourtant, certaines représentations non paramétriques ont montré leur efficacité, en particulier celle obtenue en sortie d'un réseau connexioniste; A. Robinson obtient ainsi les meilleurs taux actuels de DAP sur la base TIMIT avec un système hybride intégrant un réseau de neurones récurrent dans une structure markovienne [Robinson, 1994]. L'avantage des méthodes non paramétriques est d'éviter toute hypothèse éventuellement fausse sur la distribution des coefficients.

2.5.5.1. Systèmes hybrides et réseaux de neurones

Les réseaux de neurones [Lippmann, 1987] ont pendant un certain temps été considérés comme une alternative à la reconnaissance de parole par DTW ou par HMM. La modélisation markovienne a de nombreuses faiblesses: en particulier, l'apprentissage des modèles est généralement réalisé suivant le critère du maximum de vraisemblance, qui est faiblement discriminant; de plus, la structure des densités de probabilité est choisie *a priori* et n'est pas toujours adaptée aux données. A l'inverse, un perceptron multi-couches (PMC), disposant d'une couche d'entrée, d'une couche de sortie et d'au moins une couche cachée de taille suffisante, est capable d'estimer n'importe quelle fonction non-linéaire discriminante; les réseaux de neurones permettent ainsi une identification discriminante de trames acoustiques isolées. L'algorithme de rétro-propagation du gradient réalise l'apprentissage d'un perceptron multi-couches en minimisant l'erreur quadratique en sortie; la convergence peut être accélérée par un apprentissage stochastique proposé par Y. Le Cun et dont les propriétés ont été étudiées par L. Bottou [Bottou, 1991].

Cependant, la parole est un processus séquentiel. La prise en compte du temps a été proposée dans certains réseaux, comme les réseaux récurrents ou les réseaux à délai (*Time Delay Neural Networks* ou TDNN). A. Waibel et ses collaborateurs ont montré l'efficacité d'un TDNN pour identifier l'occlusive initiale d'une syllabe isolée [Waibel *et al.*, 1989]. Malgré ces efforts, l'alignement temporel dynamique reste plus efficace pour la segmentation de la parole continue. Actuellement, les réseaux de neurones sont le plus souvent utilisés en parole dans des architectures hybrides, associés à un module d'alignement temporel ou à des modèles de Markov [Haton, 1995]. H. Bourlard et C. Wellekens montrent que, sous certaines conditions d'apprentissage, la sortie d'un PMC est une estimation de la probabilité *a posteriori* d'une classe sachant l'observation [Bourlard & Wellekens, 1990]; cette propriété est étudiée dans [Richard & Lippmann, 1991]. Le réseau peut donc être intégré dans un modèle de Markov pour estimer de manière discriminante les probabilités d'émission associées aux états ou aux transitions. Des expériences sur la tâche *Resource Management*, constituée de phrases en vocabulaire limité et avec une syntaxe très contrainte pour l'interrogation d'une base de données navales, montrent qu'un système hybride HMM/PMC est compétitif avec les meilleurs systèmes classiques en reconnaissance de la parole continue [Bourlard & Morgan, 1993].

D'autres systèmes hybrides ont été développés. A. Robinson et F. Fallside ont intégré un réseau récurrent dans un système de décodage phonétique par alignement dynamique; la sortie du réseau estime la probabilité que la trame observée appartienne à l'une des 61 classes phonétiques d'apprentissage, et la rétroaction à l'intérieur du réseau permet de modéliser les effets contextuels à long terme de la parole [Robinson & Fallside, 1991]. Leur système présente actuellement les meilleurs taux de DAP sur la base de référence TIMIT [Robinson, 1994]. A. Mellouk et P. Gallinari construisent un système neuronal prédictif discriminant. Dans chaque état de chaque modèle phonétique, un PMC prédit la trame courante en fonction des trames adjacentes; un alignement dynamique entre les trames prédites et observées permet le décodage phonétique [Mellouk & Gallinari, 1993]. L. Devillers utilise des perceptrons multi-couches à délai (TDNN) pour estimer les probabilités d'émission dans des HMM [Devillers, 1992]; elle obtient une diminution de 20% des taux d'erreurs en mots sur la tâche *Resource Management* par rapport aux modèles purement markoviens. Y. Bengio, R. De Mori, G. Flammia et R. Kompe réalisent un système hybride, avec des classifieurs neuronaux qui estiment le lieu d'articulation ou la macro-classe phonétique; les sorties de ces réseaux constituent les observations d'un HMM [Bengio *et al.*, 1992]. Les réseaux de neurones permettent une représentation des trames acoustiques observées plus compacte que les représentations classiques par MFCC, et les taux de reconnaissance sur 8 classes phonétiques sont significativement supérieurs à ceux d'un système markovien standard. R. De Mori souligne l'importance d'une expertise acoustique et phonétique pour concevoir des réseaux de neurones adaptés à des tâches de reconnaissance de la parole [De Mori, 1993].

2.5.5.2. Intégration d'une distance perceptive

Si l'on cherche à utiliser d'autres coefficients que les MFCC, comme des coefficients issus d'un modèle perceptif associés à une distance adaptée, rien ne permet de penser qu'une représentation par multi-gaussiennes soit pertinente. La discrimination entre voyelles par les méthodes classiques d'identification est généralement peu performante. La distance APS [Caraty, 1987] semble permettre une bonne discrimination entre voyelles (cf. §1.2.3.2). Une expérience a été réalisée sur la base BDSONS pour identifier 1980 segments parmi 12 voyelles du français, et la distance APS a donné des résultats supérieurs à une distance euclidienne associée à des coefficients LPCC, ainsi qu'à la distance d'Itakura entre des modèles auto-régressifs [Caraty *et al.*, 1992]. L'intégration de la distance APS dans un système de reconnaissance par HMM a été proposée au LAFORIA dans un travail de C. Lecordier [Lecordier, 1995], mais ces recherches restent pour l'instant à un stade expérimental.

2.6. Apprentissage de modèles contextuels

Le système de référence est construit à partir du système de base pour permettre une meilleure modélisation acoustique. Les choix faits pour le système de base, à savoir la représentation du signal, la topologie des modèles, la modélisation des densités de probabilité par une combinaison linéaire de gaussiennes diagonales, et le bigramme phonétique comme modèle de langage, sont conservés. Le système de référence prend en compte la coarticulation en modélisant des allophones contextuels, et le nombre de gaussiennes par densité de probabilité est déterminé par la procédure automatique déjà décrite. Enfin, les performances en décodage acoustico-phonétique du système développé sont comparées à celles des systèmes existants.

2.6.1. Prise en compte du contexte

Les modèles phonétiques indépendants du contexte ne permettent pas une bonne représentation des caractéristiques acoustiques, en raison des phénomènes de coarticulation. Il est préférable de travailler avec des modèles tenant compte des contextes phonétiques gauche et droit. L'apprentissage de modèles contextuels est cependant plus difficile, car il n'est pas toujours possible de disposer d'un nombre de représentants suffisant pour l'apprentissage de chaque modèle d'allophone.

2.6.1.1. Mise en œuvre

Avec la prise en compte des contextes gauche et droit, le nombre de modèles peut s'élever à plusieurs dizaines de milliers. Il est difficile de disposer de suffisamment de données pour apprendre de manière robuste la totalité des modèles d'allophones possibles. Pour contourner cette difficulté, il est possible de regrouper plusieurs contextes produisant un effet similaire sur la réalisation acoustique du phonème; le contexte d'un modèle est constitué d'une classe de phonèmes plutôt que d'un seul phonème. On parle dans ce cas d'allophones généralisés (cf. §1.4.1.3). HTK offre des facilités pour l'apprentissage de modèles contextuels généralisés (dépendants du contexte phonétique gauche et/ou droit). Un algorithme de classification ascendante, utilisant une mesure de distance entre lois gaussiennes, permet de réaliser automatiquement une classification des états, et de lier toutes les probabilités d'émission des états d'une classe à une seule densité de probabilité. Cela entraîne une réduction du nombre de densités de probabilité à estimer et une estimation plus robuste de celles-ci. Nous présentons l'apprentissage de modèles phonétiques contextuels tel que nous l'avons réalisé avec HTK, en apprenant d'abord des allophones dépendant uniquement du contexte gauche ou du contexte droit, puis des allophones gauche et droit.

2.6.1.2. Allophones contextuels

Les allophones dépendants du contexte permettent une meilleure modélisation acoustique que des modèles indépendants du contexte. K.F. Lee et H. Hon rapportent une amélioration de près de 10% en valeur absolue du taux de reconnaissance phonétique sur la base TIMIT en introduisant des modèles en contexte droit [Lee & Hon, 1989]. Des allophones droits généralisés peuvent être appris avec HTK de la manière suivante:

- Apprentissage de modèles phonétiques indépendants du contexte.
- Traduction des étiquettes pour disposer d'étiquettes contextuelles; l'allophone du phonème φ en contexte phonétique droit d est par exemple noté φ_d.
- Tous les allophones droits présents dans l'ensemble d'apprentissage sont créés à partir du modèle phonétique indépendant du contexte correspondant. Les probabilités d'émission de l'état initial et de l'état central, ainsi que les probabilités de transition sont identiques pour tous les allophones d'un même phonème; seule la probabilité d'émission de l'état final est spécifique du contexte phonétique (cf. Figure 2.12).
- Ré-estimation des modèles sur de la parole continue par l'outil HERest, qui tient compte des liens entre lois de probabilités.
- Pour chaque phonème, une Classification Hiérarchique Ascendante (CHA) des états finaux de tous les modèles d'allophones, réalisée avec l'éditeur de modèles HHEd, permet de réduire le nombre de contextes; tous les allophones dans un contexte phonétique similaire sont modélisés par une seule machine.
- Ré-estimation des modèles; les allophones peuvent ensuite être améliorés par une augmentation du nombre de gaussiennes suivie de la ré-estimation des modèles. Dans notre cas, cette augmentation est contrôlée par le test du χ^2.

On dispose alors d'un nombre réduit de modèles en contexte droit. Le nombre total de modèles dépend principalement du seuil D_{IC} choisi comme distance intra-classe maximale lors de la CHA; la fusion des deux classes les plus proches est réalisée tant que la distance maximale entre deux états de ces deux classes ne dépasse pas le seuil D_{IC}. Un autre seuil $Tr_{\min}$ permet de supprimer les classes pour lesquelles le nombre de trames acoustiques participant au calcul de la loi de probabilité est trop faible. Quelques expériences sont nécessaires pour fixer ces seuils.

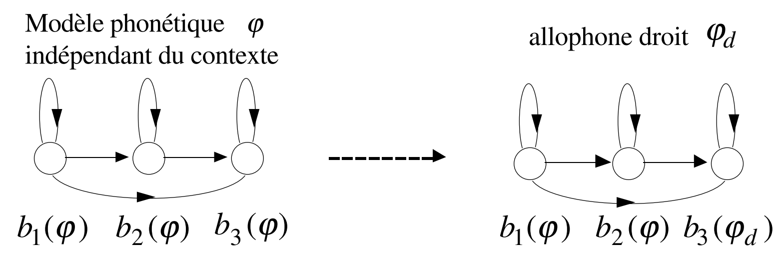

Figure 2.12. Création d'un modèle contextuel à partir d'un modèle indépendant du contexte.

La procédure décrite s'applique naturellement à l'apprentissage d'allophones en contexte gauche; mais d'autres approches pour l'apprentissage de modèles contextuels peuvent être choisies, y compris en utilisant les outils de HTK. Le regroupement des contextes phonétiques peut par exemple être obtenu à partir de connaissances *a priori* plutôt que par une classification automatique [Ljolje, 1994a].

2.6.1.3. Allophones en contexte gauche et droit

Nous avons choisi de créer des modèles d'allophones gauche et droit à partir de modèles d'allophones gauches et d'allophones droits en suivant la procédure proposée par S. Young [Young, 1992b]. Les étapes de l'apprentissage sont les suivantes:

- Apprentissage d'allophones gauches et d'allophones droits généralisés.
- Traduction des étiquettes en contexte gauche et droit.
- Création des allophones gauche et droit par fusion de modèles contextuels gauche et droit avec l'éditeur de modèles HHEd. La probabilité d'émission de l'état initial est celle du modèle contextuel gauche, celle de l'état final provient du modèle en contexte droit, et celle de l'état central reste indépendante du contexte (cf. Figure 2.13).
- Les modèles sont ré-estimés sur les phrases non segmentées par la procédure de Baum-Welch avec l'outil HERest.

Ces modèles sont en fait des modèles hybrides issus de modèles plus simples, et sont appelés quasi-triphones. La modélisation indépendante du début et de la fin du modèle permet d'estimer chaque densité de probabilité sur plus de données; il est même possible de créer des modèles pour des quasi-triphones qui n'ont pas été observés dans l'ensemble d'apprentissage. Les effets de la coarticulation sont cependant modélisés sur un seul état; il est possible que cela soit insuffisant.

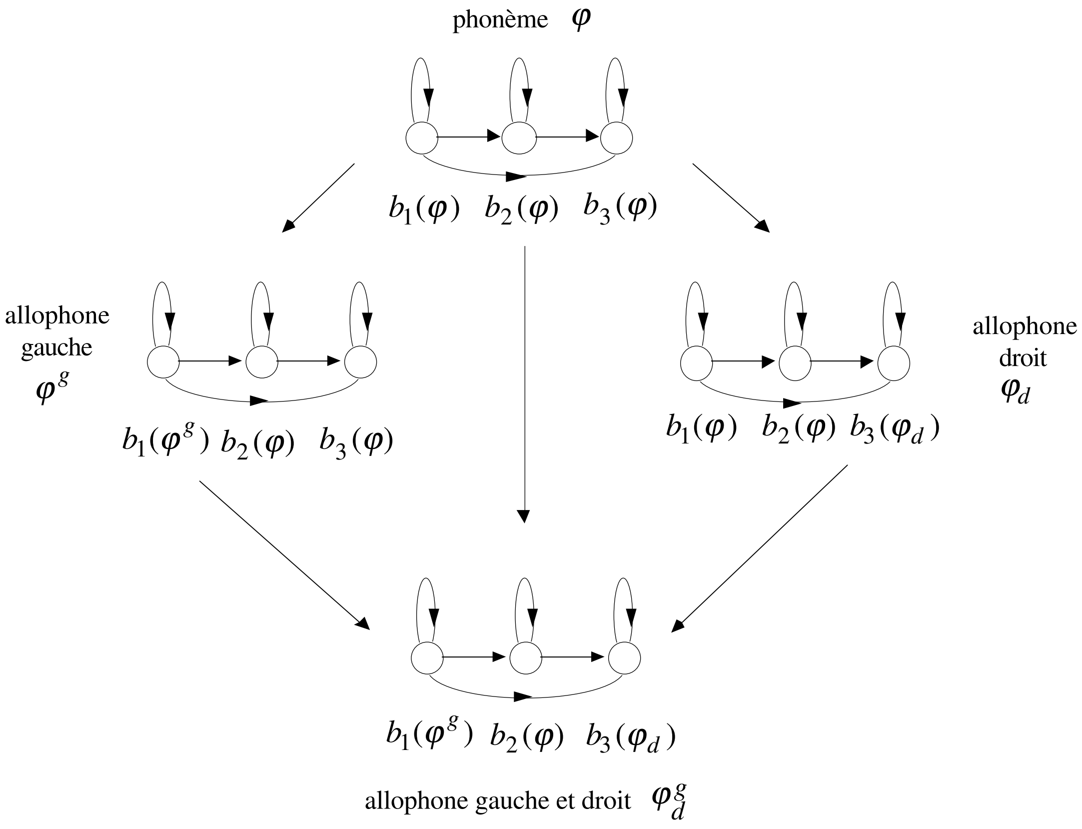

Figure 2.13. Création d'un modèle d'allophone gauche et droit à partir de modèles dépendants d'un seul contexte.

2.6.2. Apprentissage des modèles

La méthode d'apprentissage de modèles contextuels a été appliquée aux modèles phonétiques du système de base. Les gains apportés par les quasi-triphones dans le cadre de ces expériences sont médiocres, probablement en raison d'hypothèses trop fortes sur l'indépendance entre le début et la fin de l'allophone et éventuellement d'une quantité de données d'apprentissage insuffisante. Nous avons donc préféré (cf. § 2.6.2.3) de simples modèles d'allophones droits, dont les densités de probabilité ont été améliorées par la procédure automatique utilisant le test du χ^2.

2.6.2.1. Allophones droits

Si l'on considère les contextes droits des 48 classes phonétiques, l'ensemble d'apprentissage de la base TIMIT contient 1690 allophones différents sur les 48^2=2304 possibles. Nous avons appris 1690 modèles d'allophones en contexte droit en suivant la procédure décrite §2.6.1. Chaque densité de probabilité est modélisée par une simple gaussienne, et les allophones d'un même phonème ne diffèrent que par la densité de probabilité associée au dernier état. Les résultats comptés sur 39 classes phonétiques (cf. Tableau 2.11) peuvent sembler décevants, puisque le système obtenu contient un nombre élevé de paramètres (de l'ordre de 93000), alors que ses performances en reconnaissance ne sont que légèrement supérieures au système de base SB_1. La nette amélioration du taux d'identification est masquée dans le taux global de reconnaissance par l'augmentation des insertions.

Nombre de modèles	Paramètres du système (milliers)	Phonèmes identifiés (%)	Phonèmes omis (%)	Phonèmes substitués (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
Modèles indépendants du contexte						
48	7,5	58,9	13,6	27,5	5,7	**53,2**
Modèles dépendants du contexte droit						
1690	93	68,6	6,1	25,3	12,7	**55,9**

Tableau 2.11. Prise en compte des contextes droits sans regroupements contextuels.

2.6.2.2. Allophones généralisés

L'apprentissage d'un modèle pour chacun des contextes différents ne semble pas très efficace. Mais les 1690 diphones de la base ne sont pas de fréquence identique. Nous avons décompté la couverture de cette base par les allophones droits de fréquence décroissante (cf. Figure 2.14). Nous voyons ainsi que les 600 allophones les plus fréquents couvrent 90 % de la base d'apprentissage; les autres allophones sont représentés chacun par moins de 50 échantillons, ce qui ne permet pas un apprentissage robuste. Un regroupement d'allophones par contextes similaires est donc nécessaire.

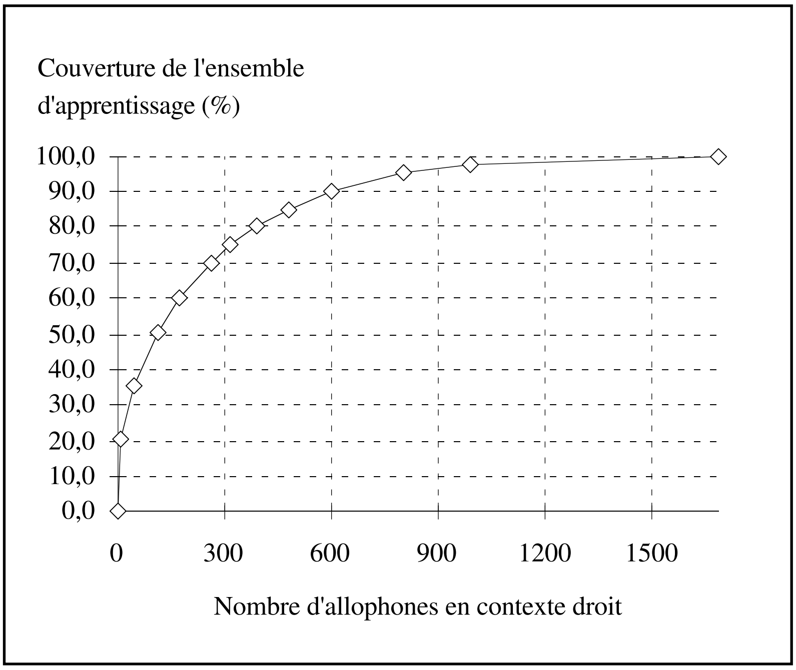

Figure 2.14. Couverture de l'ensemble d'apprentissage par les allophones droits les plus fréquents.

Nous avons appliqué la fusion des contextes proposée au § 2.6.1.2. Plusieurs valeurs du seuil inter-classes D_{IC} ont été expérimentées, avec un nombre minimal de trames par état fixé arbitrairement à 100 trames (cf. Tableau 2.12). Le seuil retenu de 0,3 entraîne une très forte réduction du nombre de modèles, de 1690 à 655 seulement.

Seuil D_{IC}	Nombre de modèles	Taux d'identification	Taux d'omission	Taux de substitution	Taux d'insertion	Taux de reconnaissance
0,4	542	67,7	6,3	26,0	13,3	**54,4**
0,3	655	68,8	6,0	25,2	13,5	**55,3**
0,2	752	68,9	5,8	25,4	13,7	**55,2**

Tableau 2.12. Choix du seuil de proximité D_{IC} pour le regroupement des contextes.

Pour une valeur de D_{IC} fixée à 0,3, nous avons ensuite fait varier le nombre minimal de trames par gaussienne $Tr_{\min}$, et constaté qu'une valeur assez élevée de 300 trames permettait de dépasser les performances des modèles contextuels sans classification, avec un nombre de modèles réduit à 404 modèles, soit le quart du nombre initial d'allophones (cf. Tableau 2.13). Le nombre d'allophones par classe phonétique varie entre un seul modèle pour représenter le phonème 'zh' et 18 modèles pour représenter tous les contextes du phonème 'iy'.

Seuil $Tr_{\min}$	Nombre de modèles	Taux d'identification	Taux d'omission	Taux de substitution	Taux d'insertion	Taux de reconnaissance
100	655	68,8	6,0	25,2	13,5	**55,3**
200	489	68,2	6,0	25,7	12,6	**55,6**
300	404	67,9	6,3	25,9	11,4	**56,5**
400	337	67,6	6,5	25,9	11,1	**56,5**

Tableau 2.13. Choix du nombre minimal de trames par état $Tr_{\min}$ pour le regroupement des contextes.

2.6.2.3. Quasi-triphones

Les 404 modèles droits obtenus après l'étape de classification ont été ré-estimés, en divisant progressivement les gaussiennes dont le test sur le χ^2 moyen ne satisfait pas le critère choisi, mais qui disposent d'au minimum 300 trames d'apprentissage (cf. §2.5.3.2). Des allophones gauches ont été appris de la même manière. Enfin, des modèles phonétiques en contextes gauches et droits ont été fabriqués par fusion des deux premières catégories de modèles suivie de quelques cycles de ré-estimations (cf. §2.6.1.3). Les taux de décodage sont donnés dans le Tableau 2.14.

Nombre total de gaussiennes	Paramètres du système (milliers)	Phonèmes identifiés (%)	Phonèmes omis (%)	Phonèmes substitués (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
404 modèles dépendants du contexte droit						
2070	108	73,2	5,3	21,5	10,4	**62,8**
383 modèles dépendants du contexte gauche						
2096	109	72,1	5,5	22,3	10,2	**61,9**
Modèles dépendants du contexte gauche et droit						
2726	142	74,0	4,4	21,6	11,9	**62,1**

Tableau 2.14. Modèles en contexte gauche et/ou droit.

Les taux de reconnaissance des modèles droits sont légèrement meilleurs que ceux des modèles gauches; cela a été observé par K.F. Lee qui attribue la différence à une meilleure identification des occlusives pré-vocaliques [Lee, 1988] et par A.-M. Derouault, pour qui l'anticipation prime sur la rémanence [Derouault, 1987]. Les quasi-triphones ont des performances inférieures aux modèles d'allophones droits, ce qui peut signifier que l'hypothèse d'une indépendance entre le début et la fin d'une réalisation est trop forte. Les systèmes les plus performants de CMU [Lee, 1988], de BBN [Schwartz *et al.*, 1989] ou d'AT&T [Ljolje, 1994a] combinent des modèles phonétiques indépendants du contexte, des allophones gauches ou droits et des allophones dépendants simultanément des contextes gauche et droit, ce qui n'est pas le cas de notre système. Au vu de ces résultats, nous avons décidé de continuer nos expériences avec les seuls modèles en contexte droit, et d'améliorer encore leurs densités de probabilité en utilisant un critère plus contraignant pour évaluer le caractère gaussien des données.

2.6.2.4. Amélioration des densités de probabilité

Nous avons repris l'apprentissage des 404 modèles d'allophones droits, en augmentant progressivement le nombre de gaussiennes par état. Le critère pour la division des gaussiennes a été modifié, de manière à n'autoriser le dépassement du seuil que pour 3 des 26 coefficients du vecteur de paramètres (D_{max}=3); un minimum de 300 trames par gaussienne est toujours nécessaire pour autoriser la division de la gaussienne (Tr_{min}=300). Les modèles ont été testés sur le noyau de test de 192 phrases ainsi que sur un ensemble de validation EA_{195} constitué de 195 phrases tirées aléatoirement dans l'ensemble d'apprentissage EA; dans les deux cas l'intervalle de confiance est légèrement supérieur à 1 %. Le nombre total de gaussiennes se stabilise à 3584 gaussiennes pour 9 cycles d'augmentation des gaussiennes et de ré-estimation des modèles, et le taux de reconnaissance global atteint un pallier à près de 72 % sur l'apprentissage et un peu plus de 65% sur le test (cf. Tableau 2.15).

Itération	Nombre total de gaussiennes	Milliers de paramètres du système	Phonèmes reconnus (validation)	Phonèmes reconnus (test)
0	500	26,0	58,7	56,5
1	946	49,2	60,2	58,4
2	1607	83,6	62,5	59,7
3	2289	119,0	64,5	61,9
4	2881	149,8	67,0	62,6
5	3314	172,3	69,6	63,3
6	3479	180,9	70,8	63,6
7	3526	183,4	71,5	64,3
8	3556	184,9	71,9	65,1
9	3584	186,4	71,9	65,1

Tableau 2.15. Taux de reconnaissance sur l'ensemble de validation EA_{195} et sur le noyau de test ET_{192} au cours de la ré-estimation des 404 modèles d'allophones droits.

L'évolution des taux en fonction du nombre de paramètres du système (cf. Figure 2.15) montre que les dernières estimations sont les plus profitables, puisque le nombre de gaussiennes augmente peu tandis que le taux de reconnaissance continue à progresser. De plus, l'amélioration est plus forte sur l'ensemble de validation, ce qui est vraisemblablement dû au fait que le choix du nombre de gaussiennes est optimisé sur l'ensemble d'apprentissage lui-même. Le système de référence utilise les meilleurs modèles obtenus pour la modélisation acoustique.

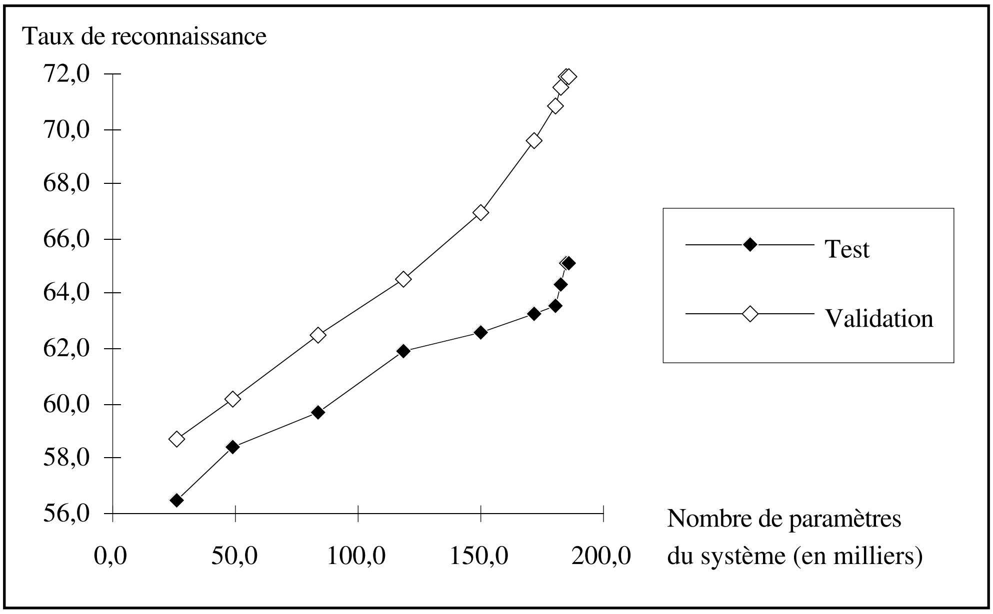

Figure 2.15. Taux de reconnaissance des modèles en contexte droit pour les ensembles de validation EA_{195} et de test ET_{192}.

Les 404 modèles contiennent 500 densités différentes qui sont représentées par un total de 3584 gaussiennes; chaque densité de probabilité utilise entre 1 et 32 gaussiennes, pour une moyenne de 7 gaussiennes. Le nombre total de gaussiennes pour tous les modèles d'une classe phonétique est donné dans le Tableau 2.16. Par exemple, la modélisation des 18 allophones de 'iy' nécessite un total de 162 gaussiennes réparties entre tous les états des modèles. Les phonèmes sont classés par ordre croissant du nombre de représentants dans l'ensemble d'apprentissage, car les deux séries de chiffres sont corrélées; cela est probablement dû au seuil imposé sur le nombre minimum de trames par gaussienne.

Modèle	zh	oy	uh	en	aw	th	ch	epi	el	y	jh	g
N_A	149	304	500	630	728	745	820	908	951	995	1013	1191
N_G	6	23	14	23	56	23	27	15	46	37	31	18
Modèle	ng	sh	ow	hh	dx	ao	ay	uw	v	b	f	w
N_A	1220	1317	1653	1660	1864	1865	1934	1952	1994	2181	2215	2216
N_G	37	55	91	51	21	101	119	80	49	31	60	72
Modèle	aa	ah	ey	ae	dh	d	p	eh	m	z	k	ax
N_A	2256	2266	2271	2292	2376	2432	2588	3277	3566	3682	3794	3892
N_G	109	79	107	116	40	32	43	106	86	118	85	63
Modèle	t	er	ih	l	iy	r	s	n	vcl	ix	sil	cl
N_A	3948	4138	4248	4425	4626	4681	6176	6896	7219	7370	8283	12518
N_G	73	158	112	136	162	119	169	119	111	94	113	148

Tableau 2.16. Nombre de représentants dans l'ensemble d'apprentissage (N_A) et nombre final de gaussiennes par classe phonétique (N_G) après l'apprentissage des 404 modèles.

2.6.2.5. Choix du facteur d'échelle

Il est possible avec HTK d'appliquer un coefficient s appelé "facteur d'échelle grammaticale" à la probabilité de transition inter-modèles qui est remplacée par la quantité suivante:

$$\tilde{P}(\lambda_j|\lambda_i) = \left[\frac{f(\varphi_i,\varphi_j)}{f(\varphi_i)}\right]^s \tag{2.14}$$

Ce facteur modifie l'influence du modèle de langage, qui est dans notre cas un simple bigramme phonétique, par rapport au modèle acoustique dans l'équation (1.4). En particulier, une valeur de s supérieure à 1 a pour effet de diminuer globalement les probabilités de transition inter-modèles, en pénalisant plus lourdement les transitions peu fréquentes. Des expériences montrent l'influence de ce facteur pour le contrôle des insertions [Ljolje, 1994a; Young & Woodland, 1994] mais la valeur optimale semble dépendre fortement des conditions expérimentales. Nous avons fait varier ce facteur pour des valeurs comprises entre 1 et 8, lors d'expériences de validation avec les modèles d'allophones droits retenus (cf. Tableau 2.17). L'augmentation de ce facteur diminue les insertions mais augmente les omissions. Le taux global de reconnaissance atteint un maximum pour une valeur de s comprise entre 3 et 5 (cf. Figure 2.16); nous privilégions le taux d'identification le plus élevé et donc une valeur faible de s (s=3).

Facteur d'échelle	Taux d'identification	Taux de substitution	Taux d'omission	Taux d'insertion	Taux de reconnaissance
1	79,9	15,2	5,0	8,0	71,9
2	79,9	14,5	5,6	5,3	74,6
3	79,4	14,2	6,5	3,9	75,5
4	78,9	13,8	7,3	3,2	75,6
5	78,1	13,8	8,0	2,8	75,4
6	76,8	14,2	9,0	2,7	74,1
8	72,5	16,2	11,3	2,3	70,2

Tableau 2.17. Application d'un facteur d'échelle au bigramme; taux de reconnaissance sur l'ensemble de validation EA_{195}.

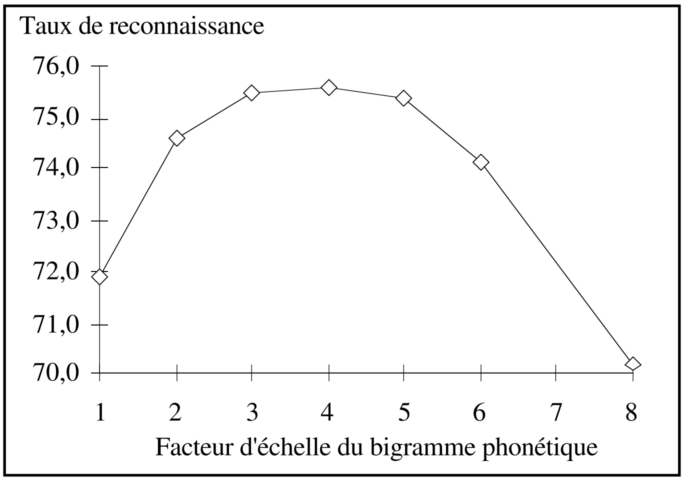

Figure 2.16. Influence du facteur d'échelle sur le taux de reconnaissance de l'ensemble de validation EA_{195}.

2.6.3. Performances du système de référence

Le système de référence est finalement constitué de 404 modèles phonétiques dépendants du contexte phonétique droit. Chaque modèle est une machine de Bakis à trois états. Les états initiaux et centraux des allophones d'un même phonème sont liés, et les 500 densités de probabilité différentes associées aux états (i.e., 48 densités pour les états initiaux, 48 pour les états centraux et 404 pour les états finaux) sont représentées par un nombre variable de gaussiennes à matrice de covariance diagonale en dimension 26, déterminé par une procédure automatique originale; ces lois utilisent un total de 3584 gaussiennes, soit en moyenne 7 gaussiennes par densité. Le système de reconnaissance contient finalement près de 190.000 paramètres. Enfin, le modèle de langage est un bigramme phonétique dont les probabilités sont élevées au cube. Ce système a été testé en décodage acoustico-phonétique sur les 168 locuteurs de l'ensemble de test complet de la base TIMIT, en comptabilisant les erreurs sur 39 classes phonétiques. L'ensemble de test complet comprend 50754 segments phonétiques; les taux de reconnaissance sont estimés avec un intervalle de confiance de l'ordre de ±0,4%. Les performances du système de référence sont comparées au système de base et aux systèmes existants.

2.6.3.1. Comparaison avec le système de base

Les taux de décodage du système de base et du système de référence sur l'ensemble de test complet sont présentés dans le tableau 2.18. Les taux du système de base sont légèrement supérieurs à ceux obtenus sur le noyau de test; cela est probablement dû au fait que l'ensemble de test complet contient des phrases ayant déjà été lues par d'autres locuteurs au cours de l'apprentissage. Par rapport au système de base, la prise en compte du contexte acoustique dans le système de référence et l'application d'un facteur d'échelle grammaticale mieux adapté permet une augmentation globale du taux de reconnaissance d'environ 7%. Cette amélioration est apportée par une réduction comparable des omissions et des substitutions.

Nombre de gaussiennes G	Paramètres du système (milliers)	Phonèmes identifiés (%)	Phonèmes omis (%)	Phonèmes substitués (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
Système de base (48 modèles)						
1	7,5	59,5	13,0	27,5	6,2	**53,3**
2	15	62,9	11,5	25,5	6,3	**56,7**
4	30	65,7	10,3	24,0	6,2	**59,5**
8	60	67,7	10,1	22,2	5,5	**62,2**
Système de référence (404 modèles)						
variable	186	74,8	6,6	18,7	5,7	**69,1**

Tableau 2.18. Taux de décodage par le système de base et le système de référence sur l'ensemble de test complet de la base TIMIT.

Cependant, la globalité des chiffres masque des performances très disparates en fonction des classes phonétiques. Le tableau 2.19 détaille les résultats du décodage pour six macro-classes phonétiques, à savoir les voyelles ou les diphtongues, les liquides ou les semi-voyelles, les nasales, les occlusives, les fricatives ou les affriquées, et enfin les silences ou les occlusions.

Classe phonétique	Nombre de segments	Phonèmes identifiés (%)	Phonèmes substitués (%)	Phonèmes omis (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
Voyelles	16580	65,3	27,9	6,8	5,9	**59,4**
Liquides	5890	73,7	16,6	9,7	7,4	**66,3**
Nasales	4434	76,0	17,5	6,4	4,6	**71,4**
Occlusives	6341	72,8	19,1	8,0	6,2	**66,6**
Fricatives	7271	73,1	22,4	4,5	3,5	**69,6**
Silences	10238	92,7	2,4	4,9	6,0	**86,7**

Tableau 2.19. Détail par macro-classes phonétiques des taux de décodage de l'ensemble de test complet par le système de référence.

Le tableau 2.20 précise les taux de substitutions intra-classes et inter-classes; pour chaque phonème substitué, la ligne représente la classe de l'étiquette phonétique attendue, et la colonne représente la classe de l'étiquette phonétique obtenue par le décodage. Les confusions sont rapportées au nombre total d'étiquettes d'une classe; la somme des taux sur une ligne est donc égale au pourcentage de substitutions du tableau 2.19.

	Voyelles	Liquides	Nasales	Occlusives	Fricatives	Silences
Voyelles	**22,9**	2,6	0,9	0,5	0,4	0,6
Liquides	8,5	**2,9**	1,1	1,9	1,9	0,3
Nasales	3,2	1,0	**9,7**	1,6	0,9	1,1
Occlusives	1,7	1,0	0,7	**11,5**	3,8	0,4
Fricatives	1,4	0,7	0,8	4,5	**13,8**	1,1
Silences	1,0	0,2	0,4	0,2	0,5	**0,0**

Tableau 2.20. Taux de substitutions intra-classes (en caractères gras) et inter-classes (en ligne: classe du phonème attendu; en colonne: classe du phonème décodé).

Le tableau 2.21 répertorie les erreurs les plus fréquentes en insertion, omission et substitution. Les substitutions entre deux phonèmes cumulent les confusions dans les deux sens possibles. Pour chaque catégorie d'erreur, le pourcentage d'erreur est rapporté au nombre total d'erreurs de décodage. Au total, les erreurs signalées dans le tableau contribuent pour plus de 50% à toutes les erreurs de décodage; il s'agit donc des plus significatives. Les insertions et omissions les plus fréquentes concernent des silences, et plus précisément des occlusions sourdes ou voisées. Enfin, le tableau 2.22 donne le détail pour chaque classe phonétique des différents taux de décodage.

Les voyelles ont les taux de reconnaissance les plus faibles, le plus souvent en raison de confusions avec d'autres voyelles: plus de 80% des cas de substitutions de voyelles sont des substitutions intra-classes; cela justifie la recherche d'un traitement mieux adapté aux voyelles proposé au §2.5.5.2. Les diphtongues sont particulièrement concernées par ce phénomène, et sont reconnues dans 20% des cas comme une voyelle non diphtonguée. Les confusions entre liquides et voyelles sont principalement dues à la confusion 'r' et 'er'. Parmi les consonnes, les fricatives sourdes sont les mieux identifiées. Les explosions des occlusives sourdes sont aussi mieux détectées que celles des occlusives voisées. Parmi toutes les macro-classes phonétiques, les nasales sont les mieux identifiées. Enfin, les silences et les occlusions sont très bien reconnus, ce qui biaise positivement le taux global de reconnaissance, car ces étiquettes représentent 20 % du total des segments de la base TIMIT. Nous avons comptabilisé ces segments dans le calcul des taux pour que nos résultats soient comparables avec ceux des systèmes existants. Lorsque les silences et les occlusions ne sont pas comptabilisés, le taux de reconnaissance global passe de 69,1% à 64,9%, soit une réduction de 4,2% en valeur absolue du taux de reconnaissance.

Insertions	Pourcentage des erreurs
sil	3,9
ih	1,5
ah	1,1
aa	0,9
l	0,8
n	0,8
r	0,7

Omissions	Pourcentage des erreurs
sil	3,2
ih	1,8
ah	1,6
r	1,3
n	1,2
l	0,9
eh	0,8
t	0,8
dx	0,7
d	0,7
dh	0,6

Substitutions	Pourcentage des erreurs	Substitutions	Pourcentage des erreurs
z → s	2,5	aa → ah	0,6
ih → ah	1,8	ih → er	0,5
ah → ih	1,6	ah → eh	0,5
r → er	1,3	b → p	0,5
ih → eh	1,2	eh → ah	0,5
ih → iy	1,2	er → ih	0,5
eh → ih	1,1	ih → ey	0,5
er → r	1,0	uw → ih	0,5
m → n	1,0	l → w	0,5
eh → ae	1,0	l → ow	0,5
n → m	0,9	ey → ih	0,4
ah → aa	0,8	ah → ow	0,4
iy → ih	0,8	ng → n	0,4
d → t	0,8	ow → ah	0,4
ae → eh	0,8	n → ng	0,4
s → z	0,7	ey → iy	0,4
g → k	0,6	v → f	0,4

Tableau 2.21. Insertions, omissions, et substitutions phonétiques contribuant pour plus de 50% au total des erreurs de décodage.

Étiquette	Nombre de segments	Phonèmes identifiés (%)	Phonèmes substitués (%)	Phonèmes omis (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
Voyelles et diphtongues						
iy	1810	82,4	13,7	3,9	4,5	77,8
ih	3939	66,6	26,3	7,1	5,8	60,8
eh	1247	48,1	41,9	9,9	7,6	40,5
ae	772	61,1	33,4	5,4	6,7	54,4
ah	2295	54,1	35,2	10,8	7,4	46,7
uw	572	53,5	37,4	9,1	2,3	51,2
uh	215	11,2	73,5	15,3	3,3	7,9
aa	1607	77,7	17,4	5,0	8,5	69,1
er	1692	75,4	20,3	4,3	2,8	72,6
ey	802	72,4	24,6	3,0	6,5	66,0
ay	686	75,4	21,3	3,4	6,3	69,1
oy	127	49,6	46,5	3,9	2,4	47,2
aw	216	38,9	55,6	5,6	4,6	34,3
ow	600	49,5	39,8	10,7	6,5	43,0
Liquides, semi-voyelles, 'h'						
l	2201	77,8	15,6	6,6	5,9	71,9
r	1849	70,6	18,7	10,7	5,6	65,1
y	376	57,2	25,3	17,6	11,4	45,7
w	903	79,2	12,7	8,1	8,1	71,1
hh	561	69,7	14,1	16,2	15,7	54,0
Nasales						
m	1406	77	17,7	5,3	4,3	72,6
n	2650	76,8	16,2	6,9	4,7	72,2
ng	378	66,9	26,2	6,9	5,0	61,9
Occlusives						
b	886	73,6	21,6	4,9	5,4	68,2
d	841	53,0	33,8	13,2	9,8	43,3
dx	634	58,8	23,2	18,0	5,2	53,6
g	452	50,2	42,5	7,3	2,4	47,8
p	957	79,9	15,6	4,5	6,1	73,9
t	1367	78,6	12,6	8,9	6,0	72,6
k	1204	89,9	6,5	3,7	6,6	83,2
Fricatives et affriquées						
z	1236	60,8	36,2	3,0	1,5	59,2
v	710	56,3	32,5	11,1	6,2	50,1
f	911	88,0	9,3	2,6	3,5	84,5
th	259	40,5	50,6	8,9	8,5	32,0
s	2172	89,5	8,7	1,8	2,4	87,1
sh	533	80,1	18,6	1,3	2,4	77,7
dh	896	58,0	30,7	11,3	5,1	52,9
ch	259	62,5	33,6	3,9	5,4	57,1
jh	295	69,5	28,5	2,0	3,7	65,8
Silences et occlusions						
sil	10238	92,7	2,4	4,9	6,0	86,7

Tableau 2.22. Taux de décodage du système de référence par classe phonétique.

2.6.3.2. *Comparaison avec d'autres systèmes*

K.F. Lee et H.W. Hon sont les premiers à avoir présenté des taux de reconnaissance phonétique sur une version prototype de la base TIMIT [Lee & Hon, 1989]. Les chercheurs de CMU utilisent des HMM à émissions discrètes, un bigramme phonétique, et les erreurs sont calculées sur la base des 39 classes phonétiques que nous avons présentées. Avec 48 modèles indépendants du contexte, le taux de reconnaissance est de 53,3%; il monte à 66,1% pour 1450 modèles dépendants du contexte droit. Les probabilités d'émission d'un modèle sont décrites par 2304 paramètres. Les conditions expérimentales sont légèrement différentes, puisque l'apprentissage est réalisé avec seulement 357 locuteurs et les tests sur 20 locuteurs.

L. Lamel et J.-L. Gauvain ont réalisé des expériences de DAP sur la base TIMIT [Lamel & Gauvain, 1993]. Les densités de probabilité associées aux 3 états des HMM sont décrites par une combinaison linéaire de gaussiennes à matrice de covariance diagonale. Avec 500 modèles phonétiques dépendants du contexte, un bigramme phonétique et des coefficients différentiels du premier ordre, ils obtiennent 69% de reconnaissance sur le noyau de test avec un système de 2,4 millions de paramètres. En rajoutant les coefficients différentiels du deuxième ordre et en utilisant un trigramme phonétique, ils atteignent 73,4% de reconnaissance sur l'ensemble de test complet, avec un système de 3,6 millions de paramètres.

S. Young, concepteur de HTK, a testé son système avec des expériences de DAP sur la base TIMIT [Young, 1992b]. Il construit des quasi-triphones par fusion de modèles contextuels gauches et droits. Les performances obtenues avec les quasi-triphones sont de 61,7% avec 4 gaussiennes par densité de probabilité; cependant des modèles indépendants du contexte avec 10 gaussiennes par densité permettent d'obtenir un taux de reconnaissance légèrement supérieur, de 62,8%. En modélisant 1748 modèles dépendants du contexte droit, en utilisant des coefficients différentiels du deuxième ordre et en testant les meilleures valeurs de seuils pour le regroupement des contextes, S. Young et P. Woodland obtiennent 72,3% de reconnaissance [Young & Woodland, 1994]. Le système contient 1142 états différents, et le maximum de reconnaissance est atteint pour 9 gaussiennes par densité de probabilité, soit approximativement 800.000 paramètres.

A. Ljolje d'AT&T a construit un système de reconnaissance avec des quasi-triphones construits par interpolation de modèles indépendants du contexte et de modèles dépendants d'un seul ou des deux contextes gauche et droit [Ljolje, 1994a]. Les densités de probabilité associées aux états sont représentées par une mono-gaussienne symétrique dans un espace de dimension 26, et les 3790 distributions du système nécessitent donc 1,4 millions de paramètres. Un modèle de durée est associé à chaque machine. Le taux de reconnaissance sur la base TIMIT est de 69,4% pour des modèles légèrement différents de ceux proposés par K.F. Lee. En particulier, les taux ne tiennent pas compte des occlusions et des silences. Il semblerait que ce mode de calcul, plus réaliste, diminue d'environ 5% le taux global de reconnaissance par rapport au décompte des autres systèmes.

T. Robinson de l'Université de Cambridge intègre un réseau de neurones récurrent à un système markovien et obtient un taux de reconnaissance remarquable de 75% sur l'ensemble de test complet [Robinson, 1994]. Ces résultats montrent l'efficacité d'une représentation non paramétrique des densités de probabilité: le réseau utilisé contient 47.400 poids, soit un nombre de paramètres beaucoup plus faible que pour des HMM de performance proche. A. Mellouk et P. Gallinari présentent des résultats légèrement meilleurs, avec une architecture connexioniste utilisant l'alignement temporel dynamique et des méthodes d'apprentissage discriminantes; ils obtiennent 75,6% de reconnaissance sur les 20 locuteurs de test de K.F. Lee [Mellouk & Gallinari, 1994].

Le tableau 2.23 récapitule les taux de reconnaissance de tous ces systèmes pour les comparer avec le système de base et le système de référence que nous avons développé. Nous voyons que les résultats obtenus sont honorables, surtout si l'on considère le nombre de paramètres du système de reconnaissance; seuls les systèmes hybrides markoviens et neuronaux permettent une modélisation acoustique plus performante et plus compacte.

Référence	Modèles	Nombre de paramètres du système	Nombre de locuteurs de test	Taux de phonèmes identifiés	Taux de phonèmes reconnus
		Système de base et de référence			
Système de base	48 phonèmes	60 k	24	66,5	61,2
Système de référence	404 allophones droits	186 k	168	74,8	69,1
		HMM discrets			
Lee *et al.*, 1989	48 phonèmes	110 k	20	64,1	53,3
"	1450 allophones droits	3,3 M	20	73,8	66,1
		HMM continus			
Young, 1992	48 phonèmes	75 k	160 phrases	71,9	62,8
"	807 quasi-triphones	212 k	160 phrases	72,5	61,7
Young *et al.*, 1994	1176 allophones droits	800 k	300 phrases	76,7	72,3
Lamel *et al.*, 1993	500 allophones	2,4 M	24	?	69,1
"	"	3,6 M	168	78,3	73,4
Ljolje, 1994	3382 quasi-triphones	1,4 M	40	74,8+5 ?	69,4+5 ?
		HMM & Réseaux de neurones			
Robinson, 1994	61 phonèmes	47 k	168	78,6	75,0
Mellouk *et al.*, 1994	39 phonèmes	120 k?	20	80,0	75,6

Tableau 2.23. Comparaison des performances des systèmes développés dans cette étude avec d'autres systèmes de reconnaissance testés sur la base TIMIT.

2.7. Conclusions

Nous avons construit un système de référence pour la reconnaissance de la parole continue par modèles de Markov cachés, utilisant des modèles phonétiques dépendants du contexte (cf. Tableau 2.24). L'étude des densités de probabilité associées aux états nous a conduit à proposer une procédure automatique originale basée sur la statistique du χ^2 pour la détermination du nombre optimal de gaussiennes par densité de probabilité, dans le cadre d'une modélisation par multi-gaussiennes; cette procédure conduit à une réduction du nombre de paramètres du système. Un certain nombre de choix ont été faits sur un système de base et n'ont pas été validés sur le système final, comme la représentation du signal par des coefficients cepstraux statiques et dérivés, ou le nombre d'états des modèles. Des axes de recherche ont été proposés sans être complètement explorés, comme par exemple la réductions du nombre de coefficients acoustiques différentiels, la définition d'une topologie des modèles adaptée aux différentes classes phonétiques, et l'amélioration de la représentation des densités de probabilité pour une meilleure discrimination entre voyelles.

Le système de référence proposé n'est donc pas le meilleur système de reconnaissance possible, mais il prend en compte l'expertise actuelle en matière de reconnaissance de la parole et présente un niveau de performance tout a fait comparable à l'état de l'art, pour un nombre de paramètres inférieur à celui de systèmes plus performants à base de HMM et une relative simplicité. Son développement a été réalisé avec la plate-forme logicielle HTK de l'Université de Cambridge, sur la base de données anglo-américaine multi-locuteurs TIMIT. Cela permet au système de référence d'être facilement reproductible, et un script en langage de commande C-shell réalisant l'apprentissage des modèles décrits dans ce chapitre est proposé en annexe A. Ce système de référence va nous permettre d'évaluer l'apport des traitements que nous proposons dans les chapitres suivants pour l'amélioration des modèles de Markov cachés, concernant les problèmes de leur adaptation au locuteur et de leur contrôle temporel.

Représentation du signal	26 coefficients (12MFCC, E, 12 ΔMFCC, ΔE)
Nombre de modèles	404 allophones droits généralisés
Topologie des modèles	modèle de Bakis à 3 états
Probabilité d'émission	multi-gaussienne multi-variable à matrice de covariance diagonale
Nombre de gaussiennes par état	variable (procédure automatique du χ^2)
Complexité du système	186.000 paramètres
Modèle de langage	bigramme phonétique (facteur d'échelle s=3)
Taux de DAP	74,8 % d'identification / 69,1 % de reconnaissance

Tableau 2.24. Récapitulatif des caractéristiques du système de référence.

3. Adaptation au locuteur

En raison de la grande variabilité inter-locuteurs, un système de reconnaissance de la parole dépendant du locuteur, c'est-à-dire conçu pour reconnaître la voix d'un ou de plusieurs locuteurs particuliers, ne peut être utilisé par un nouveau locuteur qu'après un apprentissage long. R. Schwartz et F. Kubala obtiennent 1,7% d'erreurs lexicales sur la tâche *Resource Management* (cf. §2.5.5.1) lorsque le système BYBLOS est entraîné en mode mono-locuteur sur une demi-heure de parole; ce taux passe à 10% lors de l'utilisation du système par un autre locuteur du même sexe, et à 77% pour un locuteur du sexe opposé [Schwartz & Kubala, 1992]. Ces limitations rendent souhaitable la conception de systèmes indépendants du locuteur. Les bases de données récentes contiennent les enregistrements de plusieurs centaines de locuteurs, ce qui permet d'illustrer de manière représentative la variabilité inter-locuteurs; les 630 locuteurs de la base de données TIMIT sont ainsi répartis en huit grands "dialectes" américains. Avec l'augmentation de la puissance des ordinateurs permettant de traiter ces données, les systèmes indépendants du locuteur atteignent des performances acceptables pour un nouveau locuteur. Ces performances restent cependant inférieures à celles de systèmes mono-locuteurs; avec un apprentissage sur 210 minutes de parole provenant de 109 locuteurs différents, le système BYBLOS déjà cité atteint 6% d'erreurs en mode indépendant du locuteur, au lieu de 1,7% en mode mono-locuteur. Il semble pour l'instant impossible d'estimer des modèles phonétiques universels. Avec un très grand nombre de locuteurs, les différentes réalisations d'un phonème sont dispersées dans l'espace acoustique, et se superposent avec les réalisations d'autres phonèmes, au détriment des propriétés discriminantes des modèles appris. Une adaptation du système au locuteur ou une normalisation des paramètres est nécessaire.

De nombreuses approches existent pour l'adaptation des systèmes de RAP; elles consistent par exemple en une transformation des modèles, ou en une transformation des paramètres acoustiques, soit par une traduction du dictionnaire de quantification vectorielle, soit par des normalisations spectrales. Nous nous intéressons particulièrement à une adaptation rapide basée sur une classification des locuteurs d'apprentissage. Nous étudions une distance inter-locuteurs calculée à partir de modèles auto-régressifs vectoriels, dont les performances en identification du locuteur sont remarquables [Montacié *et al.*, 1992a]; nous cherchons à déterminer si cette distance peut mettre en évidence des corrélations acoustiques entre les locuteurs qui soient pertinentes pour le décodage acoustico-phonétique. Enfin, nous proposons une généralisation de la reconnaissance par multi-modèles, qui permet de réaliser le décodage d'un locuteur de test avec les modèles de plusieurs classes de locuteurs et non plus avec ceux d'une seule classe. Il est possible par cette méthode de réaliser une adaptation rapide, supervisée ou non supervisée.

3.1. Techniques d'adaptation

L'adaptation au locuteur d'un système de RAP consiste à modifier ce système à partir d'un nombre réduit de phrases prononcées par un nouveau locuteur, afin d'améliorer les performances de reconnaissance pour ce locuteur. La normalisation, qui cherche aussi à traiter la variabilité inter-locuteurs, change l'espace de représentation des paramètres acoustiques pour un locuteur donné, de manière à disposer d'une représentation autant que possible invariante aux changements de locuteurs. Ces deux approches peuvent être complémentaires, bien qu'une normalisation parfaite rende théoriquement inutile la phase d'adaptation. En pratique, certaines méthodes d'adaptation utilisent des techniques de normalisation.

Nous allons présenter les méthodes existantes qui permettent de réaliser l'adaptation, qu'il s'agisse d'une transformation des paramètres acoustiques, d'une traduction du dictionnaire de quantification vectorielle, d'une modification des modèles du système, ou encore d'une classification des locuteurs d'apprentissage. Ces différentes méthodes d'adaptation peuvent être qualifiées de:

- supervisées ou non supervisées, suivant que le nouveau locuteur doit ou non prononcer des phrases particulières durant la phase d'adaptation;
- incrémentales, si l'adaptation se poursuit de manière continue durant l'utilisation du système par un même locuteur;
- rapides, si quelques mots ou quelques phrases sont suffisants pour l'adaptation.

L'approche qui nous semble souhaitable résulte d'un compromis entre ces différents modes: une phase d'adaptation rapide, éventuellement supervisée, est suivie d'une adaptation incrémentale non supervisée (cf. Figure 3.1). De cette manière, le système est presque immédiatement fonctionnel, et acquiert au bout d'une durée d'élocution suffisante les performances d'un système dépendant du locuteur. Nos expériences sont réalisées sur la base TIMIT, qui contient les enregistrements de dix phrases par locuteur; cela représente une quantité de données trop faible pour réaliser une adaptation incrémentale. Nous nous limitons donc principalement pour l'étude expérimentale de ce chapitre à la phase initiale d'adaptation rapide.

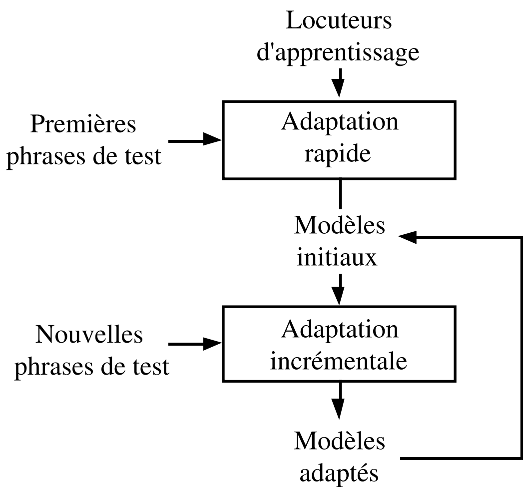

Figure 3.1. Procédure d'adaptation des modèles acoustiques.

3.1.2. Transformation des paramètres

Une méthode d'adaptation consiste à transformer les paramètres du locuteur de test de manière à les rendre plus proches des paramètres du ou des locuteurs d'apprentissage. Dans le cas d'un système dépendant du locuteur, la transformation peut être comparée à une conversion de voix du locuteur de test vers le locuteur d'apprentissage. La transformation inverse, du locuteur d'apprentissage vers le locuteur de test, est aussi possible. L'intérêt de cette deuxième solution est de réaliser tous les calculs durant la phase d'adaptation et de ne pas transformer les paramètres du locuteur de test pendant la phase de décodage. Lorsque les trames acoustiques sont quantifiées, la traduction du dictionnaire de QV permet une transformation non paramétrique de l'espace de représentation. La transformation peut aussi être réalisée par des fonctions linéaires ou non linéaires, en tenant éventuellement compte d'un modèle de production.

3.1.2.1. Traduction du dictionnaire

La quantification vectorielle a été beaucoup utilisée dans les systèmes réalisant un alignement temporel dynamique avec des formes de référence, ainsi que dans les systèmes markoviens discrets. La référence d'un mot est par exemple décrite comme une succession de prototypes du dictionnaire. Un mode d'adaptation consiste à déplacer les prototypes du dictionnaire de QV pour les rendre représentatifs du locuteur de test, sans modifier les indices des prototypes utilisés dans les références de mots ou les modèles.

L'adaptation d'un système de reconnaissance mono-locuteur par DTW est décrite dans [Shikano *et al.*, 1986; Bonneau, 1987]. Il est nécessaire de disposer de la réalisation d'un même groupe de mots par le locuteur d'apprentissage et le locuteur de test; l'adaptation est donc supervisée. Un alignement dynamique entre les réalisations des deux locuteurs permet de mettre les trames en correspondance. Une table de substitution remplace chaque prototype du locuteur de test par le prototype d'apprentissage qui lui est le plus souvent associé lors des alignements [Shikano *et al.*, 1986]. H. Bonneau propose d'utiliser une table de distances [Bonneau, 1987], qui permet une substitution dynamique des prototypes lors de l'alignement temporel dynamique entre les trames de référence et les trames de test. Soit $V=\{v_i\}$ le dictionnaire de QV du locuteur de référence, $V'=\{v'_i\}$ celui du locuteur testé, et $c(v_j,v'_k)$ le nombre de fois où les trames v_j et v'_k sont mises en correspondance pendant la phase d'adaptation. Lors de l'alignement par DTW d'un mot inconnu avec les références pour l'identification de ce mot, on ne souhaite pas calculer directement une distance entre un spectre du locuteur de test et un spectre du locuteur de référence. La distance inter-trames utilisée est alors la distance à la trame de test la plus proche parmi celles qui ont été mises en correspondance avec la trame de référence considérée pendant la phase d'adaptation:

$$d(v_i,v'_k)=\min_j d(v'_j,v'_k) \text{ avec } j \text{ tel que } c(v_i,v'_j)>0 \qquad (3.1)$$

Sur 20 locuteurs prononçant un vocabulaire d'adaptation de 136 mots et un vocabulaire de test de 104 mots, le gain en reconnaissance est très important puisque l'adaptation fait baisser le taux d'erreurs de 21,4% à 10,4%, sans toutefois atteindre le niveau du système sur le locuteur de référence qui est de 2,7% d'erreurs.

Les systèmes décrits fonctionnent en mode supervisé, ce qui est assez contraignant pour l'utilisateur. En mode non-supervisé, S. Furui a montré l'efficacité d'une classification hiérarchique descendante des trames du locuteur de test pour adapter un dictionnaire de référence à ce nouveau locuteur et mettre en correspondance les deux espaces spectraux [Furui, 1989]. Initialement, tous les prototypes du dictionnaire de référence sont translatés pour faire coïncider leur moyenne μ_V avec celle des trames d'adaptation μ_A:

$$\tilde{v}_k = v_k + \mu_A - \mu_V \qquad (3.2)$$

Les trames sont ensuite séparées en deux classes par une procédure automatique. Les prototypes sont assignés à la classe la plus proche et la procédure initiale de translation est répétée à l'intérieur de chaque classe, c'est-à-dire que les prototypes attribués à une classe sont déplacés par translation vers le centroïde de cette classe. Cependant, R. Schwartz et F. Kubala notent que ce mode d'adaptation suppose que les composantes principales des deux espaces correspondent aux mêmes phonèmes, et qu'il ne fonctionne plus efficacement lorsque les coefficients différentiels sont utilisés [Schwartz & Kubala, 1992]. H. Su et R André-Obrecht réalisent une adaptation incrémentale du dictionnaire de quantification vectorielle, par un algorithme de gradient stochastique qui déplace les vecteurs de référence de manière continue, ou par l'algorithme de Lloyd qui calcule un nouveau dictionnaire lorsqu'une quantité suffisante de données d'adaptation a été mémorisée [Su & André-Obrecht, 1988]. Des expériences de DAP sur 21 phrases courtes montrent que l'algorithme de Lloyd réduit de moitié les erreurs de reconnaissance. Une telle adaptation est non supervisée, mais elle nécessite une bonne reconnaissance initiale, qui tienne compte du niveau linguistique.

3.1.2.2. Transformations spectrales linéaires

La traduction d'un dictionnaire nécessite un nombre élevé d'observations, afin de mettre tous les prototypes du dictionnaire en correspondance. En se plaçant dans l'espace continu des paramètres, on peut chercher une fonction d'adaptation plus générale réalisant la transformation spectrale du locuteur. Le cas de fonctions linéaires a été étudié par Y. Grenier [Grenier, 1980]. Une matrice de transformation T de dimension dxd permet le passage des spectres de l'espace acoustique d'un locuteur à un autre:

$$X_2 = T.X_1 \tag{3.3}$$

où X_1 et X_2 sont respectivement une trame du locuteur de référence L_1 et du locuteur de test L_2. Chacun des d coefficients acoustiques est donc exprimé comme une combinaison linéaire des coefficients du locuteur de référence. La matrice est estimée sur un ensemble de couples de spectres par une méthode de régression multiple. Y. Grenier propose une généralisation de ce principe, en projetant les images acoustiques de chaque locuteur dans un espace commun (cf. Figure 3.2). Les transformations sont toujours réalisées par des fonctions linéaires et sont déterminées par analyse des corrélations canoniques en minimisant la distance entre les projetés [Grenier, 1980; Choukri, 1987].

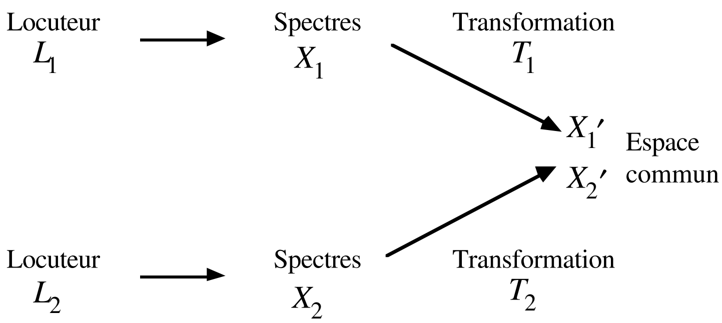

Figure 3.2. Méthodes de transformation spectrale.

En restant dans le cadre de transformations linéaires, les chercheurs d'IBM proposent une adaptation appelée transformation métamorphique, qui calcule une transformation linéaire pour chaque unité acoustique de base, en l'occurrence un fénone du système de reconnaissance d'IBM qui correspond à un état d'un ou plusieurs allophones [Bellegarda *et al.*, 1994]. Pour chaque fénone f, les trames de référence et d'adaptation sont mises en correspondance, puis normalisées de manière à avoir une moyenne nulle et l'identité comme matrice de covariance:

$$
\begin{aligned}
X_1' &= \Sigma_1{}^{-1/2}(X_1 - \mu_1) \\
X_2' &= \Sigma_2{}^{-1/2}(X_2 - \mu_2)
\end{aligned}
\tag{3.4}
$$

Une transformation orthonormale T_f est calculée pour le fénone f entre des couples représentatifs de vecteurs ainsi normalisés:

$$X_2' \approx T_f . X_1' \tag{3.5}$$

et les vecteurs d'apprentissage sont projetés dans l'espace de test:

$$
\begin{aligned}
X_2 &= \Sigma_2{}^{1/2} X_2' + \mu_2 \\
X_2 &= \Sigma_2{}^{1/2} . T_f . X_1' + \mu_2 \\
X_2 &= \Sigma_2{}^{1/2} . T_f . \Sigma_1{}^{-1/2}(X_1 - \mu_1) + \mu_2
\end{aligned}
\tag{3.6}
$$

Un nouveau dictionnaire de QV est construit avec les vecteurs obtenus. Les expériences d'adaptation avec 200 phrases par la méthode décrite montrent des résultats presque équivalents à un système mono-locuteur obtenu avec l'apprentissage de 2000 phrases; de plus, la conversion entre les voix d'hommes et de femmes ou l'utilisation de plusieurs locuteurs de référence ne semblent pas poser de problèmes.

3.1.2.3. Transformations spectrales non-linéaires

Les transformations linéaires sont probablement insuffisantes pour représenter la complexité des modifications spectrales entre locuteurs, ce qui a motivé l'étude de méthodes non linéaires. C. Montacié a expérimenté un filtrage quadratique, ainsi qu'un perceptron multi-couches (PMC) appris par rétro-propagation du gradient pour réaliser cette transformation [Montacié, 1991]. Dans le cas du filtrage polynomial quadratique, les coefficients acoustiques du locuteur de test sont calculés par une combinaison linéaire des coefficients du locuteur de référence ainsi que de leurs produits. Le vocabulaire d'adaptation est choisi sur un critère statistique de χ^2, permettant de sélectionner un sous-ensemble de mots ayant une distribution des prototypes représentative de l'ensemble du vocabulaire. Enfin, un alignement dynamique itératif permet une meilleure mise en correspondance des trames. Les résultats de cette étude montrent l'intérêt des non-linéarités du PMC.

D'autres expériences d'adaptation par des méthodes linéaires et non-linéaires ont été réalisées à l'ENST sur 20 mots anglais prononcés 26 fois par 14 locuteurs différents; le taux moyen d'erreurs diminue de près de 50% après adaptation [Tubach *et al.*, 1990]. Ces expériences confirment le léger avantage de l'utilisation d'un PMC par rapport à une régression linéaire multiple. F. Class, A. Kaltenmeier, P. Regel et K. Trottler comparent différentes méthodes d'adaptation des systèmes de RAP par DTW ou par HMM [Class *et al.*, 1990], en particulier des méthodes dérivées de celles proposées par Y.

Grenier; ils obtiennent la meilleure adaptation par une analyse en corrélations canoniques appliquée aux coefficients quadratiques.

3.1.2.4. Normalisation acoustique fondée sur la production

Des normalisations exploitant certaines caractéristiques des modèles de production ont aussi été proposées. De nombreuses techniques sont détaillées et discutées dans [Choukri, 1987], qui donne les références de travaux simplement cités ici: normalisation au moyen de diverses transformations affines du triangle vocalique, qui est l'espace de répartition des voyelles dans le plan des fréquences des deux premiers formants F_1 et F_2 (Auxerre, 1984; Chiba, 1986); normalisation relative à la longueur du conduit vocal par des transformations linéaires (Kasuya, 1976) ou non linéaires (Boe, 1980); ajustement des fréquences formantiques par des méthodes linéaires (Golibersuch, 1983) ou non linéaires (Wakita, 1979; Ainsworth, 1984; Paliwal, 1985; Blomberg, 1986; Chuang, 1986). Cette dernière approche est motivée par le fait que les variations inter-locuteurs entraînent des déplacements des pics spectraux qui dégradent la reconnaissance; Ainsworth et Paliwal réalisent l'ajustement des fréquences par alignement fréquentiel dynamique, qui est l'équivalent en fréquence de l'alignement temporel dynamique.

3.1.3. Transformation des modèles

La transformation des paramètres acoustiques du nouveau locuteur ou leur normalisation constitue un des modes d'adaptation; dans une approche markovienne, il est également possible de modifier les modèles, qu'il s'agisse de modèles discrets ou continus.

3.1.3.1. Adaptation de modèles discrets

Pour l'adaptation de modèles discrets, R. Schwartz, Y. Chow et F. Kubala estiment pour chaque classe de phonèmes ϕ une matrice de transformation linéaire T_ϕ qui permet d'exprimer les nouvelles probabilités d'émission d'un état $B'_j = \left[b'_j(k)\right]_{1 \le k \le n_V}$ à partir des probabilités existantes [Schwartz *et al.*, 1987]:

$$B'_j = T_\phi . B_j \tag{3.7}$$

On peut interpréter la transformation comme une traduction probabiliste du dictionnaire de QV spécifique à un groupe d'états; il existe donc des similitudes avec l'approche de K. Shikano. L'inconvénient de cette méthode est le grand nombre de paramètres à estimer, puisque la taille de chaque matrice est le carré du nombre de symboles du dictionnaire de QV, et qu'une matrice doit être calculée pour chaque phonème, ou au moins pour chaque classe phonétique. Dans des expériences en mode supervisé, les chercheurs de BBN adaptent un système mono-locuteur à un nouveau locuteur avec seulement deux minutes de parole; ils atteignent pour ce locuteur les performances obtenues avec un système indépendant du locuteur, dont l'apprentissage est beaucoup plus long que celui du système mono-locuteur initial [Schwartz & Kubala, 1992].

Pour adapter les modèles du système SPHINX, K.-F. Lee estime des modèles dépendants du locuteur sur quelques phrases d'adaptation et réalise un lissage des probabilités d'émission entre les modèles indépendants du locuteur et ces nouveaux modèles [Lee, 1988]. Les modèles finaux sont une combinaison linéaire de plusieurs modèles, et les pondérations affectées à chaque groupe de modèles sont apprises sur de nouvelles phrases d'adaptation par une technique de suppression et d'interpolation introduite par F. Jelinek. Un traitement supplémentaire, le lissage des paramètres par co-occurrence, permet d'améliorer la robustesse des modèles appris sur les quelques phrases d'adaptation.

3.1.3.2. Adaptation de modèles continus

Les méthodes précédentes ne sont pas applicables aux modèles de Markov continus. Pour la conversion de modèles d'un locuteur de référence à un nouveau locuteur, P. Kenny, M. Lennig et P. Mermelstein utilisent une transformation linéaire de la moyenne et de la matrice de covariance des lois de probabilité associées aux états; en particulier, la moyenne est augmentée d'un vecteur de translation estimé globalement pour tous les états d'un phonème [Kenny *et al.*, 1990]. La procédure d'apprentissage utilise une variante de l'algorithme de Baum-Welch, et les expériences montrent qu'un ré-apprentissage complet des moyennes de chaque état est préférable à l'apprentissage d'un biais global partagé par un groupe de plusieurs états.

C.-H. Lee ré-estime les modèles par *maximum a posteriori* (MAP) et exprime les nouvelles valeurs de la moyenne ou de la variance comme une somme pondérée des valeurs *a priori* et des valeurs estimées sur les nouvelles phrases [Lee *et al.*, 1991; Lee & Gauvain, 1993]. Ces études montrent que pour une durée identique d'apprentissage, l'adaptation de modèles indépendants du locuteur donne de meilleurs taux de reconnaissance qu'un apprentissage direct de nouveaux modèles dépendants du locuteur. Ce mode d'adaptation ne modifie que les modèles qui disposent de nouvelles observations; il est donc souhaitable de disposer de plusieurs prononciations des unités modélisées, ce qui peut nécessiter quelques dizaines de phrases dans le cas d'unités phonétiques.

3.1.3.3. Adaptation rapide

Une adaptation rapide des modèles est proposée par [Cox, 1995; Leggetter & Woodland, 1995; Zhao, 1994]. La représentation des corrélations entre deux états d'un même locuteur par une régression linéaire permet de prédire un son qui n'a pas été observé [Cox, 1995]. C. Leggetter et P. Woodland appliquent une matrice de régression ainsi qu'éventuellement un biais aux moyennes des densités de probabilité [Leggetter &

Woodland, 1995]; les états acoustiquement proches partagent la même fonction de transformation, et le nombre total de matrices dépend de la durée d'adaptation disponible. Les matrices de transformation sont estimées par maximum de vraisemblance, et une adaptation non supervisée d'une dizaine de secondes suffit à apporter une légère réduction des erreurs.

Enfin, Y. Zhao réalise une normalisation des paramètres acoustiques associée à une adaptation des modèles phonétiques par des modèles linéaires [Zhao, 1994]. Dans le domaine spectral logarithmique, les observations X_2^{ϕ} du locuteur de test L_2 pour la classe phonétique ϕ sont exprimées dans l'espace acoustique du locuteur standard L_1. Elles sont perturbées par un biais l_2^{ϕ} spécifique au locuteur et à la classe phonétique ainsi que par un biais acoustique h_2 global au locuteur considéré:

$$X_2^{\phi} = X_1^{\phi} + l_2^{\phi} + h_2 \tag{3.8}$$

Le biais acoustique est calculé comme la moyenne des différences de chaque trame acoustique avec la référence moyenne $\mu_1^{\phi(t)}$ correspondant à la classe phonétique de cette trame $\phi(t)$ pour le locuteur L_1:

$$h_2 = \frac{1}{T}\sum_t \left(X_2(t) - \mu_1^{\phi(t)}\right) \tag{3.9}$$

Ce biais converge vers la différence des spectres à long terme μ_1 et μ_2:

$$h_2 \rightarrow \mu_2 - \mu_1 \tag{3.10}$$

La formulation (3.9) donne une estimation du biais acoustique plus fiable que la différence des spectres à long terme lorsque la distribution des classes phonétiques n'est pas identique entre les phrases de référence et d'adaptation, ce qui est le cas pour une adaptation rapide. Le calcul du biais acoustique pour les locuteurs d'apprentissage permet d'apprendre ensuite des modèles avec des paramètres normalisés.

Le biais phonétique est la différence pour chaque classe des moyennes des trames d'adaptation et de référence, réduite du biais acoustique:

$$l_2^{\phi} = \mu_2^{\phi} - \mu_1^{\phi} - h_2 \tag{3.11}$$

Les moyennes et covariances des densités de probabilité des modèles sont recalculées par une estimation bayésienne, qui tient compte pour chaque paramètre de la quantité et de la variance des données disponibles. Cette procédure complète entraîne une réduction de 27% des erreurs de reconnaissance des mots sur la base TIMIT, en utilisant seulement 5 secondes de parole pour l'adaptation.

3.1.4. Multi-modèles

La modélisation par classes de locuteurs est fondée sur l'hypothèse qu'il existe des similitudes fortes entre certains locuteurs, et qu'il est possible d'apprendre des modèles phonétiques pour des groupes de locuteurs homogènes d'un point de vue acoustique. Pour un locuteur de test donné, il suffit alors de choisir les modèles correspondants au groupe dont il est le plus proche. On peut considérer que ce choix constitue une forme d'adaptation.

3.1.4.1. Adaptation rapide par multi-modèles

Pour tenir compte de la variabilité inter-locuteurs dans les systèmes de RAP indépendants du locuteur, une solution classique consiste à recueillir un grand nombre de réalisations d'une même unité acoustique et à réaliser une classification automatique de ces formes, ce qui permet de définir plusieurs types de prononciations. Si l'on dispose d'un système de reconnaissance par alignement temporel dynamique, ces prononciations sont représentées chacune par une référence, typiquement le centroïde de la classe, et un vote majoritaire parmi les références les plus proches de la forme de test permet de choisir le mot prononcé. Les multi-modèles constituent l'extension naturelle des références multiples aux systèmes de reconnaissance utilisant des HMM. Les locuteurs sont répartis en plusieurs classes, suivant un critère *a priori* ou par une classification automatique, et une version des modèles acoustiques est apprise pour chaque classe de locuteurs [Jouvet *et al.*, 1994]. Cela autorise une meilleure modélisation de la variabilité d'élocution des locuteurs qu'un modèle unique. Les multi-modèles permettent une adaptation rapide à un nouveau locuteur, en choisissant de manière implicite ou explicite la classe d'appartenance d'un nouveau locuteur (cf. Figure 3.3).

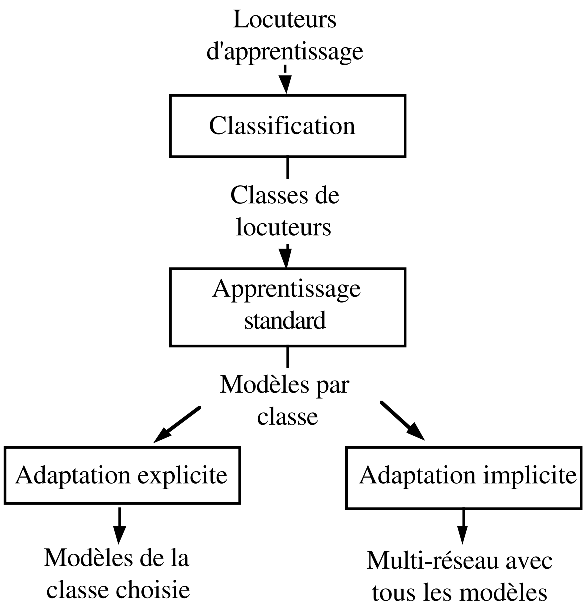

Figure 3.3. Adaptation rapide par multi-modèles.

Dans un premier mode de fonctionnement qualifié d'implicite, la structure du réseau phonétique ou syntaxique contraint le décodage à enchaîner des modèles appris sur une même classe; la classe d'appartenance du locuteur de test est donc choisie implicitement au cours de la reconnaissance. Plus précisément, la probabilité d'observation *a posteriori* de la suite de mots M que l'on cherche à maximiser en fonction des observations O peut se décomposer sur l'ensemble des classes de locuteurs L comme:

$$P(M/O) = \sum_{L} P(L, M/O) = \frac{P(M)}{P(O)} \sum_{L} P(O/L, M).P(L) \qquad (3.12)$$

puisque la suite de mots M et la classe de locuteur L sont *a priori* indépendantes. Mais le décodage est réalisé avec l'algorithme de Viterbi; on recherche en réalité le meilleur chemin parmi tous les chemins possibles, ce qui revient à faire l'approximation suivante en choisissant la classe L la plus probable et le meilleur chemin Q:

$$P(M/O) \approx \frac{P(M)}{P(O)} \max_{L,Q} P(O,Q/L,M).P(L) \qquad (3.13)$$

Dans le deuxième mode qualifié d'explicite, la classe du locuteur de test est choisie explicitement à la suite d'une phase d'adaptation supervisée ou non, et le décodage continue avec les seuls modèles de cette classe. Cette classe $\tilde{L}$ est choisie comme la classe la plus probable au vu des observations:

$$\tilde{L} = \arg\max_{L} P(L/O) = \arg\max_{L} P(O/L) \qquad (3.14)$$

en appliquant la règle de Bayes et en supposant toutes les classes de locuteurs équiprobables. Dans le cas particulier d'une adaptation supervisée où la suite de mots M_A correspondant aux phrases d'adaptation observées O_A est connue, le choix est plus précisément:

$$\tilde{L} = \arg\max_{L} P(O_A/L, M_A) \qquad (3.15)$$

Connaissant les observations acoustiques O et la classe de locuteurs $\tilde{L}$, la suite de mots $\tilde{M}$ maximisant la probabilité *a posteriori* est cherchée:

$$P(M/O,\tilde{L}) = \frac{P(O/\tilde{L},M).P(M)}{P(O/\tilde{L})} \qquad (3.16)$$

Cette formalisation généralise celle proposée par A. Imamura [Imamura, 1991].

Dans les deux cas, des modèles adaptés au locuteur de test sont choisis dès les premières phrases de test, sans pour autant nécessiter de ré-apprentissage des paramètres des modèles. Ce type d'adaptation, rapide et généralement non-supervisé, est intéressant pour traiter un nouveau locuteur; mais il n'exclut pas ensuite une adaptation incrémentale des modèles.

3.1.4.2. Classes de locuteurs pour les multi-modèles

Les modèles par genre correspondent au cas particulier où l'on choisit de ne représenter que deux classes de locuteurs en les répartissant *a priori* suivant leur sexe. Les hommes et les femmes présentent en effet des différences physiologiques importantes qui induisent des modifications des caractéristiques acoustiques. Deux modèles sont alors appris pour chaque unité phonétique, l'un sur les locuteurs masculins de l'ensemble d'apprentissage, et l'autre sur les locuteurs féminins. Lors du décodage, les deux groupes de modèles sont proposés en parallèle dans le réseau, qui impose l'enchaînement de modèles d'un même genre (cf. Figure 3.4). Le meilleur chemin emprunte l'une quelconque des deux branches du réseau, et la complexité du décodage est *a priori* doublée, mais en pratique certaines heuristiques permettent d'élaguer rapidement les chemins du genre le moins probable. Des expériences réalisées par D. Jouvet au CNET [Jouvet *et al.*, 1994] mettent en évidence la supériorité du bi-modèle sur un modèle unique, qu'il s'agisse de modèles par genre ou obtenus avec une classification automatique des locuteurs; le taux d'erreur sur les mots est réduit de 10 à 30% suivant les expériences. Des modèles régionaux ont aussi été proposés, pour tenir compte des différences de prononciation du néerlandais entre les Pays-Bas et la Flandre [Van Compernolle *et al.*, 1991], ou pour les grandes régions des États-Unis [Beattie *et al.*, 1995].

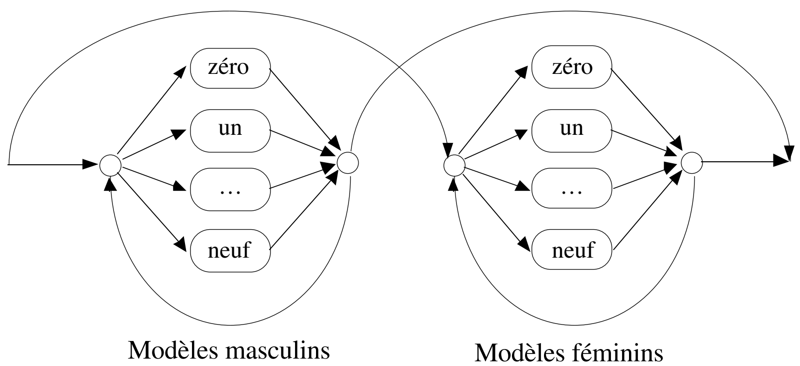

Figure 3.4. Exemple de réseau pour la reconnaissance de suites de chiffres avec des modèles par genre.

En dehors de ces cas particuliers où l'on dispose de connaissances préalables sur les groupes de locuteurs, une classification automatique des locuteurs est nécessaire. Le problème qui se pose est bien sûr le choix de critères pertinents pour la classification des locuteurs. Lorsque le nombre de classes augmente, la quantité de données d'apprentissage pour les modèles de chaque classe diminue. Un compromis est nécessaire, et conduit à choisir un nombre de classes assez faible. De telles approches ont été proposées pour des HMM discrets dans [Shikano *et al.*, 1986; Lee, 1988; Imamura, 1991]. Pour éviter un trop grand fractionnement des locuteurs, une classification hiérarchique des locuteurs est proposée dans [Mathan & Miclet, 1990; Kosaka & Sagayama, 1994; Sankar *et al.*, 1995]. De cette manière, il est possible de choisir pour chaque locuteur de test le niveau le plus adapté de la hiérarchie, allant des modèles indépendants du locuteur aux modèles les plus spécifiques. Dans les études de L. Mathan et de T. Kosaka, les meilleurs résultats en reconnaissance de mots isolés ou en reconnaissance phonétique sont obtenus pour deux niveaux hiérarchiques, soit quatre groupes de locuteurs au niveau le plus bas; ceci montre l'intérêt d'aller au-delà des bi-modèles.

La classification automatique des locuteurs nécessite la définition d'une distance inter-classes. Celle-ci peut être le taux de reconnaissance des phrases des locuteurs d'une classe par les modèles d'une autre classe si le système fonctionne en mode supervisé, ou sinon la probabilité d'émission de ces phrases [Mathan & Miclet, 1990]. Elle peut aussi provenir de la comparaison directe entre les modèles appris pour chaque classe, avec une distance basée sur la comparaison entre les lois de probabilité: perte d'entropie lors de la fusion des lois [Lee, 1988], distance de Chernoff ou de Bhattacharyya [Kosaka & Sagayama, 1994], mesure d'information de Kullback [Sankar *et al.*, 1995]. Dans le cadre des modèles connexionistes, les chercheurs de Cambridge [Cook & Robinson, 1995] proposent une classification hiérarchique descendante des phrases prononcées par les locuteurs, grâce à une mesure de sphéricité harmonique entre les trames supposées suivre une répartition gaussienne; cette classification ne leur permet cependant pas de dépasser les performances obtenues avec des modèles par genre.

3.2. Adaptation par multi-modèles

En raison de la grande variabilité inter-locuteurs, nous pensons que des modèles indépendants du locuteur appris sur la totalité de la base d'apprentissage ne constituent pas de bons modèles pour la reconnaissance; nous jugeons préférable de constituer des classes de locuteurs pour réaliser une adaptation rapide par multi-modèles. Pour définir une topologie des locuteurs et réaliser une classification automatique de ces locuteurs, nous avons choisi une distance inter-locuteurs très performante en reconnaissance du locuteur [Montacié *et al.*, 1992a]. Il nous a semblé intéressant de tester son potentiel pour une adaptation de systèmes de RAP prenant en compte la topologie de l'espace des locuteurs. Nous espérons obtenir une réduction de la variabilité inter-locuteurs à l'intérieur des classes d'apprentissage des multi-modèles. Les multi-modèles sont comparés avec des modèles standards et des modèles par genre, et les tests sont réalisés avec des modèles phonétiques indépendants du contexte sur l'ensemble de test complet.

3.2.1 Topologie des locuteurs

Nous avons choisi comme mesure de proximité entre locuteurs une distance d'Itakura symétrisée entre des modèles auto-régressifs vectoriels. Cette distance a donné de très bons résultats en identification et en vérification du locuteur [Montacié *et al.*, 1992a]. L'évaluation de la distance inter-locuteurs a été réalisée sur le système de base par des expériences préliminaires d'apprentissage sur une sélection des locuteurs d'apprentissage. Les résultats encourageants nous ont conduit à poursuivre l'étude de cette mesure et l'utiliser pour classifier les locuteurs d'apprentissage et tester l'adaptation rapide par multi-modèles.

3.2.1.1. Mesure inter-locuteurs

La mesure inter-locuteurs utilisée est calculée sur des modèles auto-régressifs vectoriels; cette représentation, plus complexe qu'une représentation auto-régressive classique du signal, est particulièrement efficace en identification et vérification du locuteur. Elle est à l'origine de la serrure vocale Orphée développée et expérimentée au LAFORIA [Montacié & Le Floch, 1992; Le Floch *et al.*, 1994].

Les modèles auto-régressifs vectoriels (ARV) permettent de représenter l'évolution d'un vecteur de paramètres. Les vecteurs de paramètres $\{X_n\}_{1 \le n \le N}$ de dimension d sont décrits par un modèle auto-régressif vectoriel d'ordre p :

$$X_n = \sum_{i=1}^{p} A_i X_{n-i} + E_n \tag{3.17}$$

où les $\{A_i\}_{1 \le i \le p}$ sont des matrices de dimension d x d, et E_n est un bruit blanc vectoriel de matrice de covariance D. Les coefficients des matrices A_i sont estimés par l'algorithme de Levinson-Whittle-Robinson, en minimisant la trace de la matrice de covariance D. Les modèles ARV ont été appliqués à des vecteurs de paramètres de parole (LPCC), principalement pour la reconnaissance du locuteur. L'ordre p des modèles a été fixé à 2 à la suite d'expériences préliminaires [Montacié *et al.*, 1992a].

Lors d'expériences d'identification du locuteur [Montacié & Le Floch, 1993], la mesure inter-locuteurs qui a donné les meilleurs résultats est une version symétrique de la mesure d'Itakura. Soient deux locuteurs L_1 et L_2, les N trames acoustiques $\{X_n\}_{1 \le n \le N}$ du locuteur L_1 sont représentées par le modèle ARV $\{A_i\}_{1 \le i \le p}$, et les M trames $\{Y_m\}_{1 \le m \le M}$ du locuteur L_2 sont représentés par le modèle $\{B_j\}_{1 \le j \le p}$. La matrice de covariance du résidu du filtrage inverse des vecteurs $\{X_n\}$ par le modèle $\{A_i\}$ est notée D_{XA}. Les matrices D_{XB}, D_{YA} et D_{YB} sont définies de manière analogue. La mesure d'Itakura vectorielle d_{IV} entre les deux locuteurs est alors:

$$d_{IV}(L_1, L_2) = \log\left(\frac{M.\mathrm{Trace}(D_{YA}.D_{XA}{}^{-1}) + N.\mathrm{Trace}(D_{XB}.D_{YB}{}^{-1})}{M + N} \right) \tag{3.18}$$

3.2.1.2. Expériences avec sélection de l'apprentissage

Avant de réaliser l'apprentissage de multi-modèles par classes de locuteurs, nous avons voulu évaluer la pertinence de la mesure par des expériences ne nécessitant pas de classification. Un modèle ARV est calculé pour chaque locuteur d'apprentissage et de test sur quelques phrases. Chaque locuteur de test L_T est comparé à l'ensemble des locuteurs d'apprentissage au moyen de la mesure d_{IV} entre les modèles ARV du locuteur de test et de chaque locuteur d'apprentissage. Nous sélectionnons ensuite un voisinage du locuteur de test dans l'ensemble des locuteurs d'apprentissage; ce voisinage est défini simplement comme un certain pourcentage des plus proches locuteurs de L_T. Si la base d'apprentissage contient un nombre très élevé de locuteurs et si elle est représentative de la variabilité inter-locuteurs, alors un voisinage convenablement choisi permet de se rapprocher des performances d'un système mono-locuteur. Les modèles phonétiques sont finalement appris sur ce sous-ensemble des locuteurs d'apprentissage (cf. Figure 3.5).

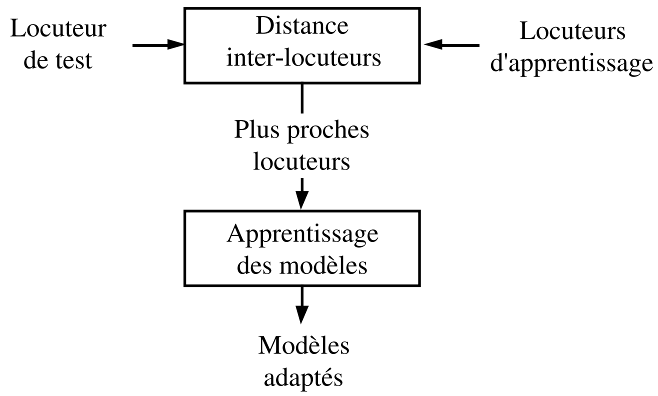

Figure 3.5. Adaptation par sélection de l'apprentissage.

Les modèles phonétiques sont adaptés au locuteur de test, dans la mesure où ils ont été appris sur des locuteurs qui lui sont proches au sens d'une certaine distance. Cette procédure évite le biais nécessairement introduit par une classification préalable des locuteurs, puisque la classe est définie pour chaque locuteur testé. Ainsi, l'influence de la mesure d_{IV} est plus facile à étudier. Cependant, il n'est pas réaliste pour une application en temps réel ou faiblement différé de réaliser l'apprentissage des modèles pour chaque locuteur de test. Dans un deuxième temps, si l'efficacité de cette approche et l'intérêt de la mesure sont démontrés, il est possible de pratiquer une classification automatique des locuteurs, et de choisir une approche plus classique d'adaptation rapide par multi-modèles.

3.2.1.3. Résultats expérimentaux

Des expériences de sélection de l'apprentissage nous ont permis d'étudier l'effet de la mesure d_{IV} [Barras *et al.*, 1994a]. Le protocole expérimental est le suivant: les 630 locuteurs de la base TIMIT prononcent deux phrases ("sa1" et "sa2") qui sont identiques pour tous les locuteurs et qui ont été exclues des expériences d'identification phonétique. Il s'agit donc d'un mode d'adaptation supervisé. Un modèle AR-vectoriel d'ordre 2 est calculé sur ces deux phrases pour chacun des locuteurs. Les paramètres acoustiques sont les 12 coefficients MFCC utilisés en reconnaissance. La distance entre deux locuteurs est calculée à partir de ces modèles. Pour un locuteur de test donné, nous calculons sa distance avec les 462 locuteurs d'apprentissage, et nous sélectionnons un certain pourcentage de ses plus proches voisins. Nous avons choisi arbitrairement pour ces expériences une sélection de 5%, 10% et 20% des plus proches locuteurs, soit respectivement 23, 46 et 92 locuteurs. Pour chacun des 24 locuteurs du noyau de test, l'apprentissage du système de base SB_1 est réalisé avec l'ensemble des locuteurs sélectionnés. Les modèles appris sont des modèles de Bakis à 3 états à une gaussienne à matrice de covariance diagonale par densité de probabilité, représentant des unités phonétiques indépendantes du contexte. Le modèle de langage est un bigramme phonétique.

Comparé à un apprentissage standard sur la totalité de l'ensemble d'apprentissage (cf. Tableau 3.1), l'apprentissage réduit à 10% des locuteurs apporte une augmentation de plus de 2% du taux de reconnaissance, ce qui est significatif au regard des 7215 phonèmes de test. Il est remarquable qu'une diminution de l'ensemble d'apprentissage améliore la qualité de la reconnaissance, alors que la tendance générale est de traiter une quantité toujours croissante de données d'apprentissage… Les coûts de calcul élevés nous ont cependant contraint à réaliser cette expérience avec un système de reconnaissance très simple et sur le noyau de test uniquement. Il n'est pas certain que l'amélioration observée soit aussi sensible sur un système plus complexe modélisant mieux les distributions acoustiques. Cependant, les résultats obtenus nous encouragent à continuer l'étude de la mesure d_{IV}, puisque la topologie choisie a effectivement amélioré la représentation des paramètres des modèles. Nous l'utilisons donc pour définir des classes de locuteurs et apprendre des multi-modèles, que nous comparons à de plus classiques bi-modèles par genre.

Locuteurs sélectionnés	Taux d'identification	Taux d'omission	Taux de substitution	Taux d'insertion	Taux de reconnaissance
5 %	61,2	12,8	25,9	6,5	**54,7**
10 %	61,9	12,8	25,3	6,0	**55,9**
20 %	61,6	13,0	25,4	5,9	**55,7**
100 %	59,1	13,6	27,3	5,6	**53,5**

Tableau 3.1. Taux de décodage obtenus par sélection de l'apprentissage avec le système de base SB_1 sur le noyau de test.

3.2.2 Classification a priori par genre

Les modèles par genre, constitués d'un modèle appris sur des locuteurs masculins et d'un modèle appris sur des locuteurs féminins, sont couramment utilisés en reconnaissance de la parole et améliorent généralement les performances par rapport aux modèles standards [Jouvet *et al.*, 1994]. Nous avons testé les bi-modèles par genre avec le système de base SB_G, constitué de modèles phonétiques indépendants du contexte. Les densités de probabilité sont représentées par une combinaison linéaire de G=1, 2, 4 ou 8 gaussiennes à matrice de covariance diagonale. Les 462 locuteurs de l'ensemble d'apprentissage sont divisés en deux sous-ensembles, l'un contenant les 326 hommes et l'autre les 136 femmes. Le réseau du décodage contraint l'enchaînement de modèles d'un seul genre (cf. Figure 3.4). Les tests sont réalisés sur les 168 locuteurs de l'ensemble de test complet, et l'intervalle de confiance est de ±0,4 %. Si l'on compare les résultats des modèles par genre avec les modèles standards pour un nombre de gaussiennes identique (cf. Tableau 3.2), les modèles par genre apportent une amélioration significative de 1% à 2% du taux de reconnaissance.

Nombre de gaussiennes	Taux d'identification	Taux d'omission	Taux de substitution	Taux d'insertion	Taux de reconnaissance	Gain
Modèles standards						
1	59,5	13,0	27,5	6,2	**53,3**	**-**
2	62,9	11,5	25,5	6,3	**56,7**	**-**
4	65,7	10,3	24,0	6,2	**59,5**	**-**
8	67,7	10,1	22,2	5,5	**62,2**	**-**
Modèles par genre						
1	61,4	12,6	26,0	6,7	**54,7**	**+1,4**
2	64,7	11,1	24,2	6,4	**58,3**	**+1,6**
4	67,2	10,2	22,6	5,8	**61,4**	**+1,9**
8	68,7	9,8	21,5	5,5	**63,2**	**+1,0**

Tableau 3.2. Taux de reconnaissance des modèles standards et des modèles par genre avec le système de base sur l'ensemble de test complet.

3.2.3 Classification automatique

Les expériences avec des modèles par genre montrent l'utilité des multi-modèles; une classification automatique des locuteurs grâce à la distance inter-locuteurs déjà étudiée permet de tester l'intérêt des modèles par classe de locuteurs [Mayrargue, 1995].

3.2.3.1. Classification des locuteurs

La mesure de similarité inter-locuteurs d_{IV}, initialement conçue pour l'identification du locuteur, semble permettre aussi de mettre en évidence des similarités phonétiques entre les locuteurs (cf. §3.2.1). Nous avons donc choisi d'utiliser cette distance pour réaliser la classification automatique des locuteurs d'apprentissage en k classes, le nombre de classes choisi devant être assez faible afin de conserver pour chacune une quantité de données d'apprentissage suffisante. La procédure expérimentée pour obtenir les classes de locuteurs est la suivante: un modèle ARV est calculé pour chaque locuteur d'apprentissage sur les phrases "sa1" et "sa2", et la mesure d_{IV} donne les distances entre tous les couples de locuteurs. A partir de la matrice des distances inter-locuteurs, une Classification Hiérarchique Ascendante (CHA) fournit un nombre k' élevé de classes ($k' >> k$). Nous avons observé que cette première phase de CHA génère des classes d'effectifs très hétérogènes. Dans une deuxième phase, les centroïdes des k classes de plus fort cardinal servent à initialiser une classification par Nuées Dynamiques (ou k-moyennes), afin de ré-attribuer les points des classes qui ont été éliminées. La classification des n_L locuteurs $E_L=\{L_i\} 1 \le i \le n_L$ est finalement réalisée de la manière suivante:

- Initialement, il y a autant de classes que de locuteurs:

$$n_C = n_L \tag{3.19}$$

  et chaque classe C_i contient un seul locuteur

$$C_i=\{L_i\},\ 1 \le i \le n_L \tag{3.20}$$

- Tant que le nombre de classes est supérieur au nombre voulu en fin de première phase, soit $n_C > k'$,
  - on recherche les deux classes les plus proches l'une de l'autre:

$$(\tilde{i},\tilde{j}) = \arg\min_{1 \le i,j \le n_C} d(C_i, C_j) \tag{3.21}$$

  - les classes choisies sont remplacées par la classe résultant de la fusion:

$$\tilde{C} = C_{\tilde{i}} \cup C_{\tilde{j}} \tag{3.22}$$

  - et le nombre total de classes est décrémenté:

$$\tilde{n}_C = n_C - 1 \tag{3.23}$$

- Seules les k classes d'effectif le plus élevé sont conservées:

$$\tilde{C}_i = C_{h(i)}, \quad 1 \leq i \leq k \tag{3.24}$$

où la fonction $h(j)$ ordonne les classes par effectif décroissant:

$$\operatorname{card} C_{h(j)} \geq \operatorname{card} C_{h(j+1)}, 1 \leq j < k' \tag{3.25}$$

- Tant que les classes sont modifiées entre deux itérations:
    - tous les locuteurs sont ré-attribués à la classe la plus proche:

$$\tilde{C}_i = \left\{ L \in E_L \;/\; \forall j, d(L, C_i) \leq d(L, C_j) \right\} \tag{3.26}$$

La distance entre deux classes est la moyenne des mesures d_{IV} entre les modèles ARV calculés sur tous les locuteurs de chacune des classes:

$$d(C_1, C_2) = \frac{1}{\operatorname{card} C_1 . \operatorname{card} C_2} \sum_{L_1 \in C_1, L_2 \in C_2} d_{IV}(L_1, L_2) \tag{3.27}$$

et la distance d'un locuteur à une classe est la moyenne des mesures entre les modèles du locuteur choisi et de chaque élément de la classe:

$$d(L, C_i) = \frac{1}{\operatorname{card} C_i} \sum\nolimits_{L' \in C_i} d_{IV}(L, L') \tag{3.28}$$

D'autres définitions sont possibles pour les distances (mesure du locuteur le plus proche, du plus lointain, du centroïde…) mais ne nous ont pas semblé *a priori* plus pertinentes.

Nous avons constitué 6 classes de locuteurs. Ce choix est motivé par la nécessité de ne conserver qu'un nombre réduit de classes, et correspond aux classes les plus stables lors de la CHA; l'examen de leurs effectifs (cf. Tableau 3.3) montre encore des disparités, avec une classe regroupant près du tiers de la base d'apprentissage (la classe 5) et une autre représentant à peine 6% de la base. D'autre part, la répartition par genre montre la très bonne discrimination entre les hommes et les femmes, puisque les femmes se retrouvent en quasi-totalité dans les deux premières classes, tandis que les hommes sont dans les quatre suivantes. Les rares locuteurs classés dans un genre opposé paraissent effectivement ambigus à l'écoute. La séparation hommes-femmes étant *a priori* la plus pertinente pour des modèles acoustiques, il est satisfaisant de la retrouver dans une classification plus fine. Par contre, il ne semble pas y avoir de corrélation entre cette classification et les dialectes de la base TIMIT. Des modèles phonétiques indépendants du contexte ont été appris sur les classes ainsi définies.

Classe	Locuteurs masculins	Locuteurs féminins	Effectif de la classe
1^f	3	93	96
2^f	0	42	42
3^h	69	1	70
4^h	75	0	75
5^h	152	0	152
6^h	27	0	27

Tableau 3.3. Effectifs des classes d'apprentissage après classification automatique des locuteurs (l'indice d'une classe constituée majoritairement d'hommes ou de femmes est suivi respectivement d'un ^{h} ou d'un ^{f}).

3.2.3.2. Adaptation explicite

Dans la phase d'adaptation explicite, les deux phrases de calibration "sa1" et "sa2" prononcées par le locuteur de test servent pour l'adaptation. Dans notre cas, cette phase d'adaptation est supervisée puisque tous les locuteurs prononcent les mêmes phrases d'adaptation, mais elle pourrait être non supervisée car la distance d_{IV} ne dépend pas du texte. La distance du locuteur de test avec les classes d'apprentissage permet d'attribuer le locuteur de test à la classe la plus proche, et de choisir l'ensemble de modèles adapté à ce locuteur. Nous donnons la répartition des locuteurs de test dans les différentes classes (cf. Tableau 3.4). La séparation par genre est bien respectée. D'autre part, aucun locuteur de test n'utilise les modèles de la dernière classe, qui est probablement d'effectif trop faible.

Classe	Locuteurs masculins	Locuteurs féminins	Nombre de locuteurs
1^f	2	42	44
2^f	1	13	14
3^h	45	0	45
4^h	29	0	29
5^h	36	0	36
6^h	0	0	0

Tableau 3.4. Classification supervisée des locuteurs de test.

Les résultats du décodage pour des modèles avec 1 à 8 gaussiennes par état (cf. Tableau 3.5) montrent un gain par rapport aux modèles standards pour un faible nombre de gaussiennes; mais ce gain disparaît avec des densités de probabilité plus complexes et un nombre élevé de gaussiennes. De plus, les résultats sont inférieurs à ceux obtenus avec des modèles par genre.

Nombre de gaussiennes	Taux d'identification	Taux d'omission	Taux de substitution	Taux d'insertion	Taux de reconnaissance	Gain
1	61,6	12,3	26,1	6,7	**54,9**	**+1,6**
2	64,8	10,9	24,3	6,6	**58,2**	**+1,5**
4	67,2	10,1	22,7	6,2	**61,0**	**+1,5**
8	68,5	9,6	22,0	6,2	**62,3**	**+0,1**

Tableau 3.5. Adaptation explicite supervisée par multi-modèles; taux de reconnaissance et gain par rapport aux modèles standards.

3.2.3.3. Adaptation implicite

Une approche implicite consiste, comme dans le cas des modèles par genre, à présenter les modèles de toutes les classes de locuteurs en parallèle dans le réseau de décodage. L'adaptation est donc simultanée au décodage puisque le meilleur chemin issu du décodage est obtenu avec la meilleure des classes. Il s'agit par nature d'une adaptation non supervisée. Les résultats sont quasiment identiques aux résultats obtenus avec un apprentissage supervisé (cf. Tableau 3.6), ce qui confirme au passage la cohérence de la classification et la pertinence acoustique de la distance choisie: les classes de locuteurs les plus proches du locuteur de test au sens de la distance ont permis d'apprendre les modèles les mieux adaptés à ce locuteur au sens de la probabilité d'émission. La suppression de la phase d'adaptation est bien sûr un avantage pour l'utilisateur; par contre, le réseau de décodage doit contenir les modèles de toutes les classes en parallèle, ce qui est plus coûteux pour la reconnaissance.

Nombre de gaussiennes	Taux d'identification	Taux d'omission	Taux de substitution	Taux d'insertion	Taux de reconnaissance	Gain
1	61,6	12,3	26,1	6,7	**54,9**	**+1,6**
2	64,8	10,9	24,4	6,6	**58,2**	**+1,5**
4	67,1	10,1	22,8	6,2	**60,9**	**+1,4**
8	68,5	9,6	21,9	6,1	**62,4**	**+0,2**

Tableau 3.6. Adaptation non supervisée par multi-modèles; taux de reconnaissance et gain par rapport aux modèles standards.

3.2.4 Conclusions

La comparaison des différentes méthodes d'adaptation rapide à base de multi-modèles, qu'il s'agisse de modèles par genre ou de modèles obtenus par classification automatique des locuteurs, et expérimentés en mode supervisé ou non supervisé, sont clairement à l'avantage des modèles par genre (cf. Tableau 3.7). Nous pensons que la distance inter-locuteurs n'est pas en cause, car les multi-modèles ont les meilleurs taux dans le cas de modèles simples, avec une gaussienne par état. Avec l'augmentation du nombre de gaussiennes, les modèles standards peuvent couvrir une plus grande variabilité, y compris celle des différentes classes de locuteurs. Par contre, la classification des locuteurs réduit l'ensemble d'apprentissage disponible pour chaque groupe de modèles; l'augmentation du nombre de gaussiennes est alors moins efficace.

Nombre de gaussiennes	Modèles standards (1 classe)	Modèles par genre (2 classes)	Adaptation supervisée (6 classes)	Adaptation non supervisée (6 classes)
1	53,3	54,7	54,9	54,9
2	56,7	58,3	58,2	58,2
4	59,5	61,4	61,0	60,9
8	**62,2**	**63,2**	**62,3**	**62,4**

Tableau 3.7. Comparaison des méthodes d'adaptation rapide sur l'ensemble de test complet pour les modèles phonétiques indépendants du contexte.

En conclusion, la classification des locuteurs permet de choisir des modèles appris sur des locuteurs effectivement plus proches, mais l'élimination complète des données d'apprentissage présentes dans les autres classes n'est pas souhaitable. Nous devons donc développer une approche mixte, combinant la notion de proximité qui semble exister, et tirant un meilleur parti de la totalité de la base d'apprentissage. Nous espérons de cette manière dépasser les performances des modèles par genre.

3.3. Généralisation des multi-modèles

L'approche par multi-modèles pose le problème d'un bon compromis entre la proximité des locuteurs d'apprentissage, qui conduirait à choisir des classes de taille réduite, et la quantité de données d'apprentissage, qui impose des classes de taille importante. Afin de surmonter cette difficulté, nous proposons une approche originale pour l'adaptation d'un système de RAP à un nouveau locuteur: l'apprentissage pondéré. La distance d_{IV} entre le nouveau locuteur testé et chaque classe de locuteurs d'apprentissage est calculée. Ces distances servent à pondérer la contribution des locuteurs d'apprentissage lors de la ré-estimation de modèles acoustiques indépendants du locuteur. La phase d'adaptation proposée est rapide et généralise le principe de l'adaptation par multi-modèles.

3.3.1 Principe d'un apprentissage pondéré

Des modèles acoustiques standards, indépendants du locuteur, sont appris sur la totalité de l'ensemble d'apprentissage suivant une procédure classique (cf. chapitre 2). Une classification des locuteurs d'apprentissage est ensuite réalisée. Pour chaque classe de locuteurs ainsi définie, les modèles sont ré-estimés par une itération de l'algorithme de Baum-Welch. La contribution d'une classe à l'estimation de chaque densité de probabilité des modèles est mémorisée à la suite de cette phase d'apprentissage; la capacité mémoire du système peut donc éventuellement limiter le nombre maximal de classes de locuteurs. La phase d'adaptation à un nouveau locuteur de test nécessite le calcul des distances entre ce locuteur et les classes d'apprentissage. Les modèles appris sur chacune des classes d'apprentissage sont combinés en fonction d'une pondération déduite de ces distances. La Figure 3.6 donne le récapitulatif des étapes lors des phases d'apprentissage et d'adaptation à un locuteur de test.

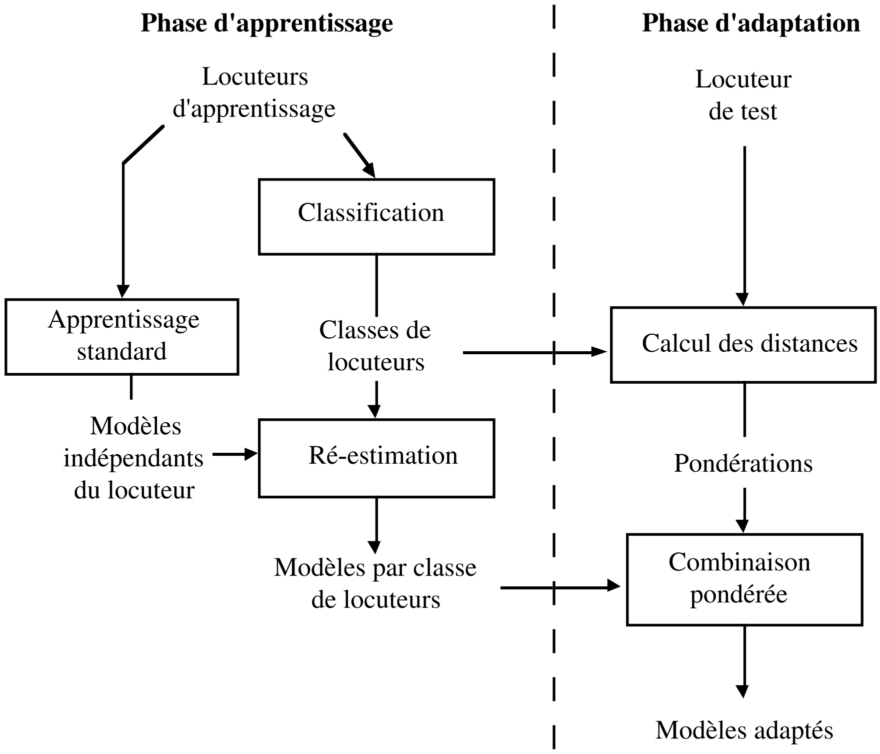

Figure 3.6. Phases d'apprentissage et d'adaptation pour un apprentissage pondéré.

3.3.1.1. Apprentissage partiel

Pour chaque classe de locuteurs, les modèles standards sont ré-estimés par une itération de l'algorithme de Baum-Welch sur les phrases d'apprentissage des locuteurs de cette classe. Ainsi, pour la classe c, le compteur des transitions vers l'état i à l'instant t est (cf. équation 1.70):

$$\omega_t^c(i) = \frac{\alpha_t^c(i)\beta_t^c(i)}{P(O_c/\lambda)} \tag{3.29}$$

d'où un nombre moyen de transitions par état:

$$\gamma_i^c = \sum_{t=1}^{T_c} \omega_t^c(i) \tag{3.30}$$

et la ré-estimation de la moyenne associée à l'état i, pour une densité de probabilité représentée par une mono-gaussienne, est donnée par (cf. équation 1.71):

$$\tilde{\mu}_i^c = \frac{1}{\gamma_i^c} S_i^c \text{ où } S_i^c = \sum_{t=1}^{T_c} \omega_t^c(i) o_t^c \tag{3.31}$$

Après cet apprentissage, nous conservons en mémoire le détail de la contribution de chaque classe aux paramètres des modèles, à savoir les quantités γ_i^c et S_i^c. Une ré-estimation classique sur l'ensemble d'apprentissage complet consisterait à cumuler ces compteurs sur l'ensemble des classes de locuteurs, soit par exemple:

$$\tilde{\mu}_i = \frac{1}{\gamma_i} \sum_{c=1}^{n_C} S_i^c \text{ avec } \gamma_i = \sum_{c=1}^{n_C} \gamma_i^c \tag{3.32}$$

Nous proposons, lors de la phase d'adaptation, de pondérer les compteurs de chaque classe en fonction de la distance du locuteur de test à cette classe; le calcul des pondérations est détaillé dans le paragraphe suivant. Si la pondération associée à la classe c est notée π_c la nouvelle moyenne de l'état i est:

$$\tilde{\mu}_i = \frac{1}{\gamma_i} \sum_{c=1}^{n_C} \pi_c S_i^c \text{ avec } \gamma_i = \sum_{c=1}^{n_C} \pi_c \gamma_i^c \tag{3.33}$$

Les matrices de covariance peuvent être ré-estimées en suivant une procédure analogue, mais il est possible aussi de conserver leur valeur initiale.

L'outil HERest de HTK permet de mémoriser les quantités intermédiaires γ_i^c et S_i^c à la suite d'une ré-estimation des modèles sur un ensemble de phrases. Ce mode de fonctionnement est initialement prévu pour fractionner l'ensemble d'apprentissage et paralléliser le fonctionnement de l'outil sur plusieurs machines. Nous l'utilisons pour calculer les quantités intermédiaires correspondant à chaque classe de locuteurs. La phase d'adaptation ne nécessite que le calcul des pondérations π_i et la combinaison linéaire des contributions de chaque classe de locuteurs à la moyenne de chaque état; cette phase est donc très rapide.

3.3.1.2. Calcul de la pondération

Une phase d'adaptation supervisée permet de calculer le modèle ARV d'un nouveau locuteur L sur quelques phrases d'adaptation, puis d'en déduire la distance du locuteur L à chaque classe d'apprentissage C_i. Dans notre cas, nous calculons cette distance comme la moyenne des distances entre le locuteur de test et chaque locuteur de la classe (cf. équation 3.28):

$$d(L, C_i) = \frac{1}{\operatorname{card} C_i} \sum\nolimits_{L' \in C_i} d_{IV}(L, L') \tag{3.34}$$

La pondération de chaque classe pour la ré-estimation des modèles doit être déduite de ces distances. Supposons que les classes sont ordonnées en fonction de la distance au locuteur de test, c'est-à-dire:

$$i \leq j \Rightarrow d(l,C_i) \leq d(l,C_j) \tag{3.35}$$

Nous souhaitons favoriser les classes les plus proches du locuteur de test; le choix le plus simple est alors une pondération binaire:

$$\begin{cases} \pi_i = 1, & \text{si } i \leq \kappa \\ \pi_i = 0 & \text{sinon} \end{cases} \tag{3.36}$$

Le nombre de classes conservées κ permet de régler l'importance de l'adaptation. Une valeur de $\kappa = n_C$ induit une pondération identique pour toutes les classes; les modèles obtenus sont identiques à ceux obtenus par une ré-estimation de Baum-Welch classique et il n'y a aucune adaptation. Avec une valeur de $\kappa = 1$, les modèles sont ré-estimés sur les locuteurs de la classe la plus proche, ce qui est similaire à l'adaptation par multi-modèles étudiée précédemment. La pondération peut être aussi une fonction décroissante avec la distance de la classe, comme une fonction gaussienne:

$$\pi_i = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{i^2}{2\sigma^2}} \tag{3.37}$$

Avec une pondération judicieusement choisie, nous espérons obtenir une estimation simultanément robuste et précise des paramètres des modèles, tenant compte de l'ensemble des données d'apprentissage ainsi que de la proximité des locuteurs.

3.3.1.3. Comparaison avec d'autres approches

D'autres approches ont été proposées pour résoudre le problème de la diminution des données d'apprentissage par classe de locuteurs dans le cas des multi-modèles. A. Ljolje réalise une "classification floue". La probabilité d'appartenance du locuteur testé à chaque classe permet de choisir la pondération attribuée aux gaussiennes des lois de probabilité [Ljolje, 1993]; mais l'amélioration des taux de reconnaissance n'est pas significative. La modification des poids attribués aux gaussiennes est aussi la solution retenue dans [Kosaka *et al.*, 1993]. Chaque locuteur L_k contribue à une des gaussiennes de chaque loi de probabilité dans l'état j par un poids g_k identique dans tous les états des modèles:

$$b_j(o) = \sum_{k=1}^{n_L} g_k \mathcal{N}(o, \mu_{jk}, \Sigma_{jk}),\ 1 \leq j \leq N \tag{3.38}$$

et les poids sont ré-estimés avec la procédure de Baum-Welch. Cette méthode d'adaptation améliore la reconnaissance des phrases de 4% par rapport au système initial pour une durée d'adaptation d'à peine 0,6 seconde.

3.3.2 Expériences

Comme dans le cas de l'adaptation par multi-modèles, l'adaptation par apprentissage pondéré doit être comparée à un apprentissage standard et à l'utilisation de modèles par genre. Les expériences portent sur le système de référence constitué de modèles dépendants du contexte.

3.3.2.1. Résultats avec les modèles par genre

L'étude des multi-modèles nous a montré que les modèles par genre apportent une amélioration des performances par rapport aux modèles standards. Cependant, les modèles par genre expérimentés au §3.2.2 ont été obtenus avec des modèles phonétiques indépendants du contexte; nous voulons réaliser un apprentissage par genre du système de référence pour évaluer l'apport de l'apprentissage pondéré.

Des modèles par genre peuvent être obtenus comme un cas particulier de l'apprentissage pondéré. Les locuteurs d'apprentissage sont séparés en deux classes en fonction du genre. Puis la distance de chaque locuteur de test aux deux classes d'apprentissage est calculée avec la mesure d_{IV}. Pour chaque locuteur de test, la plus proche des deux classes est affectée d'une pondération $\pi_i = 1$, et l'autre d'une pondération nulle. Dans 96% des cas, cette détermination automatique du genre est correcte. En fonction de ces pondérations, les modèles initiaux sont ré-estimés sur la partie de l'ensemble d'apprentissage correspondant soit aux hommes, soit aux femmes.

Les expériences ont été réalisées sur le noyau de test avec le système de référence. Lors de la ré-estimation, il est possible de modifier les moyennes, les variances, ou les poids des gaussiennes ainsi que les probabilités de transition. Nous avons comparé une ré-estimation complète des modèles, une ré-estimation conservant la valeur initiale des variances et une ré-estimation conservant les moyennes et les variances des gaussiennes (cf. Tableau 3.8). Dans ce dernier cas, les densités de probabilité ne sont modifiées que par la ré-estimation du poids des gaussiennes, ce qui est comparable à ce que proposent A. Ljolje et T. Kosaka (cf. §3.3.1.3); nous n'observons dans cette dernière expérience aucune amélioration du taux de reconnaissance. La ré-estimation de la moyenne et éventuellement de la variance apporte une amélioration de près de 1% du taux de reconnaissance, avec un léger avantage pour le cas où la variance n'est pas modifiée. Cependant, le rayon de l'intervalle de confiance à 95% est de ±1,1% et ces écarts ne sont pas significatifs.

Taux d'identification	Taux d'omission	Taux de substitution	Taux d'insertion	Taux de reconnaissance
Système de référence				
74,8	4,9	20,3	9,7	**65,1**
Ré-estimation complète des modèles				
75,6	4,9	19,5	9,8	**65,8**
Ré-estimation des modèles (variances fixées)				
75,7	5,0	19,2	9,7	**66,0**
Ré-estimation des modèles (moyennes et variances fixées)				
75,1	5,1	19,8	10,0	**65,1**

Tableau 3.8. Adaptation par genre du système de référence. Expériences sur le noyau de test avec un facteur d'échelle $s = 1$.

Nous avons choisi pour les expériences suivantes une ré-estimation des modèles en conservant la variance, car nous supposons que l'estimation initiale de la variance est plus robuste que sa ré-estimation sur un ensemble plus réduit. Sur l'ensemble de test complet et avec le facteur d'échelle $s = 3$ pour le bigramme phonétique, le taux de reconnaissance passe de 69,1% pour le système de référence à 70,0% après l'adaptation par genre (cf. Tableau 3.9). Cette augmentation est significative puisque l'intervalle de confiance est de ±0,4%, et elle est réalisée sans augmentation du nombre total de paramètres du système de reconnaissance.

Taux d'identification	Taux d'omission	Taux de substitution	Taux d'insertion	Taux de reconnaissance
Système de référence				
74,8	6,6	18,7	5,7	**69,1**
Ré-estimation des modèles (variances fixées)				
75,4	6,5	18,2	5,3	**70,0**

Tableau 3.9. Adaptation par genre du système de référence.
Expériences sur le test complet avec un facteur d'échelle $s = 3$.

Il est intéressant de constater que l'augmentation du taux de reconnaissance provient principalement d'une amélioration de la reconnaissance pour les locuteurs féminins (cf. Tableau 3.10). L'amélioration s'élève à 2% pour les femmes alors qu'elle est plus faible pour les hommes, à 0,4%; cette différence est peut-être causée par la plus faible représentation des femmes dans la base d'apprentissage (un tiers des locuteurs environ) ce qui entraîne un biais plus important pour ces locuteurs lors de l'utilisation de modèles standards.

Locuteurs testés	Nombre de segments phonétiques	Reconnaissance par des modèles standards (%)	Reconnaissance par des modèles par genre (%)	Augmentation du taux de reconnaissance
Hommes	33624	70,0	70,4	**+0,4**
Femmes	17130	67,3	69,3	**+2,0**
Totalité	50754	69,1	70,0	**+0,9**

Tableau 3.10. Adaptation par genre du système de référence. Détail des taux de reconnaissance pour les hommes et les femmes.

3.3.2.2. Ré-estimation sur les classes les plus proches

Afin que l'effet de la distance inter-locuteurs puisse être étudié avec plus de précision, les pondérations doivent être appliquées à des classes d'effectif réduit. L'algorithme de classification proposé lors de l'étude des multi-modèles (§3.2.3) a donc été utilisé pour obtenir 20 classes de locuteurs, au lieu des 6 classes déjà étudiées. Parmi ces 20 classes, 7 sont constituées quasi exclusivement de femmes, et 13 d'hommes. Les confusions hommes-femmes concernent un locuteur féminin et deux locuteurs masculins. Les effectifs des classes sont compris entre 14 et 45 locuteurs, pour une moyenne de 23 locuteurs par classe.

Une première expérience d'adaptation par apprentissage pondéré a été réalisée avec cette classification, en conservant pour chaque locuteur du noyau de test les κ plus proches classes pour la ré-estimation des modèles (cf. Tableau 3.11). Les valeurs initiales des variances des gaussiennes sont conservées. Ces expériences sont décevantes dans la mesure où les taux de reconnaissance des modèles par genre, de 66% sur le noyau de test avec un facteur d'échelle s=1, ne sont pas dépassés. La distance étudiée permet effectivement de retrouver la séparation par genre, mais rien ne prouve qu'une discrimination plus fine obtenue par cette distance soit pertinente au sens de la reconnaissance acoustique.

Nombre de classes κ	Pourcentage de l'apprentissage	Taux de reconnaissance
2	10,0	65,1
4	20,0	65,3
6	30,0	65,9
8	40,0	65,9
10	50,0	65,6

Tableau 3.11. Adaptation par ré-estimation sur les κ classes les plus proches. Expériences sur le noyau de test avec un facteur d'échelle s = 1.

Cependant, l'absence d'amélioration des taux de reconnaissance peut être due à la diminution de la quantité de données d'apprentissage. Nous allons donc chercher à mettre en évidence dans l'expérience suivante la pertinence de la notion de proximité inter-locuteurs en restant dans le cadre de la séparation *a priori* par genre, qui est la meilleure obtenue jusqu'à présent.

3.3.2.3. Classes les plus proches ou les plus lointaines

Une nouvelle expérience a été réalisée pour comparer, pour un genre donné, une adaptation sur les classes les plus proches avec une adaptation sur les classes les plus lointaines au sens de la mesure inter-locuteurs. Si la mesure n'apporte pas d'information pertinente autre que la détermination du genre, les deux types d'adaptation ne doivent pas donner de résultats sensiblement différents.

Le genre d'un locuteur de test est tout d'abord déterminé automatiquement en fonction du genre des locuteurs d'apprentissage les plus proches, et les classes d'apprentissage correspondant au genre opposé sont supprimées de l'adaptation. Dans notre cas, nous conservons 13 classes pour un locuteur identifié comme un homme et 7 classes pour un locuteur identifié comme une femme. Un apprentissage réalisé sur ces classes produit naturellement le même résultat que l'adaptation par genre présentée au §3.3.2.1. Nous réalisons ensuite l'adaptation en ré-estimant les modèles sur les κ classes les plus proches du locuteur de test, puis sur les κ classes les plus lointaines.

Nous observons une très nette supériorité de la reconnaissance par les modèles les plus proches, pour une quantité de données d'apprentissage comparable (cf. Tableaux 3.12 et 3.13). Cela signifie qu'une information utile à la reconnaissance a été apportée par la distance inter-locuteurs étudiée, en plus de la détermination du genre. L'intervalle de confiance est de 1,9% pour les résultats obtenus sur les 8 femmes, et de 1,4% pour les résultats sur les 16 hommes du noyau de test. Les différences observées pour un nombre de classe faible ($\kappa \leq 1$ pour les femmes et $\kappa \leq 4$ pour les hommes) sont donc significatives. Les performances obtenues avec des modèles par genre ne sont pas dépassées pour un nombre réduit de classes d'apprentissage, probablement en raison de la trop grande réduction de l'ensemble d'apprentissage. Cependant, les taux de reconnaissance obtenus avec les 3 plus proches classes d'hommes ou de femmes sont quasiment identiques à ceux obtenus avec la totalité de l'ensemble d'apprentissage; à l'inverse, la reconnaissance avec un apprentissage sur les classes les plus éloignées n'atteint ce maximum que pour la quasi-totalité des classes (cf. Figures 3.7 et 3.8).

3.3.3. Conclusions

Les résultats obtenus confirment la pertinence acoustique de la distance inter-locuteurs choisie, même si les conditions expérimentales ne permettent pas de proposer un apprentissage plus performant que les modèles par genre. La base de données TIMIT n'offre qu'une durée d'élocution très réduite par locuteur, de 20 secondes environ. Cette durée est probablement insuffisante pour un apprentissage robuste des densités de probabilité sur le voisinage du locuteur testé. Mais nous avons observé que, pour un locuteur de test donné et pour un nombre de paramètres constant du système de reconnaissance, la quantité de données d'apprentissage peut être réduite à une faible fraction de la taille d'origine (dans notre cas environ 15% de la base TIMIT) tout en permettant d'atteindre la qualité d'un apprentissage par genre.

Nombre de classes κ	Pourcentage des locuteurs féminins	Ré-estimation sur les classes proches	Ré-estimation sur les classes éloignées	Différence entre les taux de reconnaissance
1	14	65,9	60,9	5,0
2	29	66,7	64,0	2,7
3	43	67,8	65,6	2,3
4	57	67,3	65,9	1,4
5	71	68,1	67,1	1,0
6	86	67,8	67,7	0,1
Modèles par genre				
7	100	68,0		0,0

Tableau 3.12. Taux de reconnaissance pour les 8 locuteurs féminins du noyau de test après ré-estimation sur les κ classes de femmes les plus proches ou les plus éloignées.

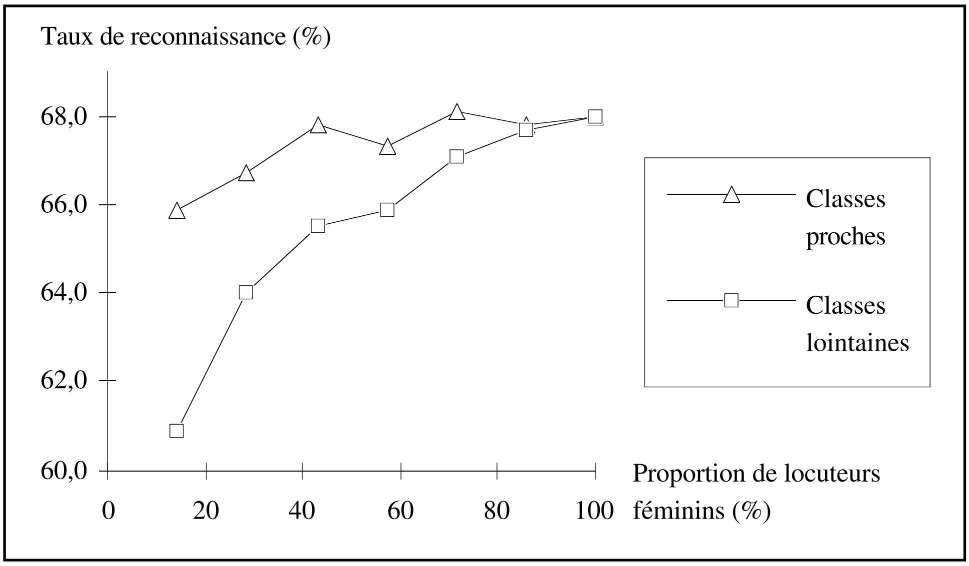

Figure 3.7. Effet de la proximité des classes d'apprentissage sur les taux de reconnaissance pour les 8 locuteurs féminins du noyau de test.

Nombre de classes κ	Pourcentage des locuteurs masculins	Ré-estimation sur les classes proches	Ré-estimation sur les classes éloignées	Différence entre les taux de reconnaissance
1	8	61,9	49,7	12,2
2	15	64,5	57,5	7,0
3	23	64,7	59,9	4,8
4	31	64,7	61,1	3,6
5	38	64,4	61,8	2,6
6	46	64,9	63,1	1,8
7	54	64,6	62,9	1,7
8	62	64,7	64,1	0,6
9	69	64,8	63,7	1,1
10	77	64,8	63,9	0,9
11	85	65,1	64,3	0,8
12	92	65,1	64,5	0,6
Modèles par genre				
13	100	65,0		0,0

Tableau 3.13. Taux de reconnaissance pour les 16 locuteurs masculins du noyau de test après ré-estimation sur les κ classes d'hommes les plus proches ou les plus éloignées.

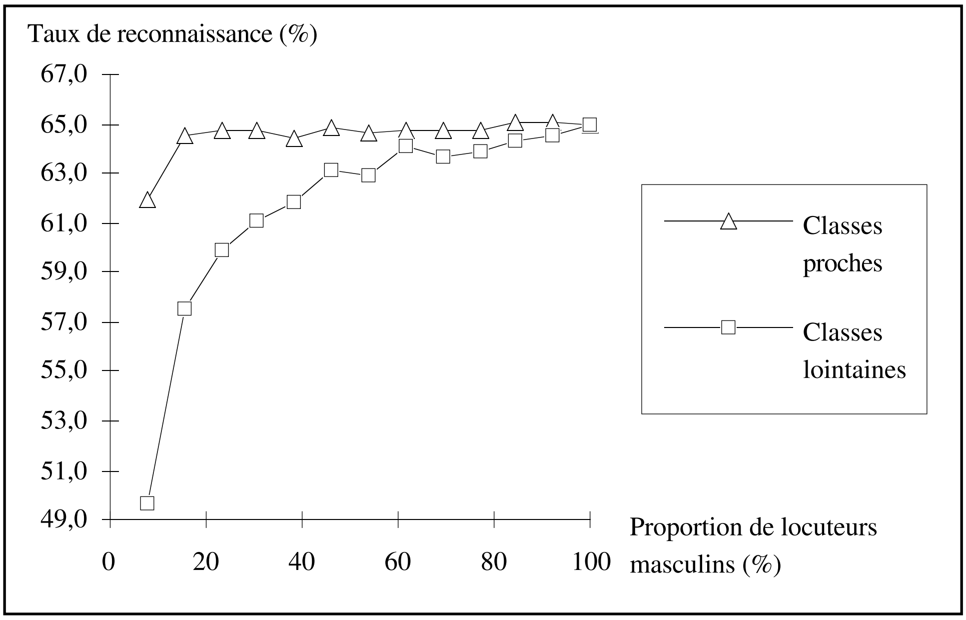

Figure 3.8. Effet de la proximité des classes d'apprentissage sur les taux de reconnaissance pour les 16 locuteurs masculins du noyau de test.

3.4. Conclusions

Nous avons montré qu'une distance inter-locuteurs initialement conçue pour l'identification du locuteur apporte une information de proximité acoustique pertinente pour la reconnaissance de la parole; cette information est plus précise qu'une simple détermination du genre. Il est donc possible de choisir, pour un locuteur de test donné, un ensemble de locuteurs d'apprentissage adaptés à ce locuteur. Nous avons proposé un nouveau mode d'adaptation, basé sur une classification préalable des locuteurs d'apprentissage au sens de cette distance inter-locuteurs et un ré-apprentissage partiel de modèles phonétiques pour chaque classe d'apprentissage. La phase d'adaptation est alors rapide puisqu'elle se limite à une combinaison pondérée des modèles partiels. Ce mode d'adaptation est une généralisation de l'approche classique par multi-modèles.

Les meilleurs taux de reconnaissance avec notre système de référence ont cependant été obtenus avec des modèles par genre. La sélection des locuteurs les plus proches entraîne en effet une réduction de la quantité de données d'apprentissage, et la faible durée d'élocution par locuteur dans la base de données TIMIT ne permet pas une estimation robuste des densités de probabilité des modèles sur le voisinage ainsi défini. Il est vraisemblable que les traitements proposés apporteront une amélioration sur des bases de données plus importantes, disposant pour chaque locuteur d'une durée d'enregistrement de plusieurs minutes. Nos expériences montrent déjà que pour un locuteur de test donné et pour une complexité fixée du système de reconnaissance, la ré-estimation des modèles sur une faible proportion de l'ensemble d'apprentissage permet d'atteindre la qualité d'un apprentissage par genre. Enfin, il serait intéressant de comparer la mesure de similarité inter-locuteurs choisie dans notre étude avec d'autres distances.

4. Contrôle temporel du décodage

Les modèles de Markov fournissent un cadre théorique général qui est à l'origine des améliorations récentes des systèmes de reconnaissance de la parole. Cependant, ils sont basés sur un certain nombre d'hypothèses qui ne sont pas toujours valides. En particulier, la probabilité de transition inter-états est indépendante du passé du processus, et cela entraîne une modélisation de la durée d'émission en décroissance exponentielle qui est jugée généralement insatisfaisante pour représenter la parole. Une modélisation explicite de la durée d'émission des modèles phonétiques a été proposée [Russel & Moore, 1985; Levinson, 1986]; ces modèles de durée sont déterminés *a priori* par une statistique sur la durée des phonèmes. Cependant, les systèmes ainsi développés sont complexes, et leur efficacité parfois réduite en reconnaissance a limité leur utilisation.

D'autres approches du contrôle temporel sont fondées sur l'analyse du signal acoustique en cours de décodage. Nous proposons ainsi de réaliser un contrôle temporel du décodage acoustico-phonétique en exploitant de manière explicite une information temporelle précise, provenant d'une analyse par ruptures de modèle. Cette analyse détecte des événements infra-phonémiques dans le signal de parole [André-Obrecht, 1988], qui sont pris en compte lors du décodage par une modification dynamique des probabilités de transitions entre les modèles acoustiques. Des expériences de DAP et de décodage lexical avec le système de référence sur la base TIMIT montrent l'apport de cette approche pour la réduction des insertions phonétiques et lexicales.

4.1. Techniques de contrôle temporel

Une des faiblesses majeures des HMM standards concerne la modélisation de la durée d'un état, qui suit par défaut une loi en décroissance exponentielle peu réaliste en parole. Des études statistiques sur la durée de réalisation d'un phonème montrent que cette durée suit plus vraisemblablement une loi Gamma [Crystal & House, 1982]. Certains auteurs ont donc introduit des modèles explicites de la durée d'émission, par exemple avec une loi de Poisson [Russel & Moore, 1985; El-Bèze, 1990] ou une loi Gamma [Levinson, 1986]. D'autres méthodes de contrôle temporel sont fondées sur l'analyse du signal acoustique en cours de décodage, et modélisent par exemple la vitesse d'élocution du locuteur de test.

4.1.1. Loi de durée implicite des modèles de Markov

La durée d'émission par une machine de Markov est contrôlée implicitement par la valeur des probabilités de transitions entre les états. Soit a_{ii} la probabilité de bouclage sur l'état courant, et $(1-a_{ii})$ la probabilité de passage à un autre état, alors la probabilité que τ trames exactement aient été émises par l'état i puis que le processus quitte cet état est:

$$p_i(\tau) = {a_{ii}}^{\tau-1}(1-a_{ii}) \tag{4.1}$$

et le nombre moyen de trames émises dans un état est alors:

$$\overline{\tau}_i = \sum_{\tau} \tau . p_i(\tau) = \sum_{\tau} \tau . {a_{ii}}^{\tau-1}(1-a_{ii}) = \frac{1}{1-a_{ii}} \tag{4.2}$$

La probabilité $p_i(\tau)$ suit une loi en décroissance exponentielle (cf. Figure 4.1) qui n'est pas nécessairement adaptée au phénomène acoustique modélisé.

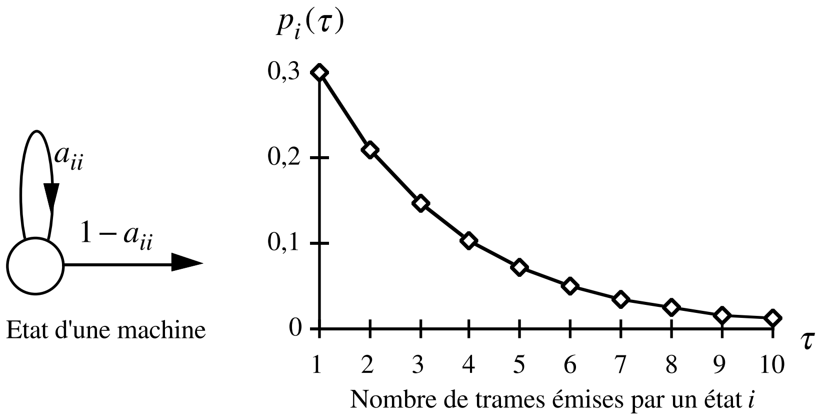

Figure 4.1. Densité de probabilité de durée associée à un état d'une machine de Markov (pour a_{ii}=0,7).

T. Crystal et A. House ont étudié la durée des unités phonétiques dans de la parole continue [Crystal & House, 1982]. Suivant leurs observations, l'histogramme de durée des phonèmes est assez bien modélisé par une distribution Gamma à deux paramètres définie par:

$$f(t) = \frac{1}{\Gamma(\nu)} \eta^{\nu} t^{\nu-1} e^{-\eta t} \tag{4.3}$$

où Γ est la fonction Gamma et où les paramètres η et ν de la distribution sont reliés à la moyenne μ et à la variance σ^2 de la durée des phonèmes par:

$$\eta = \frac{\mu}{\sigma^2} \quad \text{et} \quad \nu = \frac{\mu^2}{\sigma^2} \tag{4.4}$$

La distribution des durées dans un état d'un HMM standard n'est donc pas adaptée à la modélisation phonétique.

Lorsque la topologie des modèles acoustique est plus complexe qu'une simple machine à un seul état, une modélisation plus réaliste de la durée est possible. T. Crystal & A. House ont montré qu'une topologie gauche-droite à 6 états, avec des sauts autorisés de chaque état vers l'état final, permet une bonne modélisation de la durée d'émission des segments phonétiques [Crystal & House, 1988]. Considérons pour simplifier une cascade d'états, c'est-à-dire une chaîne de Markov gauche-droite de N états sans saut d'état, pour laquelle les probabilités de bouclage sur chaque état sont toutes identiques de valeur $a_{ii} = a$ (cf. Figure 4.2).

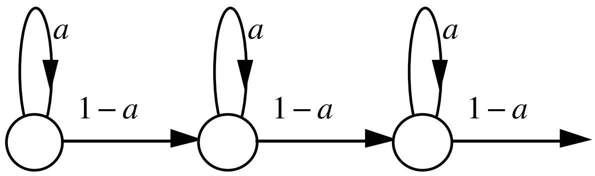

Figure 4.2. Cascade d'états de longueur N=3.

Alors la probabilité que cette chaîne émette exactement τ trames est [Wang, 1995]:

$$p_N(\tau) = C_{\tau-1}^{N-1} a^{\tau-N}(1-a)^N, \quad \tau \geq N \tag{4.5}$$

soit par exemple pour une chaîne de longueur N=3:

$$p_3(\tau) = \frac{(\tau-1)(\tau-2)}{2} a^{\tau-3}(1-a)^3, \quad \tau \geq 3 \tag{4.6}$$

La densité de probabilité de la durée totale d'émission est alors proche d'une distribution Gamma (cf. Figure 4.3). Des topologies plus complexes permettent d'approcher n'importe quelle loi de durée; ainsi, H. Ney utilise pour le système SPICOS un modèle phonétique à 6 états, pour 3 densités de probabilité différentes [Ney, 1990]. Mais une modélisation explicite des durées d'émission a aussi été proposée.

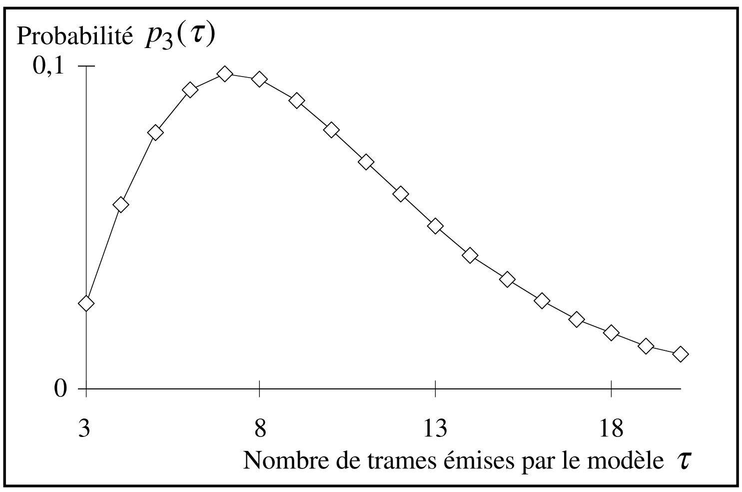

Figure 4.3. Durée d'émission pour une chaîne de Markov de longueur N=3 et une probabilité de bouclage sur chaque état a=0,7.

4.1.2. Modèle de durée explicite

Pour pallier ce qui apparaît comme une insuffisance des modèles standards, des HMM spécialisés ont été développés avec une modélisation explicite de la durée d'émission. Ces modèles entraînent malheureusement une forte augmentation du coût de calcul sans apporter toujours une amélioration décisive des taux de reconnaissance, ce qui explique probablement que l'usage des modèles de durée ne se soit pas généralisé.

4.1.2.1. Modèle à durée d'émission variable

J. Ferguson a proposé une modélisation explicite de la durée dans le cadre de modèles semi-markoviens [Ferguson, 1980]. Le bouclage sur un état n'est plus autorisé, et la probabilité que τ observations soient émises par l'état i est donnée par $p_i(\tau)$. Les lois de durée deviennent donc des paramètres supplémentaires du modèle. Les probabilités de durée sont ré-estimées avec les autres paramètres du modèle par une variante de l'algorithme de Baum-Welch. Cette approche présente cependant un certain nombre d'inconvénients. D'une part, le nombre de paramètres supplémentaires est très élevé, puisqu'il est nécessaire d'estimer une probabilité pour toutes les durées d'émission possibles dans chaque état, et l'estimation obtenue n'est fiable que si elle est réalisée sur une grande quantité de données d'apprentissage. D'autre part, le surcoût de calcul est important lors de l'apprentissage et du décodage; si le nombre maximal de trames émises dans un état est limité à D, la durée du décodage peut être multipliée par $D^2/2$.

Les algorithmes de reconnaissance ont été modifiés afin de limiter l'augmentation du coût de décodage lors de l'utilisation de HMM à émissions de durée variable. Ainsi, A. Bonafonte présente une version optimisée de l'algorithme de Viterbi pour laquelle l'augmentation de complexité est simplement proportionnelle à la durée moyenne d'émission au lieu d'être de l'ordre du carré de la durée maximale [Bonafonte *et al.*, 1995]. Le modèle de durée peut aussi être incorporé au moyen d'un post-traitement [Rabiner, 1989]. Un décodage standard est tout d'abord réalisé en recherchant les n meilleures solutions ou en constituant un graphe de mots, puis la nouvelle probabilité d'émission est recalculée pour ces phrases en intégrant *a posteriori* la modélisation de la durée. Mais cette approche introduit un délai dans la reconnaissance d'une phrase, et elle n'est pas optimale, car elle n'assure pas que la meilleure phrase au sens du critère final ait été retenue lors de la première phase sans modèle de durée.

P. Ramesh & J. Wilpon proposent des "HMM inhomogènes" (IHMM) comme alternative aux modèles de Ferguson [Ramesh & Wilpon, 1992]. Les probabilités de transition des IHMM dépendent de la durée du séjour dans l'état courant, elles sont donc décrites par $A=\{a_{ij}(\tau)\}_{1\leq i,j\leq N,\ 1\leq\tau\leq D}$. La complexité de la ré-estimation et du décodage est alors simplement augmentée d'un facteur proportionnel à D. Par une implémentation sous-optimale de l'algorithme, le coût de calcul devient identique aux HMM traditionnels, et une réduction significative des erreurs est observée en reconnaissance de suites de chiffres sur la base TI-DIGITS et en reconnaissance de mots sur la tâche *Resource Management*. S. Vaseghi a développé indépendamment un modèle identique [Vaseghi, 1995]. La modélisation de la durée associée à des contraintes de durées minimales et maximales d'émission améliore de 5,6% l'identification de lettres anglaises prononcées de manière isolée. R. Sitaram et T. Sreenivas montrent la pertinence d'une modélisation par IHMM sur une tâche de discrimination entre des occlusives, des nasales et des affriquées [Sitaram & Sreenivas, 1994].

4.1.2.2. Modélisation de la durée par une loi de probabilité

Pour réduire le nombre de paramètres à estimer dans un modèle à durée d'émission variable, il est possible de modéliser la durée par une loi de probabilité. Il suffit alors d'estimer les paramètres de la loi et non plus l'ensemble des $p_i(\tau)$ pour toutes les valeurs possibles de τ. M. Russel et R. Moore ont par exemple proposé la modélisation de la durée par une loi de Poisson [Russel & Moore, 1985]. S. Levinson utilise une distribution Gamma pour des modèles semi-markoviens intitulés CVD-HMM (*Continuously Variable Duration* HMM) [Levinson, 1986]. A. Ljolje & S. Levinson appliquent cette dernière modélisation à la parole [Ljolje & Levinson, 1991]; chaque état d'un HMM ergodique représente un phonème, et les probabilités de transition inter-états sont estimées par un bigramme phonétique. Avec un modèle phonétique à un état sans modélisation de durée, les taux d'erreurs de décodage acoustico-phonétique sur 450 phrases phonétiquement équilibrées prononcées par un seul locuteur augmentent de 50%. Dans le cadre d'expériences avec des HMM semi-continus, X. Huang applique un lissage gaussien à l'histogramme des durées pour améliorer la robustesse de l'estimation; l'utilisation du modèle de durée apporte une diminution de 14% des erreurs phonétiques lors du décodage de 98 phrases prononcées par un locuteur masculin [Huang, 1992]. C. Mitchell et L. Jamieson étudient plus généralement les densités de probabilité discrètes et continues [Mitchell & Jamieson, 1993]. Ils proposent une procédure automatique pour choisir la loi de durée la mieux adaptée à un état dans le cadre de la famille exponentielle, comprenant par exemple les versions discrètes et continues des lois gaussiennes et Gamma.

4.1.2.3. Modèle à durée d'émission bornée

Des contraintes plus simples sur la durée d'émission des modèles peuvent cependant être suffisantes. Dans le système HARPY développé à CMU, le décodage est contraint à une durée minimale et maximale dans certains états [Klatt, 1977]. M. Bush et G. Kopec obtiennent les meilleurs résultats pour une tâche de reconnaissance de chiffres enchaînés avec une durée minimale de 50 ms par segment acoustique [Bush & Kopec, 1987]. H. Gu, C. Tseng et L. Lee étudient un modèle à durée d'état bornée, pour lequel une durée minimale et maximale est imposée à chaque état [Gu *et al.*, 1991]. Ils comparent ce modèle avec un HMM standard et avec un HMM associé à un modèle de durée plus complexe comme une loi de Poisson ou une loi Gamma. De manière surprenante, les meilleurs résultats de reconnaissance sur des syllabes de mandarin sont fournis par le modèle à durée d'état bornée, avec des HMM discrets ou continus. La contrainte d'une durée phonétique minimale, déterminée pour chaque phonème, apporte une diminution de 22% du taux d'erreur lexical avec le système de RAP de 86000 mots développé à l'INRS [Gupta *et al.*, 1992]. Sur ce même système, P. Dumouchel et D. O'Shaughnessy ont expérimenté une durée phonétique bornée, et une modélisation de la durée des phonèmes par une combinaison linéaire de gaussiennes; l'expérience avec la durée phonétique bornée est la seule à apporter une légère amélioration par rapport au système standard [Dumouchel & O'Shaughnessy, 1995].

4.1.2.4. Exemples d'intégration des modèles de durée

Pour le système SPHINX de CMU, K.F. Lee introduit un modèle de durée au niveau des mots uniquement [Lee, 1988]; les modèles phonétiques ont en effet une topologie spéciale qui tient compte de manière explicite des émissions courtes, et un modèle de durée au niveau phonétique est jugé superflu. La probabilité de durée est intégrée au cours du décodage au niveau de la transition inter-états sans surcoût de calcul, mais de manière sous-optimale. L'apport du contrôle de la durée des mots est significatif pour un décodage sans contrainte grammaticale; mais cet apport disparaît avec une grammaire par paire de mots ou un bigramme lexical.

D. Burshtein incorpore la modélisation de la durée des états ou des modèles de mots au niveau de la transition inter-trames dans un modèle standard [Burshtein, 1995]. La distribution observée des durées est comparée aux lois Gamma, gaussienne ou exponentielle au sens de la mesure de Kullback-Liebler, ce qui conduit à retenir la loi Gamma pour modéliser les durées. Une pénalité de transition est ajoutée lors de chaque transition en fonction de la durée passée dans l'état ou dans le modèle. Lors d'expériences de reconnaissance de chiffres enchaînés, une réduction de 43% des erreurs est observée; elle est due principalement au contrôle de durée au niveau des états.

Finalement, il est nécessaire d'imposer une loi de durée explicite à un modèle de Markov lorsqu'un phonème est modélisé par un seul état, puisque la loi de durée exponentielle n'est pas satisfaisante. Mais S. Levinson rappelle l'équivalence formelle entre le modèle semi-markovien de Ferguson et un modèle standard à la topologie adéquate [Levinson, 1986], et il est possible de réaliser avec des HMM standards une modélisation implicite de la durée en choisissant une topologie adaptée (cf. §4.1.1). Y. Gotoh, M. Hochberg et H. Silverman remplacent chaque état d'un modèle semi-markovien par une chaîne de trois états dans un modèle standard, et ils réalisent l'apprentissage des probabilités de transition d'un modèle au moyen d'une estimation par le critère MAP [Gotoh *et al.*, 1994]. Les expériences consistent en la reconnaissance de chiffres et de lettres enchaînés. La modélisation de la distribution empirique des durées se compare à celle obtenue avec un modèle semi-markovien associé à une loi de Poisson; mais la complexité de l'apprentissage est réduite. Dans le système hybride neuronal et markovien développé par T. Robinson, la modélisation de la durée phonétique est aussi réalisée en remplaçant chaque état par une chaîne d'états [Robinson *et al.*, 1994]. La contrainte d'une durée minimale et d'une durée suivant une loi de Poisson ou une loi Gamma sont testées. Les expériences de DAP sur la base TIMIT ne montrent quasiment aucune amélioration, tandis que les erreurs de reconnaissance en mots sur la base *Wall Street Journal* sont réduites de 25% grâce à ce contrôle de durée.

4.1.3. Contrôle temporel fondé sur l'analyse du signal

La modélisation de la durée, qu'elle soit réalisée explicitement au moyen d'une loi de probabilité ou implicitement grâce à une topologie adaptée des modèles acoustiques, résulte d'une connaissance *a priori* des modèles; mais les observations restent exclusivement acoustiques et indépendantes les unes des autres. D'autres approches pour le contrôle temporel ont été proposées afin de mieux tenir compte du signal au cours du décodage. Des informations de nature segmentale sont intégrées au niveau de la représentation du signal, de l'apprentissage des modèles, du décodage, ou en post-traitement. En particulier, une analyse à débit variable est souvent réalisée à partir d'une pré-segmentation du signal. Les variations du débit d'élocution ont aussi été modélisées.

4.1.3.1. Analyse du signal à débit variable

Dans la mesure où les unités modélisées par des HMM sont des phonèmes dont les durées de réalisation sont très variables, le principe d'une analyse synchrone du signal semble peu adapté. En effet, le débit d'information n'est pas constant dans la parole, et il existe des zones quasi-stationnaires ainsi que des zones fortement transitoires. De plus, le recouvrement entre deux trames successives lors d'une analyse par fenêtre glissante rend abusive l'hypothèse markovienne d'indépendance entre trames. Une analyse du signal à débit variable a donc été proposée.

V. Le Maire, R. André-Obrecht et D. Jouvet utilisent une représentation du signal par segments adjacents [Le Maire *et al.*, 1989]; les segments correspondent à des unités infra-phonémiques obtenues par la méthode des ruptures de modèles [André-Obrecht, 1988]. Les observations sont constituées de coefficients MFCC calculés au centre de chaque segment, de coefficients différentiels aux frontières du segment ainsi que de la durée totale du segment. La comparaison avec l'approche traditionnelle par fenêtre glissante montre que l'approche segmentale permet un décodage plus rapide grâce à la réduction du nombre de trames observées, au prix d'une légère baisse des performances. G. Flammia, P. Dalsgaard, O. Andersen et B. Lindberg segmentent le signal au moyen d'une fonction de variation spectrale [Flammia *et al.*, 1992] afin de ne conserver qu'une partie des trames acoustiques. Le sous-échantillonnage ainsi réalisé sur les trames ne dégrade pas les taux de reconnaissance en identification de mots isolés.

M. Ostendorf et S. Roukos modélisent des observations de durée variable [Ostendorf & Roukos, 1989]. Les trames sont ré-échantillonnées à l'intérieur d'un segment pour obtenir un nombre constant de trames par segment; typiquement, une dizaine de vecteurs cepstraux sont émis par un segment phonétique. La segmentation est réalisée automatiquement au cours du décodage, et les résultats sont proches de ceux obtenus avec une segmentation manuelle *a priori*. J. Marcus réalise une segmentation automatique préalable du signal en "segments phonétiques" [Marcus, 1993]. La qualité de la segmentation est cruciale, puisque l'insertion ou l'omission d'une frontière dégrade immédiatement les taux de DAP; la solution retenue ici est la modélisation de multi-phones pour diminuer les risques d'omission. De plus, cette étude montre que la prise en compte de trames observées à l'extérieur d'un segment pour la représentation de ce segment apporte une meilleure identification phonétique.

F. Brugnara, R. De Mori, D. Giuliani et M. Omologo développent la théorie de modèles de Markov parallèles [Brugnara *et al.*, 1992a], constitués d'un modèle "maître" λ' et d'un modèle "esclave" λ''. Le modèle λ' est un HMM standard, tandis que les probabilités d'émission et de transition du modèle λ'' sont conditionnelles aux émissions de λ'. Ainsi, les observations acoustiques peuvent être engendrées par un modèle "esclave", qui dépend d'un contexte plus général défini au niveau du modèle "maître". Dans une expérience d'identification de nombres isolés, le signal est analysé à un débit variable. Des coefficients MFCC calculés sur la trame sont émis par le modèle "esclave", et la durée de ces trames est émise par le modèle "maître". L'analyse à débit variable puis la séparation en HMM parallèles apportent successivement une réduction de 18% puis de 29% du nombre d'erreurs par rapport à un modèle standard.

N. Suaudeau, R. André-Obrecht & B. Delyon séparent de la même façon dans les observations émises par les modèles un niveau acoustique local et un niveau phonétique global [Suaudeau *et al.*, 1992]. Les observations phonétiques sont indépendantes des émissions acoustiques, et sont modélisées globalement au niveau phonétique. Ce modèle à deux niveaux cherche, comme les HMM parallèles, à traduire dans la structure des HMM la hiérarchie existant en parole. Dans le cadre d'une analyse du signal à débit variable, la modélisation au niveau global de la durée phonétique par une loi inverse gaussienne diminue de 17% les erreurs sur une base de chiffres du CNET.

4.1.3.2. Analyse du signal à débit constant

En restant dans le cadre d'une analyse classique par fenêtre glissante, il est possible d'intégrer une information sur la nature temporelle du signal de parole. Cette information peut être exploitée au niveau de la représentation du signal, lors du décodage, ou en post-traitement.

L'utilisation des coefficients différentiels dans la représentation du signal est une façon maintenant largement répandue de prendre en compte la dynamique du signal dans la modélisation (cf. §1.2). Les représentations à débit variable déjà présentées utilisent souvent la durée du segment comme paramètre supplémentaire. A partir d'une représentation par 12 MFCC, K. Hübener ajoute des coefficients segmentaux qui mesurent la variation spectrale autour de l'instant courant [Hübener, 1995]. En utilisant 3 coefficients segmentaux calculés à différentes échelles temporelles, les taux de reconnaissance sont très proches de ceux obtenus en ajoutant les 12 coefficients cepstraux différentiels. F. Brugnara, R. De Mori, D. Giuliani et M. Omologo ajoutent à des coefficients MFCC statiques et différentiels un coefficient supplémentaire calculé par une fonction de variation spectrale entre trames [Brugnara *et al.*, 1992b]; ils réduisent ainsi de 20% les erreurs lors d'une expérience de reconnaissance de chiffres enchaînés pour un locuteur.

Nous avons proposé un contrôle temporel appliqué au cours du décodage [Barras *et al.*, 1993]. La probabilité de transition entre modèles phonétiques est modifiée en tenant compte d'une pré-segmentation par ruptures de modèles; cette approche est détaillée dans la suite de ce chapitre. C. Mitchell, M.P. Harper et L.H. Jamieson ont suivi une démarche très similaire, en comparant la segmentation manuelle de la base TIMIT avec une segmentation automatique obtenue au moyen d'une fonction de variation spectrale ou d'un changement du delta-cepstre [Mitchell *et al.*, 1995]. La connaissance de la segmentation manuelle de référence au cours du décodage permet de supprimer les omissions et les insertions, les seules erreurs possibles étant des substitutions. Mais les résultats obtenus avec les segmentations automatiques ne sont pas supérieurs à ceux d'un système standard sans segmentation préalable.

Enfin, les chercheurs du CNET réalisent un post-traitement des *n* meilleures phrases proposées par le décodage en tenant compte d'une segmentation *a priori* [Moudenc *et al.*, 1995]. Un nouveau score est calculé pour chaque phrase en considérant deux modèles en parallèle: un modèle estime la probabilité que l'identification du segment soit correcte, et un autre la probabilité qu'elle soit incorrecte. Ces probabilités sont calculées en fonction du nombre de discontinuités détectées par la segmentation *a priori* à l'intérieur du segment phonétique. Cette technique apporte une réduction des erreurs par mots comprise entre 5% et 15% sur des corpus de mots isolés enregistrés sur le réseau téléphonique.

4.1.3.3. Analyse et modélisation de la vitesse d'élocution

La prise en compte de la vitesse d'élocution est habituellement absente des modèles de durée. M. Jones et P. Woodland intègrent un modèle de durée relative comme post-traitement au décodage de la parole continue [Jones & Woodland, 1993]. Plusieurs modèles sont testés; le meilleur modèle consiste à partitionner les phrases d'apprentissage en trois groupes, en fonction de la vitesse d'élocution (rapide, moyenne ou lente). Les *n* meilleures phrases du décodage sont ensuite ré-ordonnées en fonction des probabilités de durée phonétique apprises sur celui des trois groupes correspondant à la vitesse d'élocution de la phrase. Un modèle de durée global à tous les locuteurs, utilisant un histogramme des durées phonétiques, améliore de 6% le taux de décodage lexical sur la base TIMIT par rapport à un décodage standard, et un modèle de durée à débit variable permet une augmentation de 10% du taux initial. Cela montre l'importance du débit d'élocution même pour la reconnaissance d'une base de données enregistrée dans des conditions contrôlées.

Y. Gong et W. Treurniet ont constaté une baisse significative de la durée phonétique en fonction de l'augmentation de la durée des phrases prononcées [Gong & Treurniet, 1993]. Ils proposent un changement d'échelle de la distribution des durées pour les phonèmes en fonction de la durée d'élocution. Ils réduisent ainsi de près de 50% les erreurs d'identification sur des mots isolés lorsque l'apprentissage est réalisé sur de la parole continue.

N. Suaudeau et R. André-Obrecht insistent sur la mauvaise modélisation de la prosodie dans les systèmes de reconnaissance traditionnels à base de HMM, et adaptent les lois de durée en fonction du débit d'élocution [Suaudeau & André-Obrecht, 1994]. La représentation du signal est basée sur une analyse en segments de longueur variable, et la durée des segments est émise par le processus markovien à un niveau indépendant de l'observation acoustique. Les durées phonétiques sont modélisées par une distribution gaussienne inverse. Un débit variable d'élocution est soit intégré lors de la reconnaissance (le débit est alors estimé par un filtrage de Kalman), soit utilisé en post-traitement. La modélisation de la durée entraîne une diminution de 14% des erreurs d'identification de nombres, et la prise en compte du débit d'élocution apporte une réduction supplémentaire de 6%.

4.2. Pré-segmentation du signal de parole

Plutôt que d'imposer un modèle de durée indépendant du signal et basé uniquement sur une connaissance statistique, il nous a semblé intéressant d'essayer d'ancrer le contrôle du séquencement temporel des modèles phonétiques à partir d'une information intrinsèque au signal. Nous avons étudié deux méthodes d'analyse du signal: l'analyse par ruptures de modèles et la décomposition temporelle. L'analyse par rupture de modèles permet la détection de discontinuités dans le signal de parole, qui est considéré comme une succession de zones quasi-stationnaires. La décomposition temporelle décrit le signal comme une combinaison linéaire de "cibles spectrales" pondérées par des fonctions d'interpolation. Dans les deux cas, une segmentation du signal de parole peut être déduite de ces analyses. Nous avons réalisé des expériences d'identification phonétique sur chacune des segmentations obtenues, et proposé une exploitation conjointe de celles-ci. Notre but est l'intégration d'informations temporelles apportées par ces analyses dans un décodage markovien classique pour un meilleur contrôle temporel.

4.2.1. Ruptures de modèles

La méthode des Ruptures de Modèles (RM) permet de détecter des discontinuités du signal de parole [André-Obrecht, 1988]. Dans le cadre de cette modélisation, on considère que le signal $\{s_n\}$ est constitué d'une suite de segments quasi-stationnaires, caractérisés chacun par un modèle auto-régressif d'ordre p:

$$s_n = \sum_{i=1}^{p} a_i . s_{n-i} + e_n \tag{4.7}$$

où $\{e_n\}$ est un bruit blanc gaussien de variance σ^2 constante sur le segment.

Les discontinuités sont détectées par la divergence, à l'instant n, entre un modèle à long terme M_0 calculé sur une fenêtre croissante, et un modèle à court terme M_1 calculé sur une fenêtre glissante. La distance utilisée correspond au calcul, à chaque instant n, de la somme cumulée W_n des entropies mutuelles $\{w_k\}_{1 \le k \le n}$ entre les lois conditionnelles des modèles M_0^k et M_1^k. Cette entropie se calcule par [Basseville & Benveniste, 1983]:

$$w_k = \frac{1}{2}\left\{ 2\frac{e_k^0 e_k^1}{\sigma_1{}^2} - \left[1 + \frac{\sigma_0{}^2}{\sigma_1{}^2}\right]\frac{e_k^0}{\sigma_0{}^2} + \left[1 - \frac{\sigma_0{}^2}{\sigma_1{}^2}\right]\right\} \tag{4.8}$$

où e_k^i est l'erreur de prédiction à l'instant k par le modèle $M_i = (a_1^i \ldots a_p^i, \sigma_i)$:

$$e_k^i = s_k - \sum_{j=1}^{p} a_j^i s_{k-j} \tag{4.9}$$

Le changement de modèle engendre une diminution de la moyenne de w_k. En appliquant la règle de Page-Hinkley, un biais constant δ est ajouté à w_k et la somme cumulée $\hat{W}_n$ est comparée à un seuil λ pour détecter un maximum. Ainsi, une rupture est détectée lorsque:

$$\max_{1 \le k \le n} \hat{W}_k - \hat{W}_n > \lambda \tag{4.10}$$

avec

$$\hat{W}_n = \sum_{k=1}^{n} (w_k + \delta) \tag{4.11}$$

L'instant de la rupture est celui maximisant l'expression (4.10). La statistique étudiée n'est pas symétrique, ce qui entraîne certaines omissions systématiques. Après détection d'une rupture, une deuxième analyse du signal est réalisée dans le sens rétrograde afin d'assurer la détection de certaines ruptures omises. Cette procédure complète conduit à la méthode de "divergence forward-backward" [André-Obrecht, 1988].

L'ordre des modèles, la durée d'initialisation, le biais δ et le seuil λ doivent être choisis avec soin. A la suite d'évaluations sur les phrases phonétiquement équilibrées PEQ du corpus BDSON, un ordre 2 pour les modèles semble suffisant, et la fenêtre d'initialisation est fixée à 20 ms; de plus, il paraît pertinent de distinguer les zones voisées et les zones non voisées pour fixer le biais δ et le seuil λ [André-Obrecht *et al.*, 1992]. Un test de voisement sommaire est réalisé pendant la durée d'initialisation. Les valeurs utilisées sont:

$$\begin{cases} \delta_v = 0{,}2 \\ \lambda_v = 40 \end{cases} \quad \text{pour les zones voisées}$$

et

$$\begin{cases} \delta_b = 0{,}8 \\ \lambda_b = 80 \end{cases} \quad \text{pour les zones non voisées.}$$

La segmentation par RM permet une analyse du signal à débit variable (cf. §4.1.4) [André-Obrecht, 1990; Suaudeau *et al.*, 1992]; elle a été utilisée comme pré-segmentation pour la reconnaissance [Barras *et al.*, 1992] ou l'étiquetage automatique [Parlangeau *et al.*, 1995], et pour évaluer *a posteriori* la qualité d'un décodage [Moudenc *et al.*, 1995]. Des études ont eu pour objet de caractériser les segments quasi-stationnaires ou transitoires, d'étiqueter les frontières en terme de début ou fin de voisement ou de friction [André-Obrecht & Su, 1987], et de détecter les occlusions sourdes [Malbos *et al.*, 1994].

4.2.2. Décomposition temporelle

La Décomposition Temporelle (DT) a été introduite en codage [Atal, 1983], et a eu depuis de nombreuses applications en traitement de la parole: codage, synthèse, segmentation et étiquetage, décodage acoustico-phonétique, représentation des phénomènes articulatoires (cf. par exemple [Montacié, 1991] pour une bibliographie sur le sujet). Elle représente l'évolution spectrale du signal, décrit par une suite de vecteurs spectraux observés $\{o_t\}_{t=1\ldots T}$, par un modèle d'interpolation linéaire. L'estimation d'un spectre o_t est donnée par la combinaison linéaire d'un nombre fini de vecteurs spectraux $\{u_k\}_{1\leq k\leq m}$, appelés cibles spectrales, pondérés par des fonctions d'interpolation à support compact $\{\phi_k(t)\}_{1\leq k\leq m, 1\leq t\leq T}$ qui mesurent l'influence de la cible k à l'instant t:

$$\tilde{o}_t = \sum_{k=1}^{m} u_k \; \phi_k(t) \tag{4.12}$$

L'algorithme de DT estime les cibles spectrales et les fonctions d'interpolation, sous la contrainte d'un critère d'optimalité et en imposant que les fonctions d'interpolation $\phi_k(t)$ soient positives et compactes, c'est-à-dire à support temporel borné. Pour une reconnaissance phonétique, on souhaite par exemple que la cible u_k représente un phonème dans l'espace des paramètres, et $\phi_k(t)$ l'influence de ce phonème dans la trajectoire spectrale; en pratique, des cibles dites de transition sont insérées en raison des non-linéarités lors de la réalisation des phonèmes.

Plus précisément, soient $O=(o_1 \ldots o_T)$ une matrice $d \times T$ constituée des vecteurs d'observation, $U=(u_1 \ldots u_m)$ une matrice $d \times m$ constituée des m cibles spectrales et Φ une matrice $m \times T$ telle que $\Phi_{k,t}=\phi_k(t)$. L'équation (4.12) est équivalente au système matriciel:

$$O = U\,\Phi \tag{4.13}$$

et sa résolution conduit à exprimer les fonctions d'interpolation comme une combinaison linéaire des composantes des vecteurs spectraux:

$$\tilde{\phi}_i(t) = \sum_{j=1}^{d} z_{i,j}\; o_t^j \tag{4.14}$$

où Z est une matrice $m \times d$ obtenue par pseudo-inversion de U. La décomposition temporelle nécessite de définir les coefficients de cette matrice en garantissant la positivité et la compacité des fonctions d'interpolation.

L'algorithme proposé par B. Atal consiste d'abord à rechercher des fonctions d'interpolation centrées sur chacune des trames acoustiques en minimisant une certaine mesure de compacité, à fusionner ces fonctions pour en conserver un nombre réduit, puis à calculer les cibles spectrales minimisant l'erreur quadratique E entre la trajectoire spectrale originale et la trajectoire estimée:

$$E = \sum_{t=1}^{T} \left\| o_t - \sum_{k=1}^{m} u_k\;\phi_k(t) \right\|^2 \tag{4.15}$$

Le choix d'une distance euclidienne n'est cependant pas toujours optimal, et C. Montacié a proposé une généralisation de la DT à une mesure de dissimilarité quelconque [Montacié, 1991]. La DT généralisée a ainsi été appliquée à la distance APS [Caraty & Montacié, 1992].

4.2.3. Études préliminaires

Des expériences préliminaires, réalisant l'identification phonétique à partir d'une segmentation préalable issue de deux méthodes de segmentation par DT et RM, ont permis de mettre en évidence l'intérêt d'une intégration de l'information apportée par différentes méthodes d'analyse du signal à un processus de décodage. Cette étude a été menée en collaboration avec l'ENST et l'IRISA dans le cadre d'un projet GRECO-

Reconnaissance du CNRS [Barras *et al.*, 1992] à la suite d'une première évaluation des deux méthodes [Le Floch, 1991]. La segmentation par la méthode RM a été réalisée par R. André-Obrecht, et P. Deléglise a fourni la segmentation par la méthode DT.

4.2.3.1. Base de données et paramètres

Les expériences ont été effectuées sur les corpus syllabiques SYL de la base de données BDSON, pour les enregistrements de deux locuteurs: un locuteur masculin (identifié BP) et un locuteur féminin (identifiée MD). Les corpus SYL permettent l'étude de 192 diphones consonne-voyelle, constitués par la combinaison de 16 consonnes et de 12 voyelles orales ou nasales. Chacune des 192 phrases contient plusieurs occurrences d'un des diphones. Au total, la base comprend plus de 4000 phonèmes. Les enregistrements des deux locuteurs ont été étiquetés manuellement au LAFORIA sur le principe de l'étiquetage large, c'est-à-dire que l'étiquette est placée au centre de la réalisation du phonème.

Le signal est représenté par des vecteurs de 10 coefficients MFCC, calculés à partir des énergies extraites de 24 filtres triangulaires répartis sur l'échelle de fréquence Mel. Le spectre est calculé par une technique de lissage cepstral itératif [Montacié, 1991], qui estime l'enveloppe à partir des pics stables et améliore la représentation des voix féminines.

4.2.3.2. Évaluation des méthodes de segmentation

L'analyse par la méthode RM sert à définir une segmentation du signal, chaque segment étant constitué de la zone stationnaire du signal comprise entre deux ruptures. Le nombre de segments ainsi obtenu est largement supérieur au nombre de phonèmes identifiés, puisqu'on observe en moyenne 2,2 ruptures par phonème. Cette sur-segmentation est normale, car les ruptures marquent des événements infra-phonémiques comme les explosions des occlusives, les silences, ou les débuts et fin de voisement ou de friction. Une deuxième segmentation a été déduite de l'analyse réalisée par la méthode DT: les supports temporels des fonctions d'interpolation sont choisis comme segmentation du signal. Le taux de sur-segmentation par DT est de 1,6.

Les segmentations par DT et par RM sont évaluées au moyen d'expériences de décodage acoustico-phonétique dépendant du locuteur. Pour chaque locuteur, les segments contenant une étiquette unique de l'étiquetage large sont utilisés comme segments de référence de l'étiquette considérée. Afin d'exploiter au maximum la base de données, les résultats sont calculés par validation croisée, c'est-à-dire que chaque segment est individuellement choisi pour le test tandis que tous les autres constituent l'apprentissage. Le processus d'identification est alors le suivant: chaque segment à identifier est comparé à tous les segments de référence par une comparaison dynamique (DTW) basée sur une distance euclidienne entre les vecteurs MFCC. A la suite d'un vote majoritaire sur les 3 Plus Proches Voisins (PPV), une étiquette est attribuée à ce segment si une décision est possible; sinon il est rejeté. Un alignement dynamique entre la chaîne phonétique de référence et la chaîne obtenue permet de calculer les taux habituels d'identification, substitution, omission et insertion.

4.2.3.3. Exploitation conjointe des segmentations

Les deux méthodes de segmentation sont basées sur des principes différents. Nous avons cherché à exploiter leur complémentarité en faisant l'hypothèse suivante: les segments stationnaires sont bien détectés par les deux méthodes, tandis que l'identification phonétique sur les zones transitoires est susceptible d'être différente entre les deux segmentations. Alors, une nouvelle segmentation est construite par toutes les intersections entre des segments des deux méthodes. Chaque nouveau segment "hérite" des 3 PPV de chacun des deux segments DT et RM dont il est issu. L'étiquetage est finalement réalisé par vote sur ces 6 PPV, en imposant un seuil de rejet égal à 4 (cf. Figure 4.4).

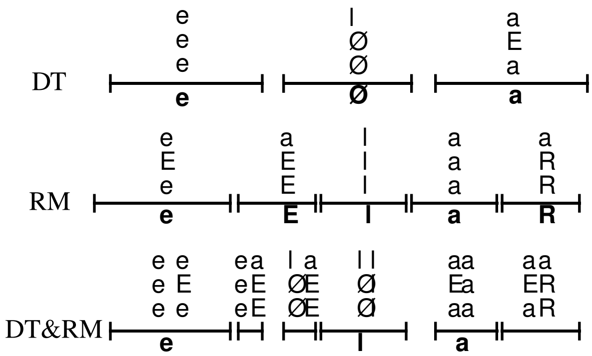

Figure 4.4. Intégration des deux méthodes DT et RM sur le segment de parole [ela].

4.2.3.4. Résultats

Les résultats du décodage sont présentés dans le Tableau 4.1 [Barras *et al.*, 1992]; il s'agit de la moyenne des résultats obtenus individuellement sur chacun des deux locuteurs. Le taux d'identification sur la segmentation par RM est élevé; le fort taux d'insertion s'explique par la sur-segmentation naturelle des ruptures. La segmentation par DT permet d'obtenir des taux d'identification comparables, pour un nombre d'insertions plus faible. Enfin, l'exploitation conjointe de ces deux segmentations permet une réduction efficace du nombre d'insertions. En exploitant la complémentarité de deux approches différentes, une amélioration des taux de reconnaissance devient possible.

Méthode de segmentation	Taux d'identification	Taux de substitution	Taux d'omission	Taux d'insertion
RM	79,6	15,4	4,9	46,1
DT	79,9	12,2	7,8	25,9
RM & DT	71,7	8,0	20,3	6,7

Tableau 4.1. Résultats de décodage acoustico-phonétique sur une pré-segmentation par RM, par DT et par combinaison des deux pour deux locuteurs de la base BDSON.

L'examen détaillé par macro-classe phonétique des taux de reconnaissance obtenus pour la segmentation par RM (cf. Tableau 4.2) révèle que les insertions sont particulièrement élevées pour les occlusives sourdes, les fricatives sonores et les liquides; pour les autres classes le taux moyen d'insertions n'est déjà plus que de 20% environ. Un traitement adapté au type d'événement détecté doit permettre une exploitation plus efficace des ruptures dans un système de décodage.

Macro-classe phonétique	Taux d'identification	Taux de substitution	Taux d'omission	Taux d'insertion
Occlusives sourdes	76 %	21 %	3 %	166 %
Occlusives voisées	72 %	24 %	4 %	32 %
Fricatives sourdes	97 %	2 %	0 %	7 %
Fricatives voisées	89 %	7 %	3 %	119 %
Nasales	79 %	17 %	4 %	27 %
Liquides	69 %	18 %	13 %	65 %
Voyelles	79 %	16 %	5 %	21 %
Voyelles nasales	89 %	9 %	3 %	12 %

Tableau 4.2. Détail des taux de DAP par macro-classe phonétique pour la pré-segmentation par ruptures de modèles.

4.2.4. Conclusions

Les expériences préliminaires réalisées avec les deux méthodes de segmentation DT et RM ont montré la possibilité de combiner des informations de différentes sources lors du décodage. Dans le cadre de la reconnaissance de la parole par HMM, la segmentation préalable du signal acoustique n'est pas nécessaire, ce qui évite les problèmes rencontrés par l'approche analytique. Pourtant, les méthodes d'analyse utilisées dans notre étude mettent en évidence des informations présentes dans le signal de parole qui ne sont pas obligatoirement détectées par le décodage markovien. Nous avons donc développé une approche originale au contrôle temporel, en intégrant certaines informations issues d'une analyse du signal au cours du décodage markovien.

Nous avons décidé de continuer cette étude avec la méthode d'analyse par RM. En effet, nous pensons que la précision temporelle de l'analyse par RM est complémentaire de l'analyse spectrale centi-seconde réalisée pour le décodage par HMM et peut être exploitée pour aider le processus de décodage; à l'inverse, la méthode DT utilise une représentation spectrale classique. De plus, les discontinuités détectées par la méthode RM peuvent dans de nombreux cas être interprétées comme résultant d'instabilités infra-phonémiques, comme des débuts ou fins de voisement, de friction, ou encore comme des explosions.

4.3. Décodage markovien avec transitions contraintes

L'utilisation d'une pré-segmentation peut sembler inutile dans le cadre d'un système markovien de reconnaissance. Pourtant, des évaluations montrent que la segmentation produite par le décodage n'est pas optimale. La comparaison avec une segmentation obtenue par la méthode des ruptures de modèles nous conduit à proposer une approche originale pour le contrôle temporel: lors du décodage de la parole continue, les probabilités de transition inter-modèles sont modifiées par une fonction de pénalité dépendant de la proximité des ruptures. Cette fonction favorise les transitions phonétiques dans le voisinage des ruptures, et les rend moins probable loin des ruptures, en tenant compte de statistiques réalisées sur le signal. Ainsi, l'information événementielle apportée par les ruptures est intégrée dans le processus de décodage, tout en restant dans le cadre classique de la recherche d'un chemin optimal, et sans introduire de coût de calcul supplémentaire.

4.3.1. Evaluation des segmentations

La segmentation issue du décodage par le système de référence et la segmentation automatique par ruptures de modèles ont été comparées à l'étiquetage de référence de la base TIMIT. Ces évaluations mettent en évidence les spécificités des deux approches et nous permettent de proposer une démarche exploitant leur complémentarité.

4.3.1.1. Critères d'évaluation

L'évaluation quantitative de la qualité d'une segmentation automatique est réalisée par comparaison avec l'étiquetage de référence de TIMIT. Deux critères sont considérés classiquement [André-Obrecht *et al.*, 1992]:

- Le taux de sur-segmentation TSS, qui est le rapport du nombre de segments proposés par la méthode sur le nombre d'étiquettes de référence;
- Le coefficient de qualité CQ_ε, qui indique le pourcentage de frontières de référence détectées à moins de ε ms par la segmentation automatique.

Il nous a semblé pertinent, lors de l'étude des segmentations, de préciser ces taux en fonction du contexte phonétique de l'étiquetage de référence. Le coefficient $CQ_\varepsilon(\varphi_1,\varphi_2)$ peut être calculé pour les transitions phonétiques entre φ_1 et φ_2. Le TSS est plus difficile à définir au niveau d'un segment phonétique, puisque les frontières ne sont pas toujours détectées. Nous avons donc utilisé un critère supplémentaire:

- Le taux de segments insérés TSI_ε, qui est le nombre moyen de transitions observées à l'intérieur des étiquettes de référence, en excluant celles qui correspondent à la détection d'une frontière à moins de ε ms.

A la différence du taux de sur-segmentation, le taux de segments insérés dépend aussi de la qualité de la segmentation. Les critères ainsi définis nous permettent d'évaluer la segmentation issue du décodage et une segmentation automatique par ruptures de modèles.

4.3.1.2. Segmentation issue du décodage

Le décodage de la parole continue par un système de reconnaissance à base de HMM au moyen de l'algorithme de Viterbi produit naturellement une segmentation des phrases décodées (cf. §1.4.4). Cependant, les systèmes de reconnaissance sont évalués exclusivement en fonction de la suite phonétique obtenue. L'étude de la segmentation issue du décodage peut être un moyen de mieux isoler les faiblesses du système, et éventuellement d'y remédier.

Le système de référence présenté au chapitre 2 est ici étudié d'un point de vue exclusivement segmental, sans tenir compte de l'identification des segments. Les valeurs des critères pour différentes valeurs de ε sont données dans le tableau 4.3, pour la segmentation résultant du décodage de l'ensemble de test complet, qui contient plus de 50.000 segments phonétiques.

ε	CQ	TSI	TSS
10 ms	53 %	44 %	0,97
20 ms	77 %	20 %	
30 ms	84 %	13 %	

Tableau 4.3. Critère de qualité, taux de segments insérés et taux de sur-segmentation pour le système de reconnaissance de référence.

Bien que le taux de sur-segmentation soit proche de 1, la proportion de transitions mal positionnées est élevée. La fréquence des insertions dépend de la durée phonétique, puisque les segments les plus longs (silences, diphtongues) sont aussi ceux qui présentent le plus d'insertions (cf. Tableau 4.4). Le quart des voyelles et la moitié des diphtongues sont ainsi mal segmentées. On peut supposer qu'une partie des erreurs de décodage seraient corrigées par une meilleure localisation des frontières.

Classe phonétique	Nombre	Durée (cs)	TSI(φ)
Silences	3385	17,4	57 %
Occlusions	6853	5,7	18 %
Occlusives sourdes	4455	5,2	12 %
Occlusives voisées	2813	2,4	3 %
Fricatives	7271	9,0	11 %
Nasales	4434	5,8	9 %
Liquides	5329	6,3	16 %
Voyelles	14149	8,5	24 %
Diphtongues	2431	14,6	47 %

Tableau 4.4. Détail par macro-classe phonétique du nombre de segments de test, de leur durée moyenne, et du taux de segments insérés au décodage (ε=20ms).

Le tableau 4.5. détaille la qualité de la reconnaissance des transitions phonétiques $CQ_{\varepsilon}(\varphi_1,\varphi_2)$ avec un seuil de 20 ms pour 8 macro-classes; les taux sont indiqués uniquement lorsque les transitions sont représentées par plus de 100 occurrences dans l'ensemble de test. De grandes disparités existent: en particulier, les transitions vers un silence sont très peu détectées, ainsi que les transitions entre liquides et voyelles ou diphtongues. Cela peut être dû à un déplacement de la frontière supérieur à 20 ms, ou à l'omission de cette frontière. A l'inverse, les occlusions, fricatives ou nasales situées après une voyelle ou une liquide sont très bien détectées. Cette évaluation suppose une segmentation de référence idéale; ce n'est bien sûr pas le cas, et il peut arriver que les erreurs proviennent d'une erreur de l'étiquetage de référence, ou d'un contexte acoustique ambigu.

$CQ_{20}(\varphi_1,\varphi_2)$	S	O	O S	O V	F	N	L	V&D
Silences	-	-	76	73	51	90	70	87
Occlusions	33	-	90	81	85	65	55	-
Occlusives Sourdes	51	-	-	-	60	-	89	78
Occlusives Voisées	-	-	-	-	-	-	81	84
Fricatives	59	91	-	-	60	90	91	92
Nasales	27	77	69	-	81	-	60	84
Liquides	8	93	-	-	90	90	58	44
Voyelles & Diphtongues	16	95	37	85	87	86	51	41

Tableau 4.5. Taux de détection à 20 ms des transitions entre phonèmes lors du décodage de la base de test complète par le système de référence (en ligne: phonème précédant la transition; en colonne: phonème suivant).

4.3.1.3. Segmentation par ruptures de modèles

Une segmentation automatique des phrases de la base TIMIT a été réalisée au moyen de la méthode RM. L'ordre des modèles reste fixé à 2, et la fenêtre d'initialisation à 20 ms. Par contre, le test de voisement développé pour la base BDSON n'est plus applicable à la base TIMIT. Nous avons choisi d'analyser toutes les phrases avec les valeurs de biais et de seuil spécifiques des zones voisées (δ_v=0,2 et λ_v=40), d'où une détection plus fréquente de ruptures dans les zones bruitées. Le taux de sur-segmentation est élevé (1,53), mais les transitions sont bien détectées: 86% des transitions sont détectées à moins de 20ms de l'étiquetage de référence, contre seulement 77% avec la segmentation issue du décodage (cf. Tableau 4.6). Il semble donc que les ruptures mettent en évidence une information qui n'est pas correctement exploitée par un système standard de reconnaissance.

ε	CQ	TSI	TSS
10 ms	67 %	85 %	1,53
20 ms	86 %	66 %	
30 ms	89 %	63 %	

Tableau 4.6. Critère de qualité, taux de segments insérés et taux de sur-segmentation pour la segmentation par RM.

L'identification des transitions phonétiques par la méthode RM, évaluée sur l'ensemble d'apprentissage complet avec un seuil ε=20 ms (cf. Tableau 4.7), révèle par exemple que la localisation des débuts de silences est nettement supérieure à celle obtenue par le décodage (cf. Tableau 4.5).

$CQ_{20}(\varphi_1,\varphi_2)$	S	O	O S	O V	F	N	L	V&D
Silences	-	-	99	98	95	91	95	98
Occlusions	58	-	97	92	96	80	84	93
Occlusives Sourdes	93	95	-	-	68	83	91	86
Occlusives Voisées	91	-	-	-	66	-	70	62
Fricatives	98	99	99	-	90	95	93	93
Nasales	78	86	80	81	89	23	56	89
Liquides	92	99	84	85	98	90	67	59
Voyelles & Diphtongues	91	99	84	86	98	93	54	40

Tableau 4.7. Taux de détection à 20 ms des transitions entre phonèmes de l'ensemble d'apprentissage par la méthode RM (en ligne: phonème précédant la transition; en colonne: phonème suivant).

Le taux de sur-segmentation élevé complique l'évaluation objective de la segmentation: avec un grand nombre de ruptures, la probabilité de trouver une transition à moins de 20 ms est nécessairement plus forte. Pour montrer la supériorité de cette segmentation sur une segmentation aléatoire, nous avons réalisé une statistique supplémentaire. Nous avons calculé la probabilité qu'une transition phonétique soit observée dans la trame courante, en fonction de la distance de la rupture la plus proche. Il apparaît que la probabilité de transition dépend très fortement de la distance de la rupture la plus proche (cf. Figure 4.5); à l'inverse, une segmentation aléatoire de même TSS produirait une distribution uniforme. Ainsi, la probabilité qu'une rupture dans la trame courante signale effectivement une transition est de plus de 25%, tandis que cette probabilité descend à environ 5% lorsque la rupture la plus proche est distante d'au moins deux trames.

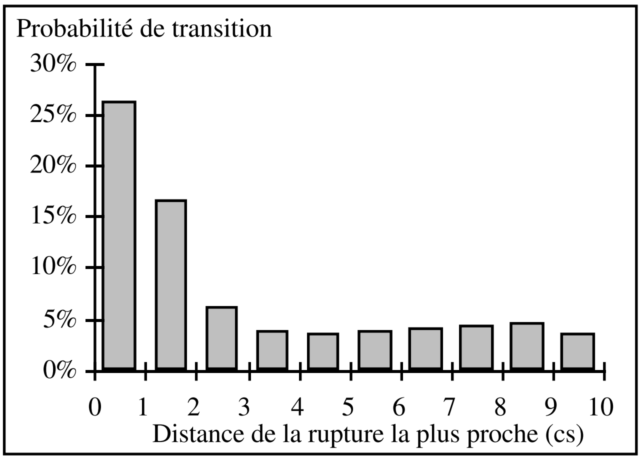

Figure 4.5. Probabilité qu'une transition soit observée dans la trame courante en fonction de la distance de la rupture la plus proche.

4.3.2. Principe des transitions contraintes

A la suite des études préliminaires présentées, nous pensons que les événements détectés par la technique RM apportent une information qui n'est pas toujours exploitée par le système de décodage. Cependant, le principe d'une pré-segmentation du signal suivie d'une identification de ces segments a montré ses limites; en effet, des omissions ou insertions de frontières sont inévitables et ces erreurs sont immédiatement répercutées sur les erreurs de décodage. L'une des supériorité des méthodes de reconnaissance globales sur les méthodes analytiques est justement d'éviter une segmentation éventuellement erronée préalable à l'identification (cf. §1.1.3).

Nous avons choisi de rester dans le cadre d'un décodage classique de la parole continue par HMM, consistant à rechercher le meilleur enchaînement de modèles en produisant simultanément une segmentation du signal. Les ruptures détectées par la technique RM dans les phrases de test sont interprétées dans le système de décodage comme des instants de transitions plus probables entre les modèles représentant les unités phonétiques (cf. Figure 4.6).

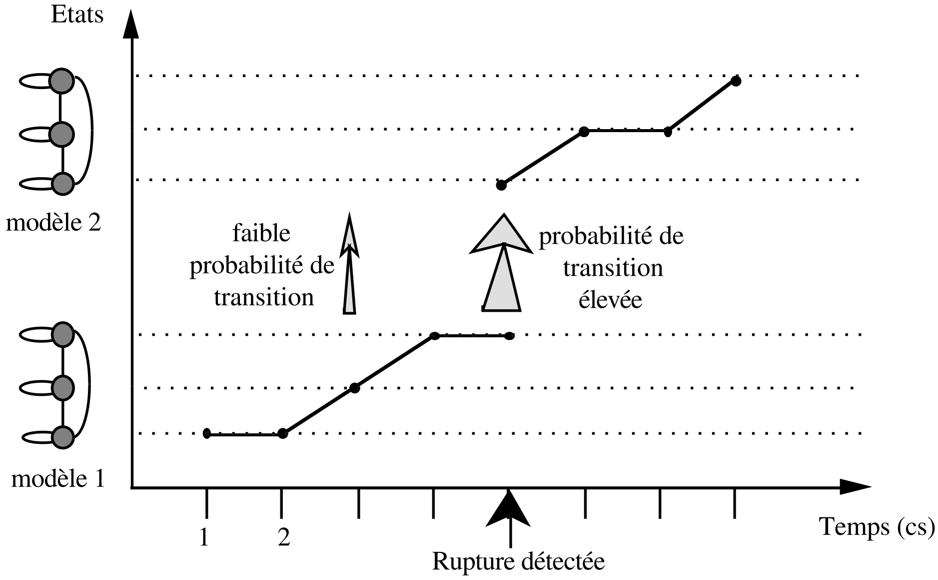

Figure 4.6. Transitions inter-modèles contraintes au cours du décodage phonétique par la proximité des ruptures.

4.3.2.1. Définition d'une pénalité inter-modèles

Lors de l'étape d'identification utilisant l'algorithme de Viterbi, la probabilité $P(\lambda_j/\lambda_i)$ de transition entre les modèles λ_i et λ_j est modifiée par un facteur multiplicatif:

$$\tilde{P}(\lambda_j|\lambda_i) = P(\lambda_j|\lambda_i)\, e^{F(t)} \tag{4.16}$$

où $F(t)$ dépend de la distance de la rupture la plus proche. Nous avons choisi une simple fonction créneau, qui applique une pénalité F_0 lorsque la trame courante ne contient pas de rupture (cf. Figure 4.7). Une valeur négative de F_0 diminue la probabilité de transition lorsqu'aucune rupture n'est observée dans la trame courante; on espère de cette manière réduire les insertions. De manière analogue, il serait possible d'augmenter la probabilité de transition pour les trames qui contiennent une rupture, dans le but de forcer une transition phonétique à cet instant et éventuellement de diminuer les omissions phonétiques; par défaut, nous n'appliquons pas de pénalité à ces trames (i.e. $F(t)$=0).

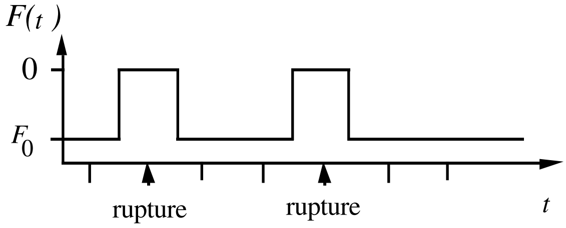

Figure 4.7. Pénalité appliquée à la probabilité de transition inter-modèles en fonction de la distance des ruptures.

Il est important de remarquer que la pénalité qui est appliquée ne dépend que de l'instant courant, et non du passé du processus dans le modèle. L'algorithme de Viterbi reste donc parfaitement optimal pour la recherche de la meilleure suite d'états en tenant compte de la contrainte introduite. Il n'y a aucun surcoût de calcul lors du décodage, en dehors de la détection des ruptures dans le signal décodé.

4.3.2.2. Choix des pénalités

La valeur de la pénalité F_0 doit dépendre de la qualité de la pré-segmentation utilisée. Notons R l'événement "il existe une rupture dans la trame courante" et T l'événement "une transition est observée dans la trame courante"; de plus, la durée moyenne des segments phonétiques en nombre de trames est notée $\overline{\tau}$. Alors, la probabilité *a priori* d'une transition phonétique au cours de la trame est:

$$P(T) = 1/\overline{\tau} \tag{4.17}$$

La proportion de ruptures signalant une transition peut être exprimée en fonction des critères CQ et TSS déjà définis comme:

$$P(T/R) = P(R/T)\frac{P(T)}{P(R)} = \frac{CQ}{TSS} \tag{4.18}$$

puisque CQ= $P(R/T)$ désigne la proportion de transitions associées à une rupture et TSS= $P(R)/P(T)$ le rapport entre le nombre de ruptures et le nombre de transition. De la même façon, la probabilité en l'absence de rupture est:

$$\begin{aligned} P(T/\overline{R}) &= \frac{P(T) - P(T,R)}{1 - P(R)} = \frac{1 - P(R/T)}{1/P(T) - P(R)/P(T)} \\ &= \frac{1 - CQ}{\overline{\tau} - TSS} \end{aligned} \tag{4.19}$$

Alors, la pénalité de transition peut être fixée à:

$$F_0 = \ln P(T/\overline{R}) - \ln P(T/R) \tag{4.20}$$

Soit, dans le cas de la segmentation par ruptures de modèles présentée au §4.3.1.3 et pour une tolérance ε=20ms sur la détection des transitions, les valeurs numériques suivantes:

$$\begin{cases} CQ = 0{,}86 \\ TSS = 1{,}53 \\ \overline{\tau} = 7{,}9 \end{cases} \Rightarrow \begin{cases} P(T/R) = 56\% \\ P(T/\overline{R}) = 2\% \end{cases} \Rightarrow F_0 = -3{,}3 \tag{4.21}$$

En pratique, les problèmes d'échelle entre les probabilités d'émission acoustiques et les probabilités du modèle de langage nous incitent à tester différentes valeurs de pénalité afin de trouver un optimum.

4.3.3. Expériences

Nous avons appliqué le contrôle temporel par transitions contraintes à des expériences de décodage acoustico-phonétique, puis à une expérience pilote de décodage lexical. Cette dernière expérience permet de nous rapprocher des conditions réelles d'utilisation d'un système de dictée automatique.

4.3.3.1. Décodage acoustico-phonétique contraint

Un décodage par transitions contraintes a été réalisé sur le noyau de test avec le système de référence. Les expériences sont d'abord menées en conservant la valeur par défaut pour le facteur d'échelle ($s = 1$). En effet, le facteur d'échelle permet de modifier de manière globale les probabilités de transition pour réduire les insertions. Son rôle est donc éventuellement concurrent de celui du contrôle temporel par transitions contraintes, et la valeur retenue pour le facteur d'échelle dans le système de référence ($s = 3$) n'est plus nécessairement adaptée aux nouvelles conditions de test.

Avec l'introduction d'une pénalité loin des ruptures, une réduction des insertions est effectivement observée au détriment d'un progression des omissions (cf. Tableau 4.8). Le taux de reconnaissance phonétique le plus élevé est obtenu pour une pénalité de -6, ce qui correspond à des taux d'insertions et d'omissions comparables, de l'ordre de 7%; cet optimum est différent de celui calculé (cf. §4.3.2.2). Au total, une réduction de 1,5% des erreurs est apportée par le contrôle temporel que nous proposons.

Pénalité F_0	Phonèmes identifiés (%)	Phonèmes substitués (%)	Phonèmes omis (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
Transitions non contraintes					
0	74,8	20,3	4,9	9,7	**65,1**
Modification de la pénalité F_0					
-2	74,1	20,1	5,8	8,2	**65,8**
-4	73,7	19,9	6,4	7,3	**66,3**
-6	73,2	19,8	7,0	6,6	**66,6**
-8	72,5	20,1	7,5	6,1	**66,4**

Tableau 4.8. Taux de décodage du noyau de test avec le système de référence par transitions contraintes pour différentes valeurs de pénalité (facteur d'échelle s=1).

Cependant, nous avons obtenu nos meilleurs résultats de DAP en augmentant l'influence du modèle de langage. Un facteur d'échelle s=3 permet une réduction importante des insertions et une amélioration du taux de reconnaissance. Dans de telles conditions, l'apport des transitions contraintes par les ruptures semble marginal (cf. Tableau 4.9), puisqu'il n'améliore plus que de 0,1% la reconnaissance. Il semble donc que les deux traitements, bien que de nature très différente, aboutissent à un contrôle concurrent des insertions.

Pénalité F_0	Phonèmes identifiés (%)	Phonèmes substitués (%)	Phonèmes omis (%)	Phonèmes insérés (%)	Phonèmes reconnus (%)
Transitions non contraintes					
0	74,1	18,7	7,2	5,3	**68,8**
Modification de la pénalité F_0					
-2	73,5	18,8	7,7	4,6	**68,9**
-4	73,0	18,5	8,5	4,2	**68,8**
-6	72,0	18,8	9,2	3,8	**68,2**

Tableau 4.9. Taux de décodage du noyau de test avec le système de référence par transitions contraintes pour différentes valeurs de pénalité (facteur d'échelle s=3).

Ces expériences montrent que l'influence du contrôle temporel par transitions contraintes est très dépendante de la pondération appliquée au bigramme phonétique utilisé relativement au modèle acoustique. Or, le réglage du facteur d'échelle reste très empirique; son augmentation a pour effet de diminuer les insertions en se rapprochant d'une distribution phonétique proche de celle observée dans l'ensemble d'apprentissage. L'amélioration des taux de reconnaissance par ce moyen ne traduit en rien une amélioration de la modélisation acoustique et nous paraît assez artificielle. De plus, un système de reconnaissance dans des conditions réelles d'exploitation fonctionne différemment, et exploite un modèle de langage plus complexe qu'un simple bigramme phonétique; l'optimisation d'un facteur d'échelle phonétique devient alors sans objet. A l'inverse, le contrôle temporel que nous proposons est indépendant du modèle de langage, et tient compte du signal acoustique décodé. Il peut donc être exploité dans un système réel. Pour mettre en évidence sa véritable influence, nous avons donc décidé de réaliser une expérience de décodage lexical et de contraindre le séquencement des modèles phonétiques à l'intérieur des mots par les ruptures de modèles.

4.3.3.2. Décodage lexical contraint

Toutes les expériences de reconnaissance présentées jusqu'ici sont des expériences de DAP. La base TIMIT est particulièrement adaptée à l'identification phonétique. Il est néanmoins possible de réaliser aussi une reconnaissance par mots, en utilisant la transcription phonétique normative des 6100 mots utilisés dans la base de TIMIT qui est distribuée avec cette base. Nous définissons un modèle de langage lexical par paires de mots; ce modèle de langage autorise la production de phrases $M=m_1m_2m_3\ldots$ telles que les paires de mots m_im_{i+1} ont été observées dans une des phrases d'apprentissage ou de test de la base TIMIT. Chaque mot est décrit comme une succession de modèles phonétiques dépendants du contexte droit; le dernier phonème du mot est modélisé par un modèle indépendant du contexte puisque le premier phonème du mot suivant n'est pas connu avant le décodage. Le décodage produit une suite de mots qui est comparée avec la suite lexicale de référence, et les taux de reconnaissance sont calculés en mots et non plus en phonèmes. Les 192 phrases du noyau de test contiennent 1570 mots, et les 1152 phrases de l'ensemble de test n'appartenant pas au noyau de test contiennent 9455 mots.

Le contrôle temporel est appliqué au cours du décodage lexical de la même manière que pour le DAP. Une pénalité est ajoutée à la probabilité d'un chemin lorsqu'une transition phonétique à l'intérieur d'un mot ou entre deux mots se produit sans que la trame courante contienne de rupture. Pour une pénalité de -4, le contrôle temporel réduit les insertions de mots, de presque 12% à moins de 9%, tout en augmentant de 2% le taux d'identification lexical (cf. Tableau 4.10), pour un gain global de 5% du taux de reconnaissance. Le rayon de l'intervalle de confiance à 95% est de 2%, et ces résultats sont significatifs.

Pénalité F_0	% mot identifiés	% mots substitués	% mots omis	% mots insérés	% mots reconnus
Transitions non contraintes					
0	85,7	13,4	0,9	12,2	**73,5**
Transitions contraintes					
-2	86,4	12,8	0,8	10,9	**75,5**
-4	87,6	11,7	0,7	9,0	**78,6**
-6	87,3	12,0	0,7	9,0	**78,2**
-8	87,2	12,2	0,6	8,9	**78,3**

Tableau 4.10. Taux de décodage lexical du noyau de test avec les modèles phonétiques du système de référence. Influence des transitions contraintes.

Finalement, nous avons réalisé le décodage lexical de l'ensemble de test complet, en excluant les 192 phrases du noyau de test déjà utilisées pour la validation du système de référence et le choix de la pénalité de transition (cf. Tableau 4.11). Le gain global est alors de 2,1% en valeur absolue, mais ces résultats restent significatifs car le rayon de l'intervalle de confiance pour cette expérience portant sur 9455 mots est de 0,8%.

Pénalité F_0	% mot identifiés	% mots substitués	% mots omis	% mots insérés	% mots reconnus
Transitions non contraintes					
0	86,5	12,7	0,9	11,1	**75,3**
Transitions contraintes					
-4	86,9	12,2	0,9	9,5	**77,4**

Tableau 4.11. Décodage lexical avec transitions contraintes de l'ensemble de test complet avec les modèles phonétiques du système de référence.

Il est certain que cette expérience est préliminaire: la prise en compte du contexte phonétique aux frontières des mots serait souhaitable, et une transcription phonétique des mots basée sur les prononciations réelles des locuteurs plutôt que sur une seule transcription normative semble indispensable; enfin, des tests avec un modèle de langage plus complexe (bigramme et trigramme lexical) devront être réalisés. Mais les résultats obtenus montrent que le contrôle temporel au niveau phonétique améliore le décodage lexical.

4.4. Conclusions

La modélisation de la parole par des HMM présente des faiblesses, en particulier en ce qui concerne une loi de durée phonétique vraisemblable. Pourtant, les tentatives pour améliorer cette modélisation ont généralement abouti à une forte complexification du système, pour des gains en reconnaissance parfois non significatifs. Nous avons proposé une approche différente, basée sur une analyse statistique fine du signal au cours du décodage. Un contrôle temporel prenant en compte les ruptures de modèles pour contraindre les transitions phonétiques est possible sans augmenter la complexité de l'apprentissage ou du décodage.

Des expériences de décodage acoustico-phonétique ont montré que le contrôle temporel ainsi réalisé permet de réduire les insertions phonétiques; cependant, une modification apportée à la pondération du bigramme phonétique peut produire elle aussi une telle réduction. Une expérience de décodage lexical a finalement été réalisée; elle apporte la confirmation que le contrôle temporel par transitions contraintes réduit aussi les insertions de mots et améliore la reconnaissance dans des conditions plus proches d'une application réelle. Dans l'avenir, d'autres méthodes d'analyse du signal pourraient être intégrées de manière analogue aux ruptures, et des méthodes plus classiques de contrôle temporel par une loi de durée *a priori* devraient être testées pour comparer leurs performances à celles du contrôle temporel développé.

La prise en compte de sources d'informations spécialisées dans un processus de décodage par modèle de Markov cachés est ainsi possible, et peut améliorer la chaîne de reconnaissance. Cette intégration reste délicate, car la structure des HMM permet facilement de rajouter de nouveaux coefficients aux observations ou des états aux modèles, mais plus difficilement de gérer des informations sur une échelle des temps différente de celle des trames acoustiques. De nombreuses études restent nécessaires pour adapter les systèmes de reconnaissance aux nombreux aspects de la parole continue.

CONCLUSIONS

La reconnaissance de la parole continue est actuellement traitée par une modélisation markovienne. Celle-ci est fondée sur une formalisation rigoureuse, mais son application à la parole nécessite malheureusement des hypothèses abusives. Nous pensons que des traitements adaptés doivent être introduits pour une meilleure modélisation de la parole. Nous nous sommes intéressés en particulier à deux points faibles des systèmes de RAP classiques: l'adaptation au locuteur et le contrôle temporel.

La très forte variabilité inter-locuteurs diminue la discrimination acoustique de modèles indépendants du locuteur, et nous avons défini une topologie des locuteurs pour adapter l'ensemble d'apprentissage au locuteur de test. Les locuteurs d'apprentissage sont classés au moyen d'une distance inter-locuteurs initialement proposée pour la reconnaissance du locuteur. Pour chaque locuteur de test, une combinaison pondérée des classes d'apprentissage permet d'obtenir des modèles adaptés à ce locuteur. Il s'agit d'une adaptation rapide qui généralise le principe des multi-modèles. Les résultats confirment la pertinence acoustique de la topologie pour la reconnaissance, et montrent l'intérêt d'une réduction intelligente de l'ensemble d'apprentissage.

La prise en compte de la durée est généralement insuffisante dans le cadre d'une modélisation classique, et nous avons proposé de contraindre l'enchaînement des modèles phonétiques au moyen d'une étude statistique préalable du signal. Une analyse par ruptures de modèles détecte des frontières entre des zones acoustiques stationnaires et transitoires. La probabilité de transition inter-modèles est alors modifiée au cours du décodage en fonction de la position des ruptures. Ce contrôle temporel apporte une réduction des insertions dans le cadre d'expériences de décodage phonétique et de décodage lexical.

Afin d'étudier ces deux points et pour montrer la pertinence des idées proposées, nous avons développé un système de référence à base de modèles de Markov cachés pour la reconnaissance de la parole continue. Ce système est constitué de modèles phonétiques dépendants du contexte, et il est représentatif des meilleures techniques actuellement disponibles. Une procédure automatique originale a été proposée pour améliorer la représentation des densités de probabilité et réduire le nombre de paramètres du système. Nous avons cherché à conserver une certaine simplicité au système qui n'est donc pas optimal, mais ses performances le classent honorablement par rapport aux systèmes développés par d'autres équipes de recherche. De plus, sa réalisation avec la plate-forme logicielle HTK permet une réelle reproductibilité.

Cependant, nous n'avons pas réalisé une intégration simultanée de toutes les améliorations proposées. Une autre limitation provient de l'évaluation du système principalement par des expériences de décodage acoustico-phonétique. L'objectif poursuivi est la constitution d'une dictée vocale, et la mise au point d'un modèle de langage plus réaliste est indispensable pour atteindre cet objectif. Le module acoustique décrit ne prendra tout son sens qu'avec cette intégration, et par des évaluations sur des bases de données plus variées.

En conclusion, nous avons constaté que la reconnaissance de la parole continue reste une tâche difficile. Un grand nombre de voies de recherche restent ouvertes pour traiter tous les aspects de la communication homme-machine [Cole *et al.*, 1995], et il est permis de se demander avec H. Bourlard si les progrès majeurs à venir de la discipline ne passent pas par une remise en cause profonde des paradigmes dominants, éventuellement au prix d'une réduction temporaire des performances des systèmes [Bourlard, 1995]…

BIBLIOGRAPHIE

R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, no. 1, pp. 29-40, 1988.

R. André-Obrecht, "Reconnaissance automatique de parole à partir de segments acoustiques et de modèles de Markov cachés", *Actes des 18èmes JEP*, pp. 212-216, Montréal, 1990.

R. André-Obrecht, *Segmentation et parole?*, Habilitation à diriger des recherches, Rennes, 1993.

R. André-Obrecht & H.Y. Su, "Expériences en vue du décodage acoustico-phonétique à partir d'une recherche statistique d'événements articulatoires et d'un codage vectoriel", *Actes des 16èmes JEP*, pp. 64-67, Hammamet, 1987.

R. André-Obrecht, G. Pérennou & N. Vigouroux, "Deux approches de l'étiquetage en événements phonétiques", *Actes des 19èmes JEP*, pp. 279-284, Bruxelles, 1992.

B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition", *Proc. ICASSP*, pp. 81-84, 1983.

B.S. Atal & S.L. Hanauer, "Speech analysis and synthesis by linear prediction of speech wave", *J. Acoust. Soc. Am.*, vol. 50, no. 2, pp. 637-655, 1971.

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

L.R. Bahl, P.F. Brown, P.V. de Souza & R.L. Mercer, "Maximum mutual information estimation in hidden Markov model parameters for speech recognition", *Proc. ICASSP*, pp. 49-52, Tokyo, 1986.

L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer & M.A. Picheny, "Acoustic Markov models used in the Tangora speech recognition system", *Proc. ICASSP*, pp. 497-500, New-York, 1988.

L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo & M.A. Picheny, "Decision trees for phonological rules in continuous speech", *Proc. ICASSP*, pp. 185-188, Toronto, 1991.

L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer & M.A. Picheny, "A method for the construction of acoustic Markov models for words", *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 4, pp. 443-452, 1993.

J.K. Baker, "The DRAGON system - An overview", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 1, pp. 24-29, 1975.

J.K. Baker, *Assessment of using a reference system for the evaluation of speech recognizers and data bases*, Report to National Bureau of Standards, 1982.

R. Bakis, "Continuous speech recognition via centisecond acoustic states", *91th. Meeting of the Acoust. Soc. Am.*, avril 1976.

C. Barras, M.-J. Caraty, P. Deléglise, C. Montacié, R. André-Obrecht et X. Rodet, "Décomposition temporelle et ruptures de modèles pour le décodage acoustico-phonétique", *Actes des 19èmes JEP*, pp. 335-340, Bruxelles, 1992.

C. Barras, M.-J. Caraty & C. Montacié, "HMM based acoustic-phonetic decoding with constrained transitions and speaker topology", *NATO ASI Workshop*, Bubion, 1993. (in *Speech Recognition and Coding. New Advances and Trends*, édité par A. Rubio Ayuso et J. López Soler, NATO ASI Series, vol. F 147, Springer-Verlag, pp. 72-75, 1995).

C. Barras, M.-J. Caraty & C. Montacié, "Contrôle temporel et sélection de l'apprentissage appliqués aux modèles de Markov cachés", *Actes du 9ème Congrès RFIA*, pp. 391-396, Paris, 1994(a).

C. Barras, M.-J. Caraty & C. Montacié, "Topologie des locuteurs et présegmentation pour le décodage acoustico-phonétique", *Actes des 20ème JEP*, pp. 491-495, Trégastel, 1994(b).

C. Barras, M.-J. Caraty & C. Montacié, "Temporal control and training selection for HMM based system", *Proc. Eurospeech*, pp. 27-30, Madrid, 1995.

M. Basseville & A. Benveniste, "Sequential detection of abrupt changes in spectral characteristics of digital signals", *IEEE Trans. Inform. Theory*, vol. IT-29, no. 5, pp. 709-724, 1983.

L. Baum, "An inequality and association maximization technique in statistical estimation for probabilistic functions of Markov processes", *Inequality*, vol. 3, 1972.

V. Beattie, S. Edmondson, D. Miller, Y. Patel & G. Talvola, "An integrated multi-dialect speech recognition system with optional speaker adaptation", *Proc. Eurospeech*, pp. 1123-1126, Madrid, 1995.

J.R. Bellegarda & D. Nahamoo, "Tied mixture continuous parameter modeling for speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-38, no. 12, pp. 2033-2045, 1990.

J.R. Bellegarda, P.V. de Suza, A. Nadas, D. Nahamoo, M.A. Picheny & L.R. Bahl, "The metamorphic algorithm: a speaker mapping approach to data augmentation", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 3, pp. 413-420, 1994.

R.E. Bellman, *Dynamic Programming*, Princeton Univ. Press, 1957.

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

E.L. Bocchieri & J.G. Wilpon, "Discriminative feature selection for speech recognition", *Computer Speech and Language*, vol. 7, pp. 229-246, 1993.

A. Bonafonte, X. Ros & J.B. Mariño, "Explicit modelling of duration in HMM: an efficient algorithm", in *Speech Recognition and Coding. New Advances and Trends*, édité par A. Rubio Ayuso et J. López Soler, NATO ASI Series, vol. F 147, Springer-Verlag, pp. 80-83, 1995.

H. Bonneau, *Quantification vectorielle et adaptation au locuteur*, Thèse de l'Université Paris Sud, 1987.

L. Bottou, *Une approche théorique de l'apprentissage connexioniste; applications à la reconnaissance de la parole*, Thèse de l'Université Paris Sud, 1991.

H. Bourlard, "Towards increasing speech recognition error rates", *Proc. Eurospeech*, pp. 883-894, Madrid, 1995.

H. Bourlard & C.J. Wellekens, "Links between Markov models and multilayer perceptrons", *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 12, no. 12, pp. 1-12, 1990.

H. Bourlard & N. Morgan, "Continuous speech recognition by connectionist statistical methods", *IEEE Trans. on Neural Networks*, vol. 4, no. 6, pp. 893-909, 1993.

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

F. Brugnara, R. De Mori, D. Giuliani & M. Omologo, "A family of parallel hidden Markov models", *Proc. ICASSP*, pp. I 377-380, San Francisco, 1992(a).

F. Brugnara, R. De Mori, D. Giuliani & M. Omologo, "Improved connected digit recognition using spectral variation functions", *Proc. ICSLP*, pp. 627-630, Banff, 1992(b).

D. Burshtein, "Robust parametric modeling of durations in hidden Markov models", *Proc. ICASSP*, pp. 548-551, Detroit, 1995.

M.A. Bush & G.E. Kopec, "Network-based connected digit recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-35, pp. 1401-1413, 1987.

T. Cacoullos, "Estimation of a multivariate density", *Annals of Inst. of Stat. Math.*, vol. 18, pp. 179-189, 1966.

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

M.-J. Caraty, *Contribution au décodage acoustico-phonétique: études de distances interspectrales et reconnaissance de cycles vocaliques*, Thèse de l'Université Paris 6, 1987.

M.-J. Caraty & C. Montacié, "Intégration de la décomposition temporelle généralisée dans un système d'apprentissage symbolique. Application à la reconnaissance des voyelles", *Actes des 19èmes JEP*, pp. 387-392, Bruxelles, 1992.

M.-J. Caraty, C. Montacié & C. Barras, "Integration of frequential and temporal structurations in a symbolic learning system", *Proc. ICSLP* , pp. 475-478, Banff, octobre 1992.

M.-J. Caraty, C. Barras, F. Lefèvre & C. Montacié, "D-DAL: un système de dictée vocale développé sous l'environnement HTK", *Actes des 21èmes JEP*, 1996 (à paraître).

F. Casacuberta, E. Vidal, B. Mas & H. Rulot, "Learning the structure of HMMs through grammatical inference techniques", *Proc. ICASSP*, pp. 717-720, Albuquerque, 1990.

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

G. Chollet, "Evaluation of ASR systems, algorithms and databases", in *Speech Recognition and Coding. New Advances and Trends*, édité par A. Rubio Ayuso et J. López Soler, NATO ASI Series, vol. F 147, Springer-Verlag, pp. 32-40, 1995.

G. Chollet & C. Gagnoulet, "On the evaluation of speech recognizers and data bases using a reference system", *Proc. ICASSP*, pp. 2026-2029, Paris, 1982.

K. Choukri, *Quelques approches pour l'adaptation aux locuteurs en reconnaissance automatique de la parole*, Thèse de l'ENST, 1987.

Y.L. Chow, M.O. Dunham, O. Kimball, M.A. Krasner, G. Kubala, J. Makhoul, P.J. Price, S. Roucos & R. Schwartz, "BYBLOS: the BBN continuous speech recognition system", *Proc. ICASSP*, pp. 89-92, Dallas, 1987.

F. Class, A. Kaltenmeier, P. Regel & K. Trottler, "Fast speaker adaptation for speech recognition systems", *Proc. ICASSP*, pp. 133-136, Albuquerque, 1990.

R. Cole et al. (collectif de 24 auteurs), "The challenge of spoken language systems: research directions for the nineties", *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 1-21, 1995.

G.D. Cook & A.J. Robinson, "Utterance clustering for large vocabulary continuous speech recognition", *Proc. Eurospeech*, pp. 219-222, 1995.

J.W. Cooley & J.W. Tukey, "An algorithm for machine calculation of complex Fourier series", *Math. Comput.*, vol. 19, pp. 297-301, 1965.

A. Corraza, R. De Mori, R Gretter & G Satta, "Language modeling using stochastic context-free grammars", *Speech Communication*, vol. 13, pp. 163-170, 1993.

S. Cox, "Predictive speaker adaptation in speech recognition", *Computer Speech and Language*, vol. 9, pp. 1-17, 1995.

T.H. Crystal & A.S. House, "Segmental durations in connected speech signals: Preliminary results", *J. Acoust. Soc. Am.*, vol. 72, no. 3, pp. 705-716, 1982.

T.H. Crystal & A.S. House, "Segmental durations in connected-speech signals: Current results", *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1553-1573, 1988.

C. d'Alessandro & C. Demars, "Représentations temps-fréquence du signal de parole", *Traitement du Signal*, vol. 9, no. 2, pp. 153-173, 1992.

S.B. Davis & P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

R. De Mori & G. Flammia, "Speaker-independent consonant classification in continuous speech with distinctive features and neural networks", *J. Acoust. Soc. Am.*, vol. JASA-94, no. 6, pp. 3091-3103, 1993.

R. De Mori, M. Galler & F. Brugnara, "Search and learning strategies for improving hidden Markov models", *Computer Speech and Language*, vol. 9, pp. 107-121, 1995.

A.-M. Derouault, "Context-dependent phonetic Markov models for large vocabulary speech recognition", *Proc. ICASSP*, pp. 360-363, Dallas, 1987.

L. Devillers, *Reconnaissance de parole continue avec un système hybride neuronal et markovien*, Thèse de l'Université Paris 11, 1992.

B. Dubuisson, *Diagnostic et reconnaissance des formes*, Hermes, 1990.

P. Dumouchel & D. O'Shaughnessy, "Segmental duration and HMM modeling", *Proc. Eurospeech*, pp. 803-806, Madrid, 1995.

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

J.D. Ferguson, "Variable duration models for speech", in *Proc. Symp. on the Application of Hidden Markov Models to Text and Speech*, édité par J.D. Ferguson , pp. 143-179, Princeton, New-Jersey, 1980.

G. Flammia, P. Dalsgaard, O. Andersen & B. Lindberg, "Segment based variable frame rate speech analysis and recognition using a spectral variation function", *Proc. ICSLP*, pp. 983-986, Banff, 1992.

G.D. Forney, "The Viterbi algorithm", *Proc. of the IEEE*, vol. 61, no. 3, pp. 268-278, 1973.

S. Furui, "Speaker independent isolated word recognition using dynamic features of speech spectrum", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 1, pp. 52-59, 1986(a).

S. Furui, "On the role of spectral transition for speech perception", *J. Acoust. Soc. Am.*, vol. 80, no. 4, pp. 1016-1025, 1986(b).

S. Furui, "Unsupervised speaker adaptation method based on hierarchical spectral clustering", *Proc. ICASSP*, pp. 286-289, Glasgow, 1989.

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

T. Galas & X. Rodet, "An improved cepstral method for source-filter system deconvolution: application to musical sounds signal", *ICMC*, Glasgow, 1990.

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

J.L. Gauvain, L.F. Lamel, G. Adda & M. Adda-Decker, "Speaker-independant speech dictation", *Speech Communications*, vol. 15, pp. 21-37, 1994.

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

L. Gillick & S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms", *Proc. ICASSP*, pp. 532-535, Glasgow, 1989.

Y. Gong & W.C. Treurniet, "Duration of phones as function of utterance length and its use in automatic speech recognition", *Proc. Eurospeech*, pp. 315-318, Berlin, 1993.

Y. Gotoh, M. Hochberg & H. Silverman, "Using MAP estimated parameters to imporove HMM speech recognition performance", *Proc. ICASSP*, pp. (I)229-232, Adelaide, 1994.

J. Goût, *Apprentissage appliqué à la reconnaissance de la parole*, Rapport de stage du DEA IARFA, Université Paris 6, 1993.

R.M. Gray, "Vector quantization", *IEEE ASSP Mag.*, vol. 1(2):4-29, 1984.

Y. Grenier, "Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur", *Actes des 11èmes JEP*, pp. 163-171, Strasbourg, 1980.

H-Y. Gu, C-Y. Tseng & L-S. Lee, "Isolated-utterance speech recognition using hidden Markov models with bounded state duration", *IEEE Trans. on Signal Processing*, vol. 39, no. 8, pp. 1743-1752, 1991.

V. Gupta, M. Lennig, P. Mermelstein, P. Kenny, P.F. Seitz & D. O'Shaughnessy, "Use of minimum duration and energy contour for phonemes to improve large vocabulary isolated-word recognition", *Computer Speech & Language*, vol. 6, pp. 345-359, 1992.

lope based distortion measures for all-pole models of speech", *Proc. ICASSP*, pp. 757-760, Tokyo, 1986.

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

F.J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform", *Proc. of the IEEE*, vol. 66, no. 1, pp. 51-83, 1978.

J.-P. Haton, "Les modèles neuronaux et hybrides en reconnaissance automatique de la parole: état des recherches", *Ecole Thématique: Fondements et Perspectives en Traitement Automatique de la Parole*, H. Méloni éd., pp. 139-153, 1995.

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

H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, vol. JASA-87, no. 4, pp. 1738-1752, 1990.

X.D. Huang, "Phoneme classification using semicontinuous hidden Markov models", *IEEE Trans. on Signal Processing*, vol. 40, no. 5, pp. 1062-1067, 1992.

X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, K.-F. Lee & R. Rosenfeld, "The SPHINX-II speech recognition system: an overview", *Computer Speech and Language*, vol. 7, pp. 137-148, 1993(a).

X. Huang, H.W. Hon, M.Y. Hwang & K.-F. Lee, "A comparative study of discrete, semicontinuous, and continuous hidden Markov models", *Computer Speech and Language*, vol. 7, pp. 359-368, 1993(b).

K. Hübener, "Using segmental coefficients in HMM speech recognition", *Proc. Eurospeech*, pp. 935-938, Madrid, 1995.

A. Imamura, "Speaker-adaptive HMM-based speech recognition with a stochastic speaker classifier", *Proc. ICASSP*, pp. 841-844, Toronto, 1991.

F. Itakura, "Minimum prediction residual principle applied to speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, pp. 67-72, 1975.

C.R. Jankovski, H.-D. Vo & R.Lippmann, "A comparison of signal processing front ends for automatic word recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4, pp. 286-293, 1995.

F. Jelinek, "Continuous speech recognition by statistical methods", *Proc. of the IEEE*, vol. 64, no. 4, pp. 532-556, 1976.

F. Jelinek, "The development of an experimental discrete dictation recognizer", *Proc. of the IEEE*, vol. 73, no. 11, pp. 1616-1624, 1985.

M. Jones & P.C. Woodland, "Using relative duration in large vocabulary speech recognition", *Proc. Eurospeech*, pp. 311-314, Berlin, 1993.

D. Jouvet, K. Bartkova & J. Monné, "On the modelization of allophones in an HMM based speech recognition system", *Proc. Eurospeech*, pp. 923-926, Gènes, 1991(a).

D. Jouvet, L. Mauuaury & J. Monné, "Automatic adjustments of the structure of Markov models for speech recognition applications", *Proc. Eurospeech*, pp. 927-930, Gènes, 1991(b).

D. Jouvet, M. Dautremont et A. Gossart, "Comparaison des multimodèles et des densités multigaussiennes pour la reconnaissance de la parole par modèles de Markov", *Actes des 20èmes JEP*, pp. 159-164, 1994.

B.-H. Juang & L.R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, no. 6, pp. 1404-1413, 1985.

B.-H. Juang, L.R. Rabiner & J.G. Wilpon, "On the use of bandpass liftering in speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 7, pp. 947-954, 1987.

J.-C. Junqua, "Utilisation d'un modèle d'audition et de connaissances phonétiques en reconnaissance automatique de la parole", *Traitement du signal*, vol. 7, no. 4, pp. 275-284, 1990.

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

S. Kapadia, V. Valtchev & S.J. Young, "MMI training for continuous phoneme recognition on the TIMIT database", *Proc. ICASSP*, pp. II.491-494, Minneapolis, 1993.

P. Kenny, M. Lennig & P. Mermelstein, "Speaker adaptation in a large-vocabulary gaussian HMM recognizer", *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 12, no. 9, pp. 917-920, 1990.

D. Klatt, "Review of the ARPA speech understanding project", *J. Acoust. Soc. Am.*, vol. JASA-62, no. 6, pp. 1345-1366, 1977.

D. Klatt, "Prediction of perceived phonetic distance from critical band spectra: a first step", *Proc. ICASSP*, pp. 1278-1281, Paris, 1982.

W. Koenig, H.K. Duhn & L.Y. Lacy, "The sound spectrograph", *J. Acoust. Soc. Am.*, vol. 18, pp. 19-49, 1946.

T. Kosaka, J. Takami & S. Sagayama, "Rapid speaker adaptation using speaker-mixture allophone models applied to speaker-independent speech recognition", *Proc. ICASSP*, pp. II 570-573, Minneapolis, 1993.

T. Kosaka & S. Sagayama, "Tree-structured speaker clustering for fast speaker adaptation", *Proc. ICASSP*, pp. I 245-248, Adelaide, 1994.

L.F. Lamel & J.L. Gauvain, "High performance speaker-independant phone recognition using CDHMM", *Proc. Eurospeech*, pp. 121-124, 1993.

C. Lecordier, *Intégration d'un module spécifique de reconnaissance des voyelles au sein d'un système de décodage acoustico-phonétique fondé sur les modèles de Markov*, Rapport de stage du DEA Mathématiques, Informatique et Applications aux Sciences de l'Homme, Université Paris 5, 1995.

C.H. Lee, C.H. Lin & B.H. Juang, "A study on speaker adptation of the parameters of continuous density hidden Markov models", *IEEE Trans. on Signal Processing*, vol. 39, no. 4, pp. 806-814, 1991.

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

C.H. Lee et J.L. Gauvain, "Speaker adaptation based on MAP estimation of HMM parameters", *Proc. ICASSP*, pp. II 558-561, Minneapolis, 1993.

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

K.-F. Lee, "Context-dependant phonetic hidden Markov models for speaker-independent continuous speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 4, pp. 599-609, 1990.

K.-F. Lee & H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, no. 11, pp. 1641-1648, 1989.

K.-F. Lee, H.-W. Hon & R. Reddy, "An overview of the SPHINX speech recognition system", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, no. 1, pp. 35-45, 1990.

F. Lefèvre, *Contribution au système de dictée vocale D-DAL; N-Best et modèles de langage*, Rapport de stage du DEA IARFA, Université Paris 6, 1995.

J.-L. Le Floch, *Comparaison de méthodes de segmentation*, Rapport de stage du DEA IARFA, Université Paris 6, 1991.

J.-L. Le Floch, C. Montacié & M.-J. Caraty, "Investigations on speaker characterization from Orphée system technics", *Proc. ICASSP*, pp. (I)149-152, Adelaide, 1994.

C.J. Leggetter & P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

V. Le Maire, R. André-Obrecht & D. Jouvet, "An acoustic-phonetic decoder based on an automatic segmentation algorithm", *Proc. Eurospeech*, pp. 392-395, Paris, 1989.

"Quand les puces ont des problèmes de syntaxe…", *Le Monde*, 5 octobre 1994.

S.E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition", *Computer Speech & Language*, vol. 1, no 1, pp. 29-45, 1986.

J.S. Liénard, "Variabilité, contraintes et spécification de la parole: un cadre théorique", *Actes du séminaire "Variabilité et spécificité du locuteur: études et applications",* édité par H. Méloni, Marseille, pp. 1-10, 1989.

R.P. Lippmann, "An introduction to computing with neural nets", *IEEE ASSP Mag.*, vol. 4, no. 2, pp. 4-22, 1987.

A. Ljolje, "Speaker clustering for improved speech recognition", *Proc. Eurospeech*, pp. 631-634, Berlin, 1993.

A. Ljolje, "High accuracy phone recognition using context clustering and quasi-triphonic models", *Computer Speech and Language*, vol. 8, pp. 129-151, 1994(a).

A. Ljolje, "The importance of cepstral parameter correlations in speech recognition", *Computer Speech and Language*, vol. 8, pp. 223-232, 1994(b).

A. Ljolje & S.E. Levinson, "Development of an acoustic-phonetic Markov model for continuous speech recognition", *IEEE Trans. on Signal Processing*, vol. 39, no. 1, pp. 29-39, 1991.

P. Lockwood & M. Blanchet, "An algorithm for the dynamic inference of hidden Markov models (DIHMM)", *Proc. ICASSP*, pp. II 251-254, Minneapolis, 1993.

J. Makhoul, "Linear prediction: a tutorial review", *Proc. of the IEEE*, vol. 63, no. 4, pp. 561-580, 1975.

F. Malbos, R. André-Obrecht & M. Baudry, "Comparaison de deux méthodes non paramétriques pour la détection des occlusives sourdes", *Actes des 20èmes JEP*, pp. 175-180, Trégastel, 1994.

J.N. Marcus, "Phonetic recognition in a segment-based HMM", *Proc. ICASSP*, pp. II 479-482, Minneapolis, 1993.

J.-J. Mariani, *ESOPE: un système de compréhension de la parole continue*, Thèse d'Etat, Paris, 1982.

J. Mariani, "Reconnaissance de la parole: progrès et tendances", *Traitement du signal*, vol. 7, no. 4, pp. 239-266, 1990.

L. Mathan & L. Miclet, "Speaker hierarchical clustering for improving speaker-independant HMM word recognition", *Proc. ICASSP*, pp. 149-152, Albuquerque, 1990.

B. Mayrargue, *Etude des propriétés de l'adaptation au locuteur dans un système de reconnaissance de la parole continue utilisant les modèles de Markov cachés*, Rapport de stage du DEA IARFA, Université Paris 6, 1995.

A. Mellouk & P. Gallinari, "A discriminative neural prediction system for speech recognition", *Proc. ICASSP*, pp. 533-536, Minneapolis, 1993.

A. Mellouk & P. Gallinari, "Discriminative training for improved neural prediction systems", *Proc. ICASSP*, pp. 233-236, Adelaide, 1994.

C.D. Mitchell & L.H. Jamieson, "Modeling duration in a hidden Markov Model with the exponential family", *Proc. ICASSP*, pp. (II)331-334, Minneapolis, 1993.

C.D. Mitchell, M.P. Harper & L.H. Jamieson, "Using explicit segmentation to improve HMM phone recognition", *Proc. ICASSP*, pp. 229-232, Detroit, 1995.

C. Montacié, *Décodage acoustico-phonétique: apport de la décomposition temporelle généralisée et de transformations spectrales non linéaires*, Thèse de l'ENST, 1991.

C. Montacié & G. Chollet, "Systèmes de référence pour l'évaluation d'applications et la caractérisation de bases de données en reconnaissance automatique de la parole", *Actes des 16e JEP*, pp. 323-326, Hammamet, 1987.

C. Montacié, P. Deléglise, F. Bimbot & M.-J. Caraty, "Cinematic techniques for speech processing: temporal decomposition and multivariate linear prediction", *Proc. ICASSP*, pp. (I)153-156, San Francisco, 1992a.

C. Montacié & J.-L. Le Floch, "AR-vector models for free-text speaker recognition", *Proc. ICSLP*, pp. 611-614, Banff, 1992b.

C. Montacié & J.-L. Le Floch, "Discriminant AR-vector models for free-text speaker verification", *Proc. Eurospeech*, pp. 161-164, Berlin, 1993.

T. Moudenc, D. Jouvet & J. Monné, "On using a priori segmentations of the speech signal in an N-best solutions post-processing", *Proc. ICASSP*, pp. 580-583, Detroit, 1995.

C.S. Myers & L.R. Rabiner, "Connected digit recognition using a level building DTW algorithm", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29(3):351-363, 1981.

H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. no. 2, pp. 263-271, 1984.

H. Ney, "Acoustic modelling of phoneme units for continuous speech recognition", *Proc. 5th EUSIPCO*, pp. 65-72, Barcelona, 1990.

L.T. Niles, "TIMIT phoneme recognition using an HMM-derived recurrent neural network", *Proc. Eurospeech*, pp. 559-562, Gènes, 1991.

NIST, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus", 1990.

Y. Normandin, "Optimal splitting of HMM gaussian mixture components with MMIE training", *Proc. ICASSP*, pp. 449-452, Detroit, 1995.

Y. Normandin, R. Cardin & R. De Mori, "High-performance connected digit recognition using Maximum Mutual Information estimation", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 299-311, 1994.

M. Ostendorf & S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-37, no. 12, pp. 1857-1869, 1989.

N. Parlangeau, R. André-Obrecht & A. Marchal, "Automatic labelling of multi-sensor speech databases: issues and perspectives", *Proc. Eurospeech*, pp. 1905-1908, Madrid, 1995.

E. Parzen, "On estimation of a probability density function and mode", *Ann. Math. Statist.*, vol. 33, pp. 1065-1076, 1962.

J. Pierce, "Whither speech recognition?", *J. Acoust. Soc. Am.*, vol. JASA-46, pp. 1049-1051, 1969.

L.R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition", *Proc. of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.

L.R. Rabiner & R.W. Schaffer, *Digital processing of speech signals*, Prentice-Hall, 1978.

L.R. Rabiner, S.E. Levinson, A.E. Rosenberg & J.G. Wilpon, "Speaker-independent recognition of isolated words using clustering techniques", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27(4):336-349, 1979.

L.R. Rabiner, J.G. Wilpon & F.K. Soong, "High performance connected digit recognition using hidden Markov models", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 8, pp. 1214-1225, 1989.

P. Ramesh & J.G. Wilpon, "Modeling state duration in hidden Markov models for automatic speech recognition", *Proc. ICASSP*, pp. (I)381-384, San Francisco, 1992.

M. Richard & R. Lippmann, "Neural network classifiers estimate Bayesian *a posteriori* probabilities", *Neural Computation*, 3(4):461-483, 1991.

A.J. Robinson, "An application of recurrent nets to phone probability estimation", *IEEE Trans. on Neural Networks*, vol. 5, no. 2, pp. 298-305, 1994.

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

T. Robinson & F. Fallside, "A recurrent error propagation network speech recognition system", *Computer Speech and Language*, vol. 5, pp. 259-274, 1991.

T. Robinson, M. Hochberg & S. Renals, "IPA: improved phone modelling with Recurrent Neural Networks", *Proc. ICASSP*, pp. (I) 37-40, Adelaide, 1994.

M. J. Russel & R. K. Moore, "Explicit modelling of state occupency in hidden Markov models for automatic speech recognition", *Proc. ICASSP*, pp. 5-8, Tampa, 1985.

H. Sakoe, "Two-level DP-matching — A dynamic programming-based pattern matching algorithm for connected speech recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27(6), pp. 588-595, 1979.

H. Sakoe & S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 26, no. 1, pp. 43-49, 1978.

A. Sankar, F. Beaufays & V. Digalakis, "Training data clustering for improved speech recognition", *Proc. Eurospeech*, pp. 503-506, 1995.

G. Saporta, *Probabilités: analyse des données et statistiques*, Technip, 1990.

R. Schwartz, Y.Chow, O. Kimball, S. Roucos, M. Krasner & J. Makhoul, "Context-dependent modeling for acoustic-phonetic recognition of continuous speech", *Proc. ICASSP*, pp. 1205-1208, Tampa, 1985.

R. Schwartz, Y.L. Chow & F. Kubala, "Rapid speaker adaptation using a probabilistic spectral mapping", *Proc. ICASSP*, pp. 633-636, Dallas, 1987.

R. Schwartz, C. Barry, Y.L. Chow, A. Derr, M.-W. Feng, O. Kimball, G. Kubala, J. Makhoul, & J. Vandegrift, "The BBN BYBLOS continuous speech recognition system", *Proc. DARPA Workshop on Speech and Natural Language*, pp. 94-99, Philadelphie, 1989.

R. Schwartz & F. Kubala, "Hidden Markov models and speaker adaptation", in *Speech Recognition and Understanding - Recent Advances*, édité par P. Laface & R. De Mori, NATO ASI Series, vol. F75, pp. 31-57, Springer-Verlag, 1992.

S. Seneff, "A joint/mean-rate model of auditory speech processing", *Journal of Phonetics*, vol. 16, no. 1, pp. 55-76, 1988.

S. Seneff & V. Zue, "Transcription and alignment of the TIMIT database", in *Getting started with the DARPA TIMIT CD-ROM: an acoustic-phonetic continuous speech database*, NIST, 1988.

K. Shikano, K.-F. Lee & R. Reddy, "Speaker adaptation through vector quantization", *Proc. ICASSP*, pp. 2643-2646, Tokyo, 1986.

R. Sitaram & T. Sreenivas, "Phoneme recognition in continuous speech using large inhomogeneous hidden Markov models", *Proc. ICASSP*, pp. (I)41-44, Adelaide, 1994.

S.P. Smith & A.K. Jain, "A test to determine the multivariate normality of a data set", *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. PAMI-10, no. 5, pp. 757-761, 1988.

F.K. Soong & A.E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 6, pp. 871-879, 1988.

S. Soudoplatoff, "Markov modeling of continuous parameters in speech recognition", *Proc. ICASSP*, pp. 45-48, Tokyo, 1986.

H.Y. Su & R. André-Obrecht, "Adaptation en cours de reconnaissance d'un dictionnaire de références phonétiques à un nouveau locuteur", *Actes des 17èmes JEP*, pp. 140-144, 1988.

N. Suaudeau, R. André-Obrecht & B. Delyon, "Modélisation de la durée globale d'un son dans un modèle de Markov caché: application à la reconnaissance de nombres", *Actes des 19èmes JEP*, pp. 189-194, Bruxelles, 1992.

N. Suaudeau & R. André-Obrecht, "An efficient combination of acoustic and supra-segmental informations in a speech recognition system", *Proc. ICASSP*, pp. I 65-68, Adelaide, 1994.

A. Tridgell, B. Millar & K.-A. Do, "Alternative pre-processing techniques for discrete hidden Markov model phoneme recognition", *Proc. ICSLP*, pp. 631-634, Banff, 1992.

J.-P. Tubach, G. Chollet, K. Choukri, C. Montacié, C. Mokbel & H. Valbret, "Adaptation au locuteur de systèmes de reconnaissance. Régression linéaire multiple et perceptrons multi-couches", *Traitement du Signal*, vol. 7, n° 4, pp. 285-292, 1990.

D. Van Compernolle, J. Smolders, P. Jaspers & T. Hellemans, "Speaker clustering for dialectic robustness in speaker independent recognition", *Proc. Eurospeech*, pp. 723-726, Genova, 1991.

S.V. Vaseghi, "State duration modelling in hidden Markov models", *Signal Processing*, vol. 41, no. 1, pp. 31-41, 1995.

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

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

T.K. Vintsjuk, "Element-wise recognition of continuous speech composed of words from a specified dictionary", *Kibernetika*, vol. 7:133-143, 1971.

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

R.A. Wagner & M.J. Fisher, "The string to string correction problem", *JACM*, 1974.

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano & K. Lang, "Phoneme recognition using time-delay neural networks", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 3, pp. 328-339, 1989.

X. Wang, "Durational modelling in HMM-based speech recognition: towards a justified measure", in *Speech Recognition and Coding. New Advances and Trends*, édité par A. Rubio Ayuso et J. López Soler, NATO ASI Series, vol. F 147, Springer-Verlag, pp. 128-131, 1995.

X. Wang, L. ten Bosch & L. Pols, "Impact of dimensionality and correlation of observation vectors in HMM-based speech recognition", *Proc. Eurospeech*, pp. 1583-1586, Berlin, 1993.

J.G. Wilpon, C.-H. Lee & L.R. Rabiner, "Connected digit recognition based on improved acoustic resolution", *Computer Speech and Language*, vol. 7, pp. 15-26, 1993.

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

H. Yé, M.-J. Caraty, L.-J. Boë & D. Tuffelli, "Structural (phonetic) evaluation of dissimilarities functions used in speech recognition", *Proc. Eurospeech*, pp. (2) 404-407, Paris, 1989.

S.J. Young, *HTK Version 1.4: Reference Manual and User Manual*, Cambridge University Engineering Department - Speech Group, 1992(a).

S.J. Young, "The general use of tying in phoneme-based HMM speech recognizers", *Proc. ICASSP*, pp. 569-572, San Francisco, 1992(b).

S.J. Young, *The HTK hidden Markov model toolkit: design and philosophy,* Cambridge University Engineering Department - Technical report 152, 1994.

S.J. Young & P.C. Woodland, "State clustering in hidden Markov model-based continuous speech recognition", *Computer Speech and Language*, vol. 8, pp. 369-383, 1994.

Y. Zhao, "A speaker-independent continuous speech recognition system using continuous mixture gaussian density HMM of phoneme-sized units", *IEEE Trans. on Speech and Audio Processing*, vol. 1, no. 3, pp. 345-361, 1993.

Y. Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker independent continuous speech recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 3, pp. 380-394, 1994.

E. Zwicker, "Subdivision of the audible frequency range into critical bands", *J. Acoust. Soc. Am.*, vol. 33, pp. 248, 1961.

ANNEXE A. APPRENTISSAGE AVEC HTK

Nous décrivons dans cette partie les détails techniques de la mise en œuvre de HTK pour la construction d'un système de reconnaissance de la parole continue indépendamment du locuteur. L'apprentissage de modèles phonétiques sur la base de données TIMIT est réalisé par un programme écrit en langage de commande C-shell.

A.1. Utilisation de HTK

Avant de détailler les scripts réalisant l'apprentissage des systèmes de reconnaissance présentés au chapitre 2, nous abordons l'utilisation d'HTK d'un point de vue pratique en donnant quelques exemples concrets illustrant l'apprentissage de modèles phonétiques indépendants du contexte[2]. L'environnement de travail est une station sous le système d'exploitation UNIX, avec comme langage de commande le C-shell. La distribution 1.4A de HTK est installée dans le répertoire identifié par la variable d'environnement `$HBIN`, et le répertoire contenant les programmes exécutables de HTK a été ajouté dans la liste des chemins d'accès du système (variable système `$path`). Enfin, la base de données TIMIT est accessible dans le répertoire désigné par la variable d'environnement `$TIMIT`.

A.1.1. Signaux de parole

Les 630 locuteurs américains de la base TIMIT sont séparés entre l'ensemble d'apprentissage et l'ensemble de test (répertoires `train` et `test`), et sont répartis en 8 "dialectes régionaux" (sous-répertoires `dr1` à `dr8`). Pour chaque locuteur, un répertoire du nom du locuteur contient les 10 phrases qu'il a prononcées; elles sont identifiées par le nom du corpus (`sa`, `si` et `sx`) suivi d'un numéro. Quatre fichiers sont disponibles par phrase: le texte lu (suffixe `.txt`), le signal enregistré (`.wav`), l'étiquetage de ce signal par mot (`.wrd`) et par phonème (`.phn`). La base est donc structurée de manière hiérarchique:

$$\texttt{\$TIMIT}/\begin{cases}\texttt{test}\\ \texttt{train}\end{cases}/\begin{cases}\texttt{dr1}\\ \vdots\\ \texttt{dr8}\end{cases}/\begin{cases}\texttt{f*}\\ \texttt{m*}\end{cases}/\begin{cases}\texttt{sa*}\\ \texttt{si*}\\ \texttt{sx*}\end{cases}\begin{cases}\texttt{.txt}\\ \texttt{.wav}\\ \texttt{.phn}\\ \texttt{.wrd}\end{cases}$$

[2]Pour plus de clarté, tous les noms de commandes, de fichiers ou de variables ainsi que les lignes de commandes et les sorties d'exécution sont imprimées dans cette annexe en police `courier`.

Les fichiers de signal (.wav) sont au format NIST, c'est-à-dire constitués d'un en-tête en ASCII de 1024 caractères suivi des échantillons sur 16 bits. Nous les avons compressés avec l'utilitaire shorten, écrit et distribué par A.J. Robinson de l'Université de Cambridge. Les outils fournis par HTK ont ensuite été modifiés pour que la décompression des signaux avec shorten soit automatique et transparente. L'outil HList affiche la valeur des échantillons d'un fichier de signal:

```
> HList -F NIST $TIMIT/test/dr1/faks0/sa1.wav

File: (…)/test/dr1/faks0/sa1.wav [NIST]
     Num Samples: 63488       Sample Period: 62.5 usecs
     Sample Size: 2 bytes     Sample Kind:   WAVEFORM

     0:  3  5  2  2  1  2  3  1  2  2
     10: 4  4  4  3  4  5  3  4  5  1
     20: …
```

imprime la valeur des échantillons du signal, et le fichier au format NIST est décompressé en mémoire avant affichage.

L'outil HSLab, destiné à l'étiquetage manuel d'un signal, peut servir à afficher simplement le signal temporel dans une fenêtre X-Windows. Une restitution médiocre du son est possible sur une SUN; nous avons apporté une modification à cet outil pour obtenir la restitution sur un PC sous UNIX équipé d'une carte SoundBlaster.

```
> HSLab -F NIST $TIMIT/test/dr1/faks0/sa1.wav
```

A.1.2. Fichiers de paramètres

Les paramètres retenus pour représenter le signal sont 12 coefficients MFCC et l'énergie du signal. Les fichiers de paramètres des 6300 phrases de la base TIMIT sont calculés et stockés dans un seul répertoire identifié par la variable $PARAM, avec pour suffixe .mfcc. Comme plusieurs locuteurs peuvent prononcer la même phrase, l'identité du locuteur a été rajoutée en début du nom du fichier afin d'éviter les confusions:

```
> HCode -m -n 12 -p 24 -e -f 10.0 -w 25.0 -h -k 0.97
          $TIMIT/test/dr1/faks0/sa1.wav
          $PARAM/faks0.sa1.mfcc
```

A partir des 13 paramètres déjà calculés, les outils de HTK peuvent calculer automatiquement les dérivées premières et secondes de ces coefficients lors de la lecture des fichiers, ce qui permet d'économiser de la place sur le disque. La variable d'environnement HCOERCE force la conversion à la valeur désirée; ainsi, pour obtenir des coefficients différentiels du premier ordre, on exécute:

```
> setenv HCOERCE MFCC_E_D
```

Les variables HDELTA et HACCS indiquent le nombre de trames utilisées pour le calcul des vitesses et des accélérations; les valeurs par défaut sont de deux trames. L'effet de la conversion peut être constaté avec `HList`. Il n'est pas nécessaire de spécifier un format car les fichiers générés par HTK contiennent un en-tête spécial:

```
> HList $PARAM/faks0.sa1.mfcc

File: (…)/faks0.sa1.mfcc [HTK]
     Num Samples:  395   Sample Period: 10000.0 usecs
     Sample Size: 104 bytes   Sample Kind: MFCC_E_D
     0: 1_ -7.7736 … 12_  -0.0903   E_  -0.3598
        1/  0.9533 … 12/   0.4855   E/   0.0140
     1: …
```

A.1.3. Fichiers d'étiquettes

L'étiquetage phonétique fourni avec la base TIMIT (fichiers `*.phn`) est très fin (61 classes phonétiques en incluant les silences), et certains allophones sont regroupés avant l'apprentissage. L'éditeur lexical `HLEd` applique des commandes simples (fusion, suppression d'étiquette) à un fichier d'étiquettes. Un script fourni avec HTK permet d'obtenir les 48 modèles phonétiques proposés par K.F. Lee:

```
> more $HBIN/HTKEtc/Scripts/KFLset.led

     S
     D q
     R uw ux
     …
     R sil h# #h pau
```

La liste des 48 phonèmes obtenus après l'application de ce script est donnée dans:

```
> more $HBIN/HTKEtc/PhoneLists/KFLphones

     iy
     ih
     …
     sil
```

Les fichiers d'étiquettes de la base TIMIT sont dans un format spécifique à cette base. Ils sont reconnus par HTK à condition de préciser le format lors de l'appel à l'éditeur lexical:

```
> HLEd -G TIMIT (…)/KFLset.led (…)/sa1.phn
```

Un fichier d'étiquettes au format HTK est alors généré avec le suffixe `.lab`. Les étiquettes des phrases de la base TIMIT sont traduites avec ce script, et rangées dans un unique répertoire (`$LABEL`) en suivant la même convention que pour les fichiers MFCC.

A.1.4. Apprentissage des modèles

Tous nos modèles phonétiques utilisent comme prototype un modèle de Bakis à 3 états. Il faut créer un fichier décrivant ce prototype dans une syntaxe propre à HTK, en indiquant les probabilités d'émission des états (décrites par des gaussiennes) ainsi que les probabilités de transition et en leur donnant une valeur par défaut. Il n'existe pas d'outil

HTK pour générer automatiquement ces fichiers, mais un prototype de modèle gauche-droit à 3 états avec une gaussienne diagonale par état, pour des paramètres `MFCC_E_D` (12 MFCC + énergie+ 13 coefficients dérivés), est distribué avec HTK:

```
> more $HBIN/HTKEtc/Prototypes/med26dm1

<BeginHMM>
 <NumStates> 5 <VecSize> 26 <MFCC_E_D> <nullD> <diagC>
     <State> 2
          <Mean>      26  0.0 … 0.0
          <Variance> 26  1.0 … 1.0
     <State> 3
          <Mean>      26  0.0 … 0.0
          <Variance> 26  1.0 … 1.0
     <State> 4
          <Mean>      26  0.0 … 0.0
          <Variance> 26  1.0 … 1.0
     <TransP> 5
          0.0  1.0  0.0  0.0  0.0
          0.0  0.6  0.4  0.0  0.0
          0.0  0.0  0.6  0.4  0.0
          0.0  0.0  0.0  0.7  0.3
          0.0  0.0  0.0  0.0  0.0
<EndHMM>
```

Un script est ensuite appliqué à ce modèle avec l'éditeur de modèles `HHEd` afin de rajouter une transition entre l'état initial et final et obtenir le modèle de Bakis souhaité.

Pour chaque modèle de phonème, le prototype est utilisé pour apprendre les lois de probabilité associées. L'apprentissage est décomposé en deux étapes. L'outil `HInit` permet une initialisation rapide du modèle avec l'algorithme de Viterbi. L'outil `HRest` applique ensuite l'algorithme de Baum-Welch et donne une ré-estimation plus fine des lois de probabilité. L'apprentissage du modèle du phonème `iy` à partir d'un prototype `proto` est réalisé par les instructions:

```
> HInit -l iy -L $LABEL -S train -o hmm/iy proto
> HRest -l iy -L $LABEL -S train hmm/iy
```

Le fichier `train` doit contenir la liste des fichiers de paramètres de la base d'apprentissage. Les segments contenant l'étiquette choisie `iy` sont déterminés grâce aux fichiers d'étiquettes contenus dans le répertoire `$LABEL`. Après apprentissage, un fichier `iy` contenant les paramètres du modèle est créé dans le répertoire `hmm`. Il est possible de choisir le nombre d'itérations (option `-i`) et d'obtenir plus de détails sur la convergence de l'apprentissage en demandant la trace (option `-T`):

```
> HInit -i 10 -T 1 -l iy -L $LABEL -S train \
   -o hmm/iy proto
```

```
Initialising  HMM proto . . .
 States   :   2  3  4 (width)
 Mixes  s1:   1  1  1 ( 26  )
 Kinds    :  DIAGC MFCC_E_D
 SegLab   :  iy
 maxIter  :  10
 maxCIter :  16
 epsilon  :  0.000100
 minSeq   :  3
 Updating :  Means Variances MixWeights

  4626 Observations Loaded
Iteration 1: Average LogP = -117.87
Iteration 2: Average LogP = -111.35  Change = 6.51388
…
Iteration 10: Average LogP = -110.88  Change = 0.00019
Estimation aborted at iteration 10
```

A.1.5. Décodage acoustico-phonétique

Le décodage acoustico-phonétique des phrases de test est réalisé avec l'outil `HVite`, qui applique l'algorithme de Viterbi pour trouver la chaîne phonétique la plus probable correspondant à un fichier de paramètres. Dans le cas d'une tâche de DAP, tout enchaînement de phonèmes est autorisé, ce qui s'exprime dans la syntaxe de HTK par le fichier suivant:

```
> more $HBIN/HTKEtc/Networks/KFLnet

     $phn = iy | ih | b | d | … | sil;
     (<$phn>)
```

Il faut fournir à l'outil `HVite` le nom du répertoire contenant les modèles phonétiques (`hmm`), le nom d'un répertoire existant pour enregistrer les fichiers d'étiquettes résultant de la reconnaissance (ici `reco`), la liste des fichiers de paramètre pour le test (`test`), la liste des modèles phonétiques utilisés (`phones`), et la syntaxe du réseau (`net`):

```
> mkdir reco
> HVite -d hmm -L reco -S test phones KFLnet
```

Nous obtenons dans le répertoire `reco` la segmentation automatique des fichiers de test, dans des fichiers au suffixe `.rec`.

L'utilisation d'un bigramme phonétique, qui donne la probabilité de transition en fonction des fréquences observées, permet en général d'améliorer le décodage. Ce bigramme peut être estimé sur l'ensemble d'apprentissage par `HLStats`:

```
> HLStats -b bigram -S train_lab phones
```

Le fichier `train_lab` doit contenir la liste des fichiers de labels de l'ensemble d'apprentissage. Le poids de ces probabilités au cours du décodage peut être ajusté par le facteur d'échelle grammatical (option `-s`), ce qui permet de réduire les insertions:

```
> HVite -b bigram -s 3.0 -d hmm -L reco -S test phones
net
```

La segmentation obtenue après le décodage acoustico-phonétique est comparée avec l'étiquetage de référence avec l'outil `HResults` par un alignement dynamique:

```
> HResults -L $LABEL phones reco/*.rec
```

```
------------------------ Overall Results -------------
PHRASE: %Correct=0.00 [ … ]
PHONE:  %Corr=68.50, Acc=58.72 [ … ]
------------------------------------------------------
```

Les résultats fournissent les taux d'identification correcte des phrases complètes (`%Correct`) et des phonèmes (`%Corr`). Le deuxième pourcentage phonétique (`%Acc`) indique le pourcentage global de reconnaissance en tenant compte des insertions phonétiques: c'est celui qui indique les performances réelles du système.

On peut fusionner plusieurs classes lors des résultats, si deux modèles différents correspondent en fait à des allophones d'un même phonème; dans l'exemple suivant, les confusions entre `aa` et `ao` ne sont pas comptabilisées comme des erreurs:

```
> HResults -e aa ao … -L $LABEL phones reco/*.rec
```

A.1.6. Amélioration des modèles

La densité de probabilité utilisée pour estimer la probabilité d'émission des modèles est une simple loi normale à matrice de covariance diagonale. L'utilisation de multi-gaussiennes permet d'améliorer la modélisation des probabilités. Bien qu'il soit possible d'apprendre directement un modèle à multi-gaussiennes, il semble plus efficace d'augmenter progressivement le nombre de gaussiennes à partir d'un modèle simple. L'éditeur de modèles HHEd permet la modification simultanée de tous les modèles:

```
> more mixup_hed

MU 4 {*.state[2-4]}

> HHEd -d hmm_mix2 -e hmm_mix4 mixup_hed phones
```

Il faut ensuite procéder au réapprentissage de tous les modèles phonétiques:

```
> HRest -l $ph -L $LABEL -S train hmm_mix4/$ph
```

Ces deux étapes sont regroupées dans l'outil `RestUp` que nous avons développé à partir de `HRest`; il intègre un test du χ^2 pour choisir les gaussiennes à diviser:

```
> RestUp -k 3 -r 300 -a 8 -l $ph -L $LABEL -S train \
     -o hmm_mix8/$ph hmm_mix1/$ph
```

Après convergence de l'estimation, les gaussiennes multi-variables dont au moins 3 coefficients échouent au test du χ^2 (option `-k`) et pour lesquelles le nombre de trames d'apprentissage est supérieur à 300 (option `-r`) sont divisées. Le processus est répété jusqu'à convergence ou jusqu'à atteindre un maximum de 8 gaussiennes par densité de probabilité (option `-a`)

```
Reestimating HMM hmm_mix1/iy . . .
 States   :   2  3  4 (width)
 Mixes  s1:   1  1  1 ( 26  )
 Kinds    :  DIAGC MFCC_E_D
 SegLab   :  iy
 MaxIter  :  10
 Epsilon  :  0.000100
 Updating :  Transitions Means Variances MixWeights

4626 Examples loaded, Max length = 44, Min length = 3
Ave LogProb at iter 1 = -102.26266 using 4626 examples
Ave LogProb at iter 2 = -102.26276 using 4626 examples
change =   -0.00010
Estimation converged at iteration 2

Augmentation du nombre de gaussiennes par test du Chi2
Etat 2 flux 1 mixt 1 poids 12097.7 chi2 185.5 -> divise
Etat 3 flux 1 mixt 1 poids 20934.4 chi2 103.7 -> divise
Etat 4 flux 1 mixt 1 poids 15694.8 chi2 252.1 -> divise

Ave LogProb at iter 1 = -105.58366 using 4626 examples
 (…)
Ave LogProb at iter 10 =  -91.16864 using 4626 examples
change =    0.05386
Estimation aborted at iteration 10

Augmentation du nombre de gaussiennes par test du Chi2
Etat 2 flux 1 mixt 1 poids  5952.4 chi2 33.8 -> divise
Etat 2 flux 1 mixt 2 poids  5912.3 chi2 43.6 -> divise
Etat 3 flux 1 mixt 1 poids 11126.9 chi2 32.8 -> divise
Etat 3 flux 1 mixt 2 poids  8508.6 chi2 44.8 -> divise
Etat 4 flux 1 mixt 1 poids  6469.7 chi2 54.3 -> divise
Etat 4 flux 1 mixt 2 poids 10756.4 chi2 54.1 -> divise

 (…)
```

Pour l'apprentissage de modèles phonétiques dépendants du contexte, nous renvoyons au manuel de référence de HTK. La ré-estimation de ces modèles est réalisée en modèles connectés par l'outil `HERest`, et nous avons intégré à cet outil l'augmentation automatique du nombre de gaussiennes au moyen de l'outil `ERestUp`, qui est l'analogue en modèles connectés de `RestUp`.

A.2. Scripts d'apprentissage des modèles

Nous présentons les scripts permettant l'apprentissage et l'évaluation du système de base puis du système de référence. Le système de base est constitué de modèles phonétiques indépendants du contexte, et le système de référence prend en compte le contexte phonétique droit. Les scripts sont écrits en C-shell sous UNIX; ils utilisent les structures de contrôle spécifiques à ce langage de commande, comme l'alternative par '`if`' ou la répétition par '`foreach`', ainsi que la gestion des variables du shell et des variables d'environnement par '`set`' et '`setenv`'. Des commandes UNIX standards, comme la commande '`awk`' pour traiter séquentiellement les lignes d'un fichier texte, sont aussi largement utilisées. Enfin, la gestion des sous-programmes n'est pas assurée en langage C-shell; nous avons choisi de simuler cette fonction au moyen d'un alias appelé '`call`'. Le début de la procédure est défini par un label, et la fin est marquée par un '`exit`' qui entraîne le retour à la ligne d'appel. Les valeurs des variables d'environnement définies au moyen de la commande '`setenv`' sont transmises aux sous-programmes appelés; les valeurs des variables du shell sont quant à elles locales à une procédure. La variable d'environnement `TRACE` permet d'obtenir l'écho à l'écran des commandes exécutées.

```
########################################################################
# Scripts de génération de modèles phonétiques par HTK
# C. Barras, LAFORIA, 1995
########################################################################

########################################################################
# Gestion des appels de sous-programmmes
#=======================================================================
alias call '$SHELL -e $0'
if ($?TRACE) set echo
if ($#argv >= 1) goto $1

########################################################################
# Programme principal
#=======================================================================
main:

   # Définition des variables d'environnement
   #####################################

   # Conversion forcée des fichiers de paramètres par les outils de HTK
   setenv HCOERCE        MFCC_E_D

   # Répertoires
   #setenv HBIN          …/HTK_V1.4A
   #setenv TIMIT         …/timit
   setenv LABEL          lab
   setenv PARAM          mfcc

   # Fichiers
   setenv PHONES         $HBIN/HTKEtc/PhoneLists/KFLphones
   setenv NET            $HBIN/HTKEtc/Networks/KFLnet
   setenv TRAIN          train
   setenv TEST           test
```

```
   setenv LISTE         liste
   setenv PROTO         proto
   setenv BIGRAM        bigram

   # Système de base avec un nombre croissant de gaussiennes
   ####################################
   call init
   foreach m (1 2 4 8)
      setenv MIX        $m
      setenv HMM        mono_$m
      call monophones
      call reco
      setenv HMM_PREC   $HMM
   end

   # Initialisation des variables pour les modèles contextuels
   ####################################
   setenv NEWLAB        droit.lab
   setenv PHYSIC        droit.phy
   setenv LOGIC         droit.log
   setenv NET           droit.net
   setenv OCC           droit.occ
   setenv SCALE         3.0
   setenv RO            300
   setenv TC            0.3

   call init_context

   # Création des biphones
   ####################################
   setenv HMM_PREC      mono_1
   setenv HMM           droit_1
   call biphones

   # Classification des contextes
   ####################################
   setenv HMM_PREC      $HMM
   setenv HMM           droit_clas
   call classif

   # Augmentation du nombre de gaussiennes
   ####################################
   foreach i (1 2 3 4 5 6 7 8 9)
      setenv HMM_PREC   $HMM
      setenv HMM        droit_mu$i
      call mixup
   end
exit

#########################################################################
# Initialisation
#========================================================================
init:
   # Liste des phrases 'si' et 'sx' de la base TIMIT sous la forme:
   #    set(train/test) dialect(dr?) speaker(f*/m*) sentence(si*/sx*)
   ####################################
   if (! -e $LISTE) then
      foreach dr ($TIMIT/{train,test}/dr*)
         ls $dr/*/s[ix]*.wav | tr './' ' ' | \
            awk '{print $(NF-4),$(NF-3),$(NF-2),$(NF-1)}' >>$LISTE
      end
   endif

   # Calcul des paramètres par HCode
   ####################################
   if (! -e $PARAM) then
      mkdir $PARAM
      foreach f ("`cat $LISTE`")
         # Lit chaque ligne de la liste comme un tableau de dimension 4
```

```
            set f=($f)
            set wav = $TIMIT/$f[1]/$f[2]/$f[3]/$f[4].wav
            set par = $PARAM/$f[3].$f[4].mfcc

            # Pré-emphasage 0,97; Fenêtrage Hamming 25 ms pas 10 ms;
            # 24 filtres en échelle Mel -> 12 coefficients MFCC + énergie
            HCode -m -n 12 -p 24 -e -f 10.0 -w 25.0 -h -k 0.97 -F NIST $wav $par
         end
      endif

      # Traduction des étiquettes par HLEd
      ####################################
      if (! -e $LABEL) then
         mkdir $LABEL
         setenv LED $HBIN/HTKEtc/Scripts/KFLset.led
         foreach f ("`cat $LISTE`")
            set f=($f)
            set phn = $TIMIT/$f[1]/$f[2]/$f[3]/$f[4].phn
            set lab = $LABEL/$f[3].$f[4].lab
            cat $phn >$lab
            HLEd -G TIMIT -L $LABEL $LED $lab
         end
      endif

      # Liste des fichiers d'apprentissage et de test
      ####################################
      awk '/train/ {printf("'$PARAM'/%s.%s.mfcc\n",$3,$4)}' $LISTE >$TRAIN
      awk '/test/ {printf("'$PARAM'/%s.%s.mfcc\n",$3,$4)}' $LISTE >$TEST

      # Calcul du bigramme phonétique par HLStats
      ####################################
      if (! -e $BIGRAM) then
         awk '/train/ {printf("'$LABEL'/%s.%s.lab\n",$3,$4)}' $LISTE >tmp_lst
         HLStats -b $BIGRAM -S tmp_lst $PHONES
         rm tmp_lst
      endif

      # Crée un prototype de Bakis à 3 états
      ####################################
      if (! -e $PROTO) then
         # HMM a trois états sans saut de l'état central
         cp $HBIN/HTKEtc/Prototypes/med26dm1 $PROTO
         echo "AT 2 4 0.1 {*.transP}"       > tmp_hed
         echo $PROTO                        > tmp_lst
         HHEd tmp_hed tmp_lst
         rm tmp_hed tmp_lst
      endif
   exit

   ########################################################################
   # Initialisation pour les modèles en contexte
   #=======================================================================
   init_context:
      # Etiquettes en contexte
      ####################################
      if (! -e $NEWLAB) then
         echo G            > tmp_led
         mkdir $NEWLAB
         awk '/train/ {printf("'$LABEL'/%s.%s.lab\n",$3,$4)}' $LISTE >tmp_train
         HLEd -n $PHYSIC -L $NEWLAB- S tmp_train tmp_led
         awk '/test/ {printf("'$LABEL'/%s.%s.lab\n",$3,$4)}' $LISTE >tmp_test
         HLEd -L $NEWLAB -S tmp_test tmp_led
         rm tmp_led tmp_train tmp_test
      endif

      # Réseau pour le décodage
      ####################################
      if (! -e $NET) then
         awk '{if(NR==1) printf("$phn = "); else printf("\n| "); printf($1)} \
```

```
            END {printf(" ;\n( << $phn >> )\n");}' $PHYSIC          >  $NET
      endif
   exit

   ########################################################################
   # Estimation de monophones en modèles isolés
   #=======================================================================
   monophones:
      if (! -d $HMM) then
         mkdir $HMM

         # Première estimation par HInit
         if ($MIX == 1) then
            foreach ph (`cat $PHONES`)
               HInit -T 1 -i 20 -l $ph -L $LABEL -S $TRAIN -o $HMM/$ph $PROTO
            end

         # ou augmentation du nombre de gaussiennes
         else
            echo MU $MIX "{*.state[2-4].mix}" > tmp_hed
            HHEd -d $HMM_PREC -e $HMM tmp_hed $PHONES
            rm tmp_hed
         endif

         # puis ré-estimation par HRest
         foreach ph (`cat $PHONES`)
            HRest -T 1 -i 30 -l $ph -L $LABEL -S $TRAIN -o $HMM/$ph $HMM/$ph
         end
      endif
   exit

   ########################################################################
   # Teste les modèles en reconnaissance
   #=======================================================================
   reco:
      if (! $?RECO) setenv RECO reco_$HMM
      if (! -d $RECO) then
         mkdir $RECO
         if (! $?NEWPHON) setenv NEWPHON $PHONES
         if (! $?SCALE) setenv SCALE 1.0
         HVite -t 30.0 -T 1 -b $BIGRAM -s $SCALE -d $HMM -L $RECO -S $TEST \
            $NEWPHON $NET

         HResults -s -e sil cl -e sil vcl -e sil epi -e l el -e n en -e sh zh \
            -e aa ao -e ih ix -e ah ax -L $LABEL $PHONES $RECO/*.rec
      endif
   exit

   ########################################################################
   # Réestimation en modèles connectés
   #=======================================================================
   reest:
      repeat 2
         HERest -T 1 -t 500.0 -d $HMM -e $HMM -s tmp_stat -L $NEWLAB \
            -S $TRAIN $NEWPHON
      end
      call reco
   exit

   ########################################################################
   # Biphones a partir des monophones
   #=======================================================================
   biphones:
      if (! -d $HMM) mkdir $HMM

      # Script pour lier les états
      echo UF macros                                       > tmp_tie
      echo CL \"$PHYSIC\"                                  >>tmp_tie
      foreach ph (`cat $PHONES`)
```

```
     echo "TI ${ph}L {($ph,$ph+*).state[2]}"             >>tmp_tie
     echo "TI ${ph}C {($ph,$ph+*).state[3]}"             >>tmp_tie
     echo "TI ${ph}P {($ph,$ph+*).transP}"               >>tmp_tie
   end
   HHEd -d $HMM_PREC -e $HMM tmp_tie $PHONES
   rm tmp_tie

   # Réapprentissage et test des modèles
   setenv NEWPHON $PHYSIC
   call reest

   # Enregistre les statistiques d'occupation des états pour la classification
   mv tmp_stat $OCC
exit

#######################################################################
# Biphones généralisés regroupés par contexte
#======================================================================
classif:
   if (! -d $HMM) mkdir $HMM

   # Script pour classifier les états
   echo RO $RO \"$OCC\"                                  >tmp_clus
   foreach ph (`cat $PHONES`)
      echo TC $TC "${ph}R {($ph,$ph+*).state[4]}"        >>tmp_clus
   end
   echo CO \"$LOGIC\"                                    >>tmp_clus
   HHEd -d $HMM_PREC -e $HMM tmp_clus $PHYSIC
   rm tmp_clus

   # Réapprentissage et test des modèles
   setenv NEWPHON $LOGIC
   call reest
exit

#######################################################################
# Augmentation du nombre de gaussiennes
#======================================================================
mixup:
   if (! -d $HMM) mkdir $HMM

   # Procédure standard avec un nombre fixe de gaussiennes
   #   echo MU $MIX "{*.state[2-4].mix}"                 >tmp_mu
   #   HHEd -d $HMM_PREC -e $HMM tmp_mu $PHYSIC
   #   rm tmp_mu

   # Procédure automatique utilisant la statistique du chi-carré
   ERestUp -r 300 -a 32 -k 3 -T 1 -t 500.0 -d $HMM_PREC -e $HMM \
      -L $NEWLAB -S $TRAIN $LOGIC

   # Reapprentissage et test des modeles
   setenv NEWPHON $LOGIC
   call reest
exit
```

TABLE DES ABRÉVIATIONS

ACP	Analyse en Composantes Principales
ALD	Analyse Linéaire Discriminante
API	Alphabet Phonétique International
APS	Appariement de Pics Spectraux
AR	Auto-Régressif
ARMA	Auto-Régressif à Moyenne Ajustée
ARV	Auto-Régressif Vectoriel
AT&T	American Telephone and Telegraph
BACC	Bark Auditory Cepstral Coefficient (coefficient cepstral en échelle Bark)
BBN	Bolt, Beranek & Newman
CHA	Classification Hiérarchique Ascendante
CMU	Carnegie Mellon University
CNET	Centre National d'Etudes des Télécommunications
CUED	Cambridge University Engineering Department
CQ	Critère de Qualité
DAP	Décodage Acoustico-Phonétique
DT	Décomposition Temporelle
DTW	Dynamic Time Warping (alignement temporel dynamique)
ENST	Ecole Nationale Supérieure des Télécommunications
FFT	Fast Fourier Transform (transformée de Fourier rapide)
HMM	Hidden Markov Models (modèles de Markov cachés)
HTK	HMM Toolkit (boîte à outils pour modèles de Markov cachés)
IBM	International Business Machines
ICASSP	International Conference on Acoustics, Speech and Signal Processing
IEEE	Institute of Electrical and Electronics Engineers
JEP	Journées d'Etudes sur la Parole
LAFORIA	Laboratoire de Reconnaissance de Formes et d'Intelligence Artificielle
LIMSI	Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
LPCC	Linear Predictive Cepstral Coefficient
MAP	Maximum A Posteriori
MFCC	Mel Frequency Cepstral Coefficient (coefficient cepstral en échelle Mel)

MIT	Massachussets Institute of Technology
MLE	Maximum Likelihood Estimation (estimation par maximum de vraisemblance)
MMI	Maximum Mutual Information
PLP	Perceptual Linear Prediction (prédiction linéaire perceptive)
PMC	Perceptron Multi-Couches
PPV	Plus Proches Voisins
QV	Quantification Vectorielle
RAP	Reconnaissance Automatique de la Parole
RM	Ruptures de Modèles
SB	Système de Base
TDNN	Time Delay Neural Network (réseau de neurones à délai)
TSI	Taux de Segments Insérés
TSS	Taux de Sur-Segmentation

TABLE DES NOTATIONS

α	facteur de pré-accentuation
$\alpha_t(i)$, $\beta_t(i)$	variables avant et arrière de Baum
a_i, A_i	coefficients d'un modèle AR ou ARV
$A = \{a_{ij}\}$	probabilités de transition d'un modèle
$B = \{b_j(k)\}$	probabilités d'émission d'un modèle
$c(v,v')$	nombre d'appariement de deux symboles
C	classe de locuteurs
C_k	coefficient cepstral
δ	biais de l'entropie mutuelle
δ_1, δ_2, δ_3	pondérations du trigramme
δ_{APS}	distorsion entre pics spectraux
ΔC, ΔE	coefficients différentiels
d	dimension des vecteurs de paramètres
D	durée maximale d'émission d'un état
D_{IC}	seuil inter-classes
$D_{\max}$	tolérance sur le test du χ^2
$D()$	distance
d_E	distance euclidienne
d_I, d_{IV}	mesure d'Itakura, d'Itakura vectorielle
d_M	distance de Mahalanobis
d_P	distance pondérée
$\mathcal{E}$	espace de représentation continu
ε	tolérance sur la segmentation
E_0, E_i	énergie d'une trame, en sortie d'un banc de filtres
E_L	ensemble de locuteurs
E_m, E_M	ensemble des mots du vocabulaire, des phrases du langage
E_Q	ensemble des chemins possibles
e_n, E_n	bruit blanc gaussien scalaire ou vectoriel
$\phi_k(t)$	fonction d'interpolation
$\Phi(t,i)$	probabilité partielle du chemin optimal
F	nombre de filtres
$F(t)$, F_0	fonction de pénalité
f	nombre de flux
$f(\varphi_1 \ldots \varphi_n)$	fréquence d'une suite de phonèmes
G	nombre de gaussiennes
g_k	poids d'une gaussienne
γ_i	nombre moyen de transitions vers l'état i
h_c	nombre d'échantillons dans un intervalle
h_n	échelle du noyau de Parzen
I_c	intervalle
φ, φ^g, φ_d	phonème, phonème en contexte gauche g, en contexte droit d
κ	nombre de classes conservées
$K(\ )$	noyau de Parzen
λ, Λ	modèle de Markov caché, modèle composite
λ	seuil de détection des ruptures
l^ϕ	biais phonétique
L	locuteur
L_r	distance de Minkovski
m	mot
$m(t,j)$	mémoire des meilleurs chemins du décodage
M	suite de mots
M_0, M_1	modèles AR
μ	vecteur moyenne

n	nombre d'éléments d'un ensemble
n_Δ, $n_{\Delta\Delta}$	nombre de trames pour le calcul des coefficients différentiels
n_λ,	nombre de modèles
n_L, n_C	nombre de locuteurs, de classes
n_m	nombre de mots
n_V	nombre de symboles
N	nombre d'états d'une machine
N_{omi}, N_{ide}, N_{sub}, N_{ins}, N_{ref}	nombre de segments omis, identifiés, substitués, inserés, de référence
$\mathcal{N}(o,\mu,\Sigma)$	loi normale
$O = (o_1 \ldots o_T)$	suite d'observations acoustiques
π_c	pondération attribuée à la classe c
p	ordre d'une analyse auto-régressive
$p_i(\tau)$	probabilité de durée d'émission d'un état
$P()$	probabilité
P_+, P_-	bornes de l'intervalle de confiance
$Q = (q_1 \ldots q_T)$	suite d'états dans un modèle
r_{ij}, r_k	coefficient de corrélation, d'auto-corrélation
R	matrice de corrélation ou d'auto-corrélation
R_m	référence du mot m
$\mathbb{R}$	ensemble des réels
s	facteur d'échelle du bigramme
s_n	échantillon de signal
σ^2, σ_{ij}	variance, covariance
$S = \{s_i\}$	ensemble des états d'une machine
S_k	spectre du signal
Σ	matrice de covariance
τ, $\overline{\tau}$	durée d'émission par un état, durée moyenne
t, T	instant, durée d'une suite d'observations
T_{ide}, T_{omi}, T_{sub}, T_{ins}, T_{rec}	taux d'identification, d'omission, substitution, insertions, reconnaissance
T_ϕ	matrice de transformation spectrale pour la classe ϕ
Tr_{min}	nombre minimal de trames par gaussienne
u_k	cible spectrale
$V = \{v_k\}$	ensemble des symboles observables
w_n	fenêtre de pondération
w_k, W_k	entropie mutuelle de modèles, entropie cumulée
$\omega_t(i)$	probabilité d'une transition vers l'état i à l'instant t
x	échantillon réel
X, Y	vecteur de paramètres
Z_n	statistique suivant un χ^2 calculée sur n valeurs

TABLE DES FIGURES

TABLE DES TABLEAUX

TABLE DES MATIÈRES