THÈSE
présentée pour obtenir le titre de Docteur
de l'Université Pierre et Marie Curie

Spécialité: INFORMATIQUE

Sujet :

ESTIMATION DE PROBABILITE NON-PARAMETRIQUE POUR LA RECONNAISSANCE MARKOVIENNE DE LA PAROLE

Présentée par :

Fabrice LEFEVRE

Soutenue le 17 janvier 2000 devant le jury composé de :

Renato DE MORI	Président
Régine ANDRÉ-OBRECHT Laurent MICLET	Rapporteurs
Thierry DENOEUX Jean-Yves JAFFRAY	Examinateurs
Marie-José CARATY Claude MONTACIÉ	Directeurs de thèse

Résumé

Les modèles de Markov cachés sont à l'origine de la majorité des avancées récentes en reconnaissance de la parole continue. Ces modèles gèrent les distorsions temporelles du signal de parole en s'appuyant sur des densités de probabilité pour modéliser les distorsions en fréquence. Nous proposons d'améliorer leur performance à l'aide d'un estimateur de densité de probabilité très efficace : l'estimateur des K plus proches voisins. La règle de décision déduite de cet estimateur présente l'intérêt d'avoir une erreur asymptotique de classification faible (proche de l'erreur optimale de Bayes) et qui diminue avec l'augmentation de K. De plus, il est discriminant par construction.

L'estimateur des K plus proches voisins est évalué comme opérateur de reconnaissance statique de spectres à court-terme de parole. Ses performances se révèlent supérieures à celles de l'estimateur de l'état de l'art à base de sommes pondérées de gaussiennes. Les adaptations nécessaires à son intégration dans un système de reconnaissance markovien sont ensuite développées. L'introduction de coefficients d'appartenance mesurant le degré d'association entre un vecteur de référence et un état markovien permet le développement d'un protocole d'apprentissage optimal basé sur une version de l'algorithme EM (*Expectation-Maximisation*) dont nous montrons la convergence selon le critère du maximum de vraisemblance. Les options de la procédure d'apprentissage sont évaluées afin d'obtenir un protocole optimal.

Cette étude a abouti à la construction d'un système markovien K plus proches voisins de base qui est évalué sur la base de données TIMIT. Ses performances sont comparées à celles d'un système markovien fondé sur l'estimateur gaussien. Deux approches permettant un apport de connaissance dans le système sont finalement traitées : l'introduction d'une information temporelle dans l'espace de représentation et l'adaptation des références.

Les performances du système markovien K plus proches voisins sont encourageantes. Des recherches complémentaires devront être menées afin de les élever au niveau de l'état de l'art. En outre, l'estimateur des K plus proches voisins en offrant une alternative à l'estimateur de l'état de l'art permet, par comparaison, de mieux mettre en évidence l'influence réelle de l'estimateur de probabilité dans les systèmes markoviens de reconnaissance de la parole.

Mots –clés : reconnaissance de la parole continue, modèles de Markov cachés, K plus proches voisins, algorithme EM

Sommaire

INTRODUCTION

Les campagnes d'évaluation de l'ARPA (*Advanced Research Projects Agency*) constituent une bonne référence pour mesurer les progrès réalisés ces dix dernières années en reconnaissance automatique de la parole (RAP) continue. De 1987 à 1992, l'évaluation portait sur la reconnaissance multilocuteur de la parole lue avec un vocabulaire fermé de 1000 mots (*Ressource Management*). Les taux d'erreur en mots ont été réduits jusqu'à 4%. Cinq ans plus tard, en 1997, elle portait sur la transcription d'émissions d'information radio-télédiffusées (*Broadcast News*) avec des vocabulaires de 65.000 mots dans des conditions allant de la parole propre (présentateur enregistré en studio) jusqu'à de la parole fortement dégradée (interventions téléphoniques, interview en extérieur…). Les taux d'erreurs globaux sur cette tâche se situent actuellement autour de 20%. Parallèlement, les ressources utilisées pour la mise au point des systèmes de reconnaissance ont été en constante évolution : elle sont passées de quelques heures à plusieurs centaines d'heures de parole pour les données acoustiques et de quelques dizaines à plusieurs centaines de millions de mots pour les corpus linguistiques. La RAP continue a ainsi atteint une certaine maturité qui se concrétise, notamment, par l'existence de systèmes de dictée vocale commerciaux destinés au grand public (*ViaVoice* d'IBM, *NaturallySpeaking* de Dragon…).

Toutefois, ces bons résultats doivent être relativisés. Si nous considérons, par exemple, la transcription d'émissions télévisuelles, une étude réalisée dans le cadre du projet *Informedia* du CMU a montré que les taux de transcription correcte peuvent varier de 80% pour un présentateur enregistré en studio à 25% pour un documentaire [Hauptmann, 1995]. L'influence sur la reconnaissance de certaines contraintes comme la qualité du signal ou la complexité du langage reste importante. Afin d'y remédier des améliorations doivent être apportées aux différents composants des systèmes de reconnaissance.

Pour la réalisation des systèmes de RAP continue, trois axes de recherche principaux peuvent être dégagés :

- *l'analyse du signal* a pour but d'extraire l'information utile du signal de parole. Pour cela, l'influence des phénomènes parasites (bruit, musique de fond ou encore influence du locuteur) doit être diminuée. Les analyses les plus courantes sont le calcul du spectre à court terme de segments de parole, le calcul de l'énergie ou encore la détermination du taux de passages par zéro. Elle peut aussi servir à la détection d'informations particulières comme les ruptures de stationnarité ;

- *la reconnaissance des formes* (RdF) permet la construction d'un opérateur de reconnaissance qui utilise les informations issues de l'analyse du signal pour transformer le continuum acoustique en une chaîne de symboles. Les symboles en RAP sont les mots ou les unités sub-lexicales (phones, phonèmes, syllabes…) ;
- *l'intelligence artificielle* (IA) apporte les connaissances sur le domaine symbolique. Elle intervient classiquement en interaction avec le processus de RdF (modèles de langages, gestion du dialogue…) mais des post-traitements peuvent être envisagés (analyse sémantique, détection de sujet…).

La pluridisciplinarité de la RAP ne se limite évidemment pas aux domaines précités. Pour être plus exhaustif, il faut leur adjoindre l'acoustique, la phonétique, la linguistique et même les statistiques, les probabilités, la théorie de l'information ou encore l'algorithmique. Chacun de ces axes possède ses propres problématiques. La recherche d'un mode d'interaction le plus efficace possible constitue un des enjeux majeurs de la RAP. Aussi, l'étude d'un point particulier d'un système de RAP doit se faire en relation avec l'ensemble des autres composantes. Au cours de cette étude, nous nous intéresserons plus particulièrement à la RdF et à la construction de l'opérateur de reconnaissance.

Actuellement, les opérateurs de RAP continue les plus performants reposent sur les méthodes statistiques et sont mis en œuvre au travers des modèles de Markov cachés (*Hidden Markov Model*, HMM). L'introduction des HMM en parole a constitué un élément fondamental permettant les avancées de ces dernières années. De nombreuses voies sont encore ouvertes pour l'amélioration de leur qualité. Parmi celles-ci, le renforcement de la discrimination entre modèles nous paraît l'une des plus prometteuses.

Les principales solutions proposées pour compenser le manque de discrimination dans les modèles de Markov interviennent dans la phase d'apprentissage des modèles. Elles se révèlent d'une utilisation complexe et leur coût de calcul est élevé. Une alternative consiste à introduire la discrimination localement dans la définition des modèles. Parmi les méthodes envisagées, l'utilisation des réseaux de neurones comme estimateur de probabilité discriminant s'est révélée relativement efficace bien que, pareillement, coûteuse et complexe à mettre en œuvre. La solution, proposée dans cette étude, est l'utilisation d'un estimateur de probabilité simple, efficace et discriminant : l'estimateur des K plus proches voisins (K-ppv). La règle de décision déduite de cet estimateur a une erreur de classification asymptotique faible (au plus égale à deux fois l'erreur bayesienne) et qui décroît lorsque K croît. La principale limitation à son emploi, un coût de calcul important, a pu être levée par le développement d'algorithmes rapides. Les travaux présentés dans cette étude constituent la première introduction de l'estimateur des K-ppv dans un système markovien de RAP continue.

L'application de la modélisation markovienne à la RAP continue est présentée dans le chapitre 1. Le cadre général de la modélisation probabiliste à l'aide des HMM est introduit. Les modèles markoviens sont des outils offrant une grande souplesse d'utilisation. Trois groupes principaux de techniques permettant leur mise en œuvre ont été distingués : la définition des modèles, les techniques d'apprentissage et les techniques de décodage. Pour chacun de ces groupes, les techniques intervenant dans la réalisation des systèmes de l'état de l'art sont présentées et discutées.

Parmi les techniques performantes, le renforcement de la discrimination entre modèles est abordé plus précisément dans le chapitre 2. Les techniques de l'état de l'art proposent de compenser le manque de discrimination par une réestimation des paramètres des modèles à l'aide d'un critère discriminant. Les défauts de ces techniques sont une mise en œuvre complexe et l'absence de garantie de convergence de la procédure d'apprentissage. Il est envisageable d'obtenir la discrimination globale des modèles de Markov par la discrimination locale des vecteurs acoustiques au niveau des états markoviens. Différentes approches sont présentées : transformations linéaires de l'espace de représentation, réseaux de neurones et estimateurs de probabilité non-paramétriques. L'estimateur non paramétrique des K-ppv offre une alternative qui n'a jamais été étudiée en RAP continue. Nous proposons de l'introduire comme estimateur de probabilité des états des HMM. Au préalable, l'évaluation de ses performances sur une tâche d'identification locale de la base de données TIMIT est réalisée. Les résultats sont comparés à ceux de l'estimateur de l'état de l'art à base de modèles de sommes pondérées de gaussiennes (*Gaussian Mixture Model*, GMM).

L'estimateur des K-ppv est adapté pour être introduit dans le formalisme markovien. Les vecteurs acoustiques de référence, parmi lesquels sont recherchés les K-ppv, doivent être associés aux états des HMM et non plus seulement à une classe phonétique. Pour cela, les coefficients d'appartenance mesurant le degré d'association entre un vecteur et un état markovien sont introduits dans le chapitre 3. Deux formules de réestimation (locale et propagée) sont présentées et testées pour adapter l'algorithme de Baum-Welch à l'apprentissage des coefficients d'appartenance. Les options de la procédure d'apprentissage sont discutées : le choix du type de probabilité conditionnelle utilisée (l'estimateur des K-ppv permettant l'expression aussi simplement des vraisemblances que des probabilités *a posteriori*), la gestion des vecteurs d'apprentissage de probabilités nulles, différents modes d'initialisation des coefficients d'appartenance et finalement, l'opportunité de la réestimation connectée. Cette étude aboutit à la définition d'un système de base HMM/K-ppv reposant sur un apprentissage itératif de type EM dont nous montrons la convergence. Ce système est comparé à un système de base HMM/GMM sur un décodage acoustico-phonétique (DAP) de la base de données TIMIT.

Dans le chapitre 4, des traitements permettant un apport de connaissance dans le système HMM/K-ppv sont finalement proposés afin d'améliorer ses performances. Les traitements proposés sont issus de l'état de l'art des systèmes HMM/GMM : adaptation de l'espace de représentation (coefficients delta) et adaptation des références (modélisations contextuelles ou par macro-classes). Des alternatives, concernant principalement l'extension des techniques précédentes à des versions non-supervisées, sont parallèlement étudiées.

Bien que les principes de l'estimation par les K-ppv aient été énoncés il y a plus de quarante ans, son application sur des problèmes de grande dimension a été retardée par un coût de calcul élevé. Son étude sur une tâche aussi complexe que la RAP continue constitue en ce sens une avancée importante. Les résultats obtenus confirment l'intérêt de l'approche. De plus, le cadre formel que nous avons développé pour la construction du système HMM/K-ppv constitue une référence à partir de laquelle pourront émerger des solutions nouvelles permettant d'élever ses performances au niveau de l'état de l'art.

1. SYSTEME MARKOVIEN DE RECONNAISSANCE DE LA PAROLE CONTINUE

Depuis une dizaine d'années, les modèles de Markov cachés sont devenus incontournables dans le domaine de la RAP continue. Ils fournissent un cadre simple et complet pour la mise en œuvre de l'approche statistique de la reconnaissance.

La modélisation des séquences de vecteurs acoustiques suppose la prise en compte de deux niveaux de distorsions imbriqués. Entre deux réalisations acoustiques d'une même unité symbolique, les vecteurs spectraux se déforment dans l'espace de représentation et les séquences de vecteurs spectraux subissent des transformations temporelles complexes. Dans l'approche statistique de la reconnaissance, le premier niveau est pris en charge par les estimateurs de probabilité qui permettent de déterminer les parties de l'espace de représentation les plus probables associées aux classes définies *a priori*. La gestion des distorsions temporelles est obtenue au moyen de la programmation dynamique. Les modèles de Markov permettent une représentation *ad hoc* de ce processus doublement stochastique.

Un des avantages majeurs de la modélisation markovienne est d'être auto-organisatrice. Sa mise en œuvre est presque entièrement automatisée au travers de deux algorithmes qui permettent l'un le décodage d'une séquence acoustique selon le principe du maximum *a posteriori* (décision bayesienne) et l'autre l'apprentissage itératif convergent d'un modèle à partir des données selon le critère du maximum de vraisemblance.

Les différents choix relatifs à la mise en œuvre d'un système de reconnaissance markovien sont présentés. Trois catégories de techniques ont été distinguées en fonction de la problématique qu'elles visent à résoudre :

- *la définition des modèles* : le choix de l'unité de modélisation, la définition de la topologie et le choix de l'estimateur de probabilité ;
- *l'apprentissage des modèles* : la réestimation connectée, le partage des paramètres et l'adaptation au locuteur[1] ;
- *le décodage de modèles connectés* : les modèles de langage, la définition du réseau de décodage et de la structure de recherche, le choix de l'algorithme de décodage, le

---

[1] L'adaptation au canal de transmission n'est pas abordée dans cette étude.

décodage par passes multiples et, finalement, l'intégration d'un mécanisme de contrôle temporel.

A partir des choix opérés sur ces trois niveaux, il est possible de réaliser un système de RAP continue.

1.1. Les modèles de Markov cachés, opérateurs de reconnaissance segmentale

Introduits en traitement de la parole depuis les années 75 [Baker, 1975; Jelinek, 1976], les modèles de Markov cachés ont pris depuis une importance prépondérante en RAP. Devenu un outil incontournable de la RAP, leur emploi n'y est cependant pas limité. Ainsi, les méthodes décrites dans ce chapitre sont généralisables à d'autres problèmes de reconnaissance segmentale. On peut citer, entre autre, la reconnaissance de caractères manuscrits pour laquelle les HMM constituent un outil très utilisé (voir, par exemple [Bunke, Roth *et al.*, 1995]).

En RAP, les HMM sont des modèles de production. Ils modélisent les variations statistiques dans le débit de parole, la prononciation et les différences entre locuteurs. Les HMM ont l'avantage d'offrir des procédures automatiques pour le décodage et l'apprentissage. Leur utilisation en RAP s'intègre dans le cadre plus général des méthodes statistiques de reconnaissance.

1.1.1. Méthodes statistiques de reconnaissance de la parole

Les méthodes statistiques de reconnaissance reposent sur une formalisation simple issue de la théorie de l'information et qui est aujourd'hui classique pour décomposer le problème de la RAP continue [Jelinek, 1976].

Dans le cadre de ces méthodes, le continuum de parole est transformé en une séquence de vecteurs acoustiques que nous appellerons aussi observations ou trames. Chacun de ces vecteurs est typiquement une représentation compacte du spectre à court terme du signal pour une période de l'ordre de la dizaine de milli-secondes. Ainsi, une phrase d'une dizaine de mots prononcée en quelques secondes sera représentée par quelques centaines de vecteurs acoustiques. Le choix de l'espace de représentation, qui définit la transformation appliquée au signal, est une problématique qui n'est pas abordée dans notre étude. Une présentation rapide des différents espaces de représentation utilisés en RAP est donnée dans l'Annexe A ; ainsi que les motivations qui nous ont conduit à retenir l'espace des MFCC (*Mel Frequency Cepstral Coefficients*) comme espace de représentation dans notre étude.

Une fois le signal convertit en une séquence de vecteurs acoustiques O, on cherche la phrase $\tilde{M}$ (suite de mots) la plus probable parmi toutes les phrases M possibles soit :

$$\tilde{M} = \arg\max_{M} P(M/O) \tag{1-1}$$

En utilisant la règle de Bayes, il est possible d'écrire la probabilité qu'une phrase *M* corresponde aux observations comme :

$$P(M/O) = \frac{P(O/M)P(M)}{P(O)} \tag{1-2}$$

Puisque *P(O)* ne dépend pas de *M*, l'équation (1-1) est équivalente à :

$$\tilde{M} = \arg\max_{M} P(O/M)P(M) \tag{1-3}$$

où

- $P(O/M)$, la vraisemblance des observations *O*, étant donnée la phrase *M*, est estimée par une modélisation des éléments du vocabulaire ;
- *P(M)*, la probabilité *a priori* de la phrase *M*, est estimée par un modèle linguistique qui peut comprendre des niveaux lexicaux, syntaxiques et sémantiques.

Le problème de reconnaissance est donc transféré du maximum de la probabilité *a posteriori* au maximum de la probabilité conjointe. L'approche statistique permet d'intégrer les niveaux numériques (analyse du signal) et symboliques (IA) dans un seul processus de décision. Ces niveaux sont classiquement représentés par les HMM.

1.1.2. Présentation des HMM

Les HMM offrent une réponse statistique au problème de la reconnaissance des formes segmentale. Leur intérêt est de fournir un cadre formel rigoureux pour le développement de techniques adaptées à la RAP.

1.1.2.1. Reconnaissance segmentale

Un des problèmes fondamentaux de la reconnaissance des formes consiste à trouver un espace de représentation et une métrique associée pour lesquels les représentations des données d’une même classe sont plus proches les unes des autres que n’importe quelle représentation d’une donnée d’une autre classe.

Dans le cadre de la reconnaissance de la parole, l'espace de représentation pourra être l’espace des MFCC associés à une distance euclidienne. Cette approche nous permet donc de reconnaître des vecteurs acoustiques représentant de courts segments de parole. Or, les formes que l'on cherche à reconnaître en RAP ne sont pas les vecteurs acoustiques mais leur concaténation en séquences ayant une interprétation linguistique (phrases, mots, syllabes, phonèmes…). L'opérateur de reconnaissance doit donc être capable de discriminer des vecteurs acoustiques mais aussi les séquences de vecteurs acoustiques. L'opérateur élémentaire de reconnaissance des formes statistique fondé sur l'estimation des fonctions de densité de probabilité (fdp) de chaque classe doit donc être intégré dans un opérateur plus complexe.

Fondamentalement, le problème de reconnaissance dans un espace de représentation isomorphe à $\Re^d$ (où d est la dimension de l'espace considéré) se transforme en un problème de reconnaissance dans l'espace $\left(\Re^d\right)^T$ (où T est la longueur du segment à reconnaître). La longueur T pouvant varier, cet espace est éminemment complexe. Une des premières solutions pour résoudre cette difficulté dans le cadre de la reconnaissance de mots isolés fut la programmation dynamique (*Dynamic Time Warping*, DTW) [Vintsyuk, 1971].

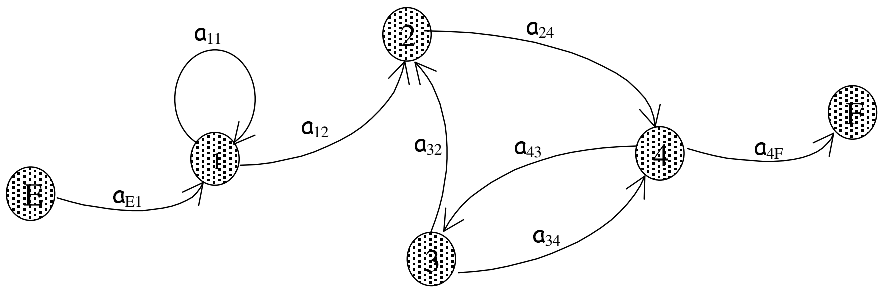
La DTW identifie la séquence d'observations représentant un mot ou une phrase en la considérant globalement et en la comparant avec des références enregistrées. Ainsi, pour chaque mot ou phrase qui peut être distingué, il existe une ou plusieurs occurrences enregistrées. Lorsqu'une entité est présentée, elle est comparée aux références au sens d'une distance donnée. Or, la durée d'un mot étant variable d'une prononciation à l'autre et les déformations n'étant pas linéaires en fonction du temps, la comparaison doit rapprocher au mieux les séquences. Afin d'éviter l'explosion combinatoire de la recherche exhaustive de toutes les déformations possibles, le principe d'optimalité de Bellman [Bellman, 1957] est appliqué avec des contraintes sur l'extension des sous-solutions [Sakoe & Chiba, 1978]. L'alignement optimal est alors obtenu en exploitant de façon récursive le fait qu'un chemin optimal est l'extension d'un sous-chemin optimal.

Des généralisations de la DTW (cf. §1.4.3.2) ont permis d'agrandir son champ d'application à la reconnaissance de mots enchaînés [Sakoe, 1979 ; Myers & Rabiner, 1981 ; Bridle, Brown *et al.*, 1982]. Son application restait pourtant mal aisée du fait de l'obligation de conserver dans le système un exemplaire de chaque prononciation type. Les HMM ont apporté une réponse satisfaisante à ce problème en permettant de modéliser les déformations type de façon probabiliste au sein d'un unique modèle.

1.1.2.2. Formalisme markovien

Un modèle de Markov caché est une machine à états finis qui change d'état à chaque unité de temps. Quand au temps t, on accède à l'état j, un vecteur o_t est engendré avec la probabilité $b_j(o_t)$ avec b_j la fdp associée à l'état j. De même, la transition d'un état i à un état j est probabiliste et prend la valeur discrète a_{ij}. Ainsi, nous appellerons le doublet $\lambda=(A,B)$, avec $A=\{a_{ij}\}$ matrice de transition et $B=\{b_j\}$ vecteur des fdp associées aux états, un modèle de Markov caché. La matrice de transition permet de représenter λ sous forme de graphe orienté (voir Figure 1-1). Il est possible d'associer les fdp aux transitions et non plus aux états des HMM. Les deux approches peuvent être montrées équivalentes. Formellement, il manque pour définir complètement λ le vecteur des probabilités initiales des états.

Dans le cadre de la RAP continue, il est commun de définir deux états supplémentaires représentant l'entrée (E) et la sortie (F) du modèle. Ces états ont la particularité de ne pas émettre de vecteurs, aucune fdp ne leur est associée et leur franchissement est asynchrone. De plus, ils ne sont généralement pas considérés dans le décompte des états des modèles. Ainsi, le modèle de la Figure 1-1 est un modèle à 4 états. Ces états servent principalement à connecter les modèles : l'état de fin étant lié à l'état initial du modèle suivant par une transition asynchrone. Il est à noter que leur emploi fixe implicitement le vecteur des probabilités initiales : il vaut 1 pour l'état d'entrée et 0 pour tous les autres états.



$$A = \begin{vmatrix} 0 & a_{E1} & 0 & 0 & 0 & 0 \\ 0 & a_{11} & a_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{24} & 0 \\ 0 & 0 & a_{32} & 0 & a_{34} & 0 \\ 0 & 0 & 0 & a_{43} & 0 & a_{4F} \end{vmatrix} \qquad B = \begin{vmatrix} b_1 \\ \vdots \\ b_4 \end{vmatrix}$$

Figure 1-1 Exemple de modèle de Markov caché.

Une fois choisie la topologie d'un HMM, sa mise en œuvre va nécessiter de résoudre 3 problèmes :

- le problème de *l'évaluation de la vraisemblance* : considérant un modèle et une séquence d'observations, comment mesurer la probabilité que ce modèle ait engendré cette séquence ?
- le problème du *décodage* : considérant un modèle et une séquence d'observations, quelle est la suite d'états la plus probable ayant produit les observations ?
- le problème de *l'apprentissage* : considérant un modèle et un ensemble de séquences d'observations, quels sont les paramètres du modèle qui maximisent la probabilité que l'ensemble des séquences ait été engendré par le modèle ?

1.1.3. Mise en œuvre des HMM

La solution au problème de l'évaluation de la vraisemblance nous donne un moyen de mesurer l'adéquation d'une séquence d'observations à un modèle. Ainsi, nous pouvons décider du meilleur modèle selon la règle de Bayes. Résoudre le problème du décodage permettra de segmenter les séquences par la recherche de la séquence d'états de vraisemblance maximale. Enfin, l'apprentissage doit permettre d'adapter automatiquement un HMM à un ensemble de données particulier.

1.1.3.1. Evaluation de la vraisemblance d'un modèle

La probabilité qu'une séquence de vecteurs d'observation $O = o_1, o_2, .., o_T$ soit engendrée, étant donné un modèle λ, à travers le chemin (ou séquence d'états) Q est simplement calculée comme le produit des probabilités de transition et d'émission le long de Q. Un exemple d'alignement d'une séquence d'observations sur les états d'un HMM est donné dans la Figure 1-2 (page suivante) :

$$Q = q_0, q_1, q_2, \ldots, q_t, \ldots, q_T, q_{T+1} = E, 1, 1, \ldots, 2, \ldots, 3, F \tag{1-4}$$

$$P(O, Q / \lambda) = a_{E1} b_1(o_1) a_{11} b_1(o_2) \ldots b_2(o_t) \ldots b_3(o_T) a_{3F} \tag{1-5}$$

Cette probabilité est le produit de deux vraisemblances :

- la vraisemblance du chemin dans le modèle $P(Q / \lambda) = a_{q_0 q_1} \times \prod_{t=1}^{T} a_{q_t q_{t+1}}$
- la vraisemblance des observations le long de ce chemin $P(O / Q, \lambda) = \prod_{t=1}^{T} b_{q_t}(o_t)$

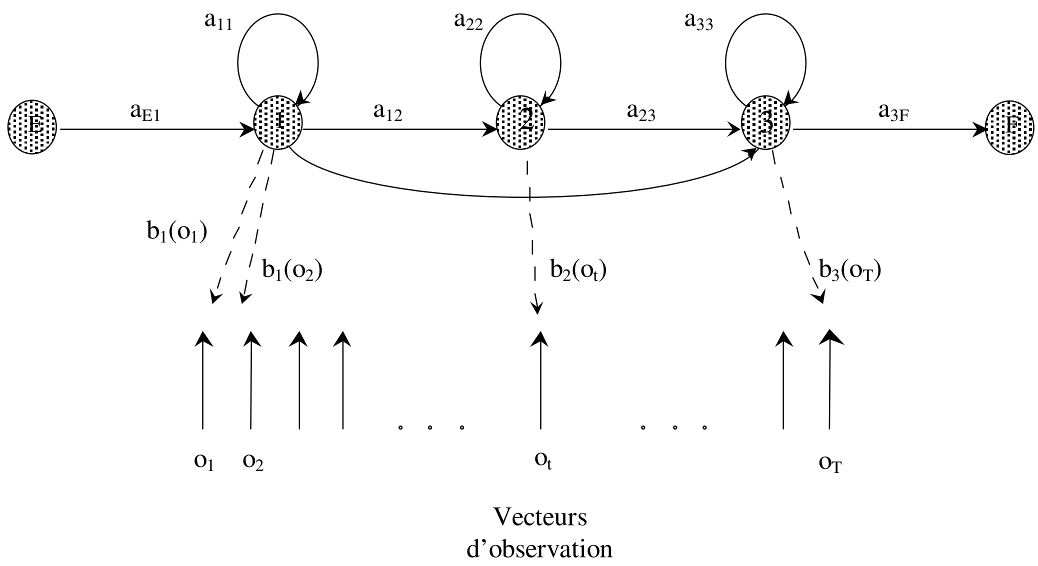


Figure 1-2 Génération d'une séquence de vecteurs d'observation par un HMM gauche-droit à 3 états.

En pratique, seule la séquence d'observation *O* est connue et la séquence d'états sous-jacente est cachée, d'où l'appellation modèles de Markov *cachés*. La séquence d'états *Q* étant inconnue, la vraisemblance est obtenue en sommant les probabilités (1-5) sur toutes les séquences possibles *Q* :

$$P(O/\lambda) = \sum_{Q} a_{q_0 q_1} \prod_{t=1}^{T} b_{q_t}(o_t)\, a_{q_t q_{t+1}} \tag{1-6}$$

Une alternative utile à l'équation (1-6) consiste à ne calculer que la séquence de probabilité maximale :

$$\hat{P}(O/\lambda) = \max_{Q} \left\{ a_{q_0 q_1} \prod_{t=1}^{T} b_{q_t}(o_t)\, a_{q_t q_{t+1}} \right\} \tag{1-7}$$

Les algorithmes de Baum-Welch et Viterbi offrent des procédures récursives pour calculer respectivement les équations (1-6) et (1-7). En pratique, la procédure de Viterbi sera utilisée lors du décodage car elle fournit, en plus de l'estimation de la vraisemblance du segment, un alignement des observations sur les états. L'apprentissage des modèles est traditionnellement initialisé par la procédure de Viterbi et réestimé par celle de Baum-Welch.

Deux variables sont introduites qui permettent l'estimation des probabilités directe et rétrograde. La variable directe α est définie comme la probabilité que les observations jusqu'à l'instant *t* ait été émises par le modèle λ à *N-2* états (les états E et F sont respectivement le 1er et le $N^{ième}$ état) et que l'état atteint en *t* soit *i* :

$$\alpha_t(i) = P(o_1 \ldots o_t, q_t = i | \lambda) \tag{1-8}$$

Le calcul de α est opéré par une récurrence sur le temps en partant de l'état initial E au temps 0 :

$$\alpha_1(i) = \begin{cases} 1 \text{ si } i = 1 \\ a_{1i} b_i(o_1) \text{ si } 1 < i < N \end{cases} \tag{1-9}$$

```
Pour t de 1 à T,
      Pour j de 2 à N-1,
```

$$\alpha_t(j) = \left[\sum_{i=2}^{N-1} \alpha_{t-1}(i) a_{ij} \right] b_j(o_t) \tag{1-10}$$

Pour arriver finalement, à l'état final F au temps T :

$$P(O/\lambda) = \alpha_T(N) = \sum_{i=2}^{N-1} \alpha_i(T) a_{iN} \tag{1-11}$$

Cette méthode développe l'ensemble des chemins du modèle en parallèle permettant ainsi une formulation simple de l'exploration de la matrice temps/états sous la contrainte des transitions autorisées entre états.

L'estimation directe est suffisante pour obtenir la probabilité définie par (1-6). Toutefois, l'apprentissage des modèles sera facilité par l'introduction de la probabilité rétrograde. Cette dernière est définie comme la probabilité que les trames suivant o_t aient été émises sachant que o_t a été émise par i :

$$\beta_t(i) = P(o_{t+1} \ldots o_T / q_t = i, \lambda) \tag{1-12}$$

Le calcul de β est opéré par une récurrence sur le temps en partant de l'état final F au temps T :

$$\beta_T(i) = \begin{cases} 1 \text{ si } i = N \\ a_{iN} \text{ si } 1 < i < N \end{cases} \tag{1-13}$$

```
Pour t de T à 1,
      Pour i de 2 à N-1,
```

$$\beta_t(i) = \sum_{j=2}^{N-1} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j) \tag{1-14}$$

Pour arriver finalement, à l'état initial E au temps 1 :

$$P(O/\lambda) = \beta_1(1) = \sum_{j=2}^{N-1} a_{1j} b_j(o_1) \beta_j(1) \tag{1-15}$$

L'utilisation conjointe des variables directe et rétrograde permet de calculer la probabilité de l'émission d'une trame sur un état par rapport à tous les chemins possibles :

$$P(O, q(o_t) = i/\lambda) = \alpha_t(i) \beta_t(i) \tag{1-16}$$

Cette probabilité sera utilisée lors de l'apprentissage des modèles par l'algorithme de Baum-Welch (voir §1.1.3.3). Ces variables permettent l'estimation de la vraisemblance de la séquence sur le modèle en tenant compte de l'ensemble des chemins. Nous avons déjà signalé qu'il peut être nécessaire de recourir à une estimation ne tenant compte que du chemin de vraisemblance maximale. C'est notamment le cas lorsqu'il est nécessaire de trouver la séquence d'état ayant le plus probablement généré le segment acoustique.

1.1.3.2. Le problème du décodage

Le décodage de Viterbi [Viterbi, 1967; Forney, 1973] permet de déterminer l'alignement optimal d'une forme acoustique sur un modèle de Markov, c'est-à-dire le chemin qui conduit à la plus forte probabilité d'émission de la forme considérée. Le principe de l'algorithme est proche de l'estimation directe introduite dans le paragraphe précédent : la sommation est remplacée par un opérateur *max* pour calculer l'équation (1-7).

Appelons $\Phi_t(j)$ la probabilité maximale sur tous les chemins aboutissant à l'état j au temps t, l'algorithme de Viterbi permet de calculer $\{\Phi_t(j),\ \forall j\}$ à partir de $\{\Phi_{t-1}(j),\ \forall j\}$. A tous les instants, et pour tous les états du modèle, on emploie la formule suivante, explicitant le fait qu'un chemin de longueur t résulte du prolongement d'un chemin de longueur $t-1$ par une transition entre deux états et l'émission d'une trame :

$$\Phi_t(j) = \max_i \{\Phi_{t-1}(i) \times a_{ij}\} \times b_j(o_t) \qquad (1\text{-}17)$$

avec

$$\Phi_1(j) = \begin{cases} 1 \text{ si } j = 1 \\ a_{1j} b_j(o_1) \text{ si } 1 < j < N \end{cases} \qquad (1\text{-}18)$$

et

$$\Phi_T(N) = \max_i \{\Phi_T(i) \times a_{iN}\} \qquad (1\text{-}19)$$

On vérifie qu'en calculant $\Phi_T(N)$ (T : nombre de trames ; N : état final F), on retrouve bien l'équation (1-7).

La reconnaissance d'une forme acoustique est rendue possible par l'évaluation de la probabilité d'émission des observations des modèles de toutes les formes acoustiques. Cela suppose l'existence d'un modèle, au moins pour chaque forme acoustique, éventuellement construit par concaténation de modèles acoustiques élémentaires, et l'apprentissage des paramètres de ces modèles.

1.1.3.3. Le problème de l'apprentissage

Les paramètres des modèles sont les probabilités de transition entre états et les probabilités d'émission associées aux états. La topologie du modèle (le nombre d'états des modèles et les transitions autorisées entre ces états) est supposée 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 par le critère du maximum de vraisemblance (*Maximum Likelihood Estimation*, MLE), obtenue par l'algorithme de Baum-Welch [Baum, 1970].

L'estimation par MLE consiste à choisir le jeux de paramètres $\tilde{\lambda}_{MLE}$ de sorte à rendre maximale la probabilité d'émission des observations O par le modèle :

$$\tilde{\lambda}_{MLE} = \arg\max_{\lambda} P(O/\lambda) \qquad (1\text{-}20)$$

Une résolution analytique directe n'est pas possible, mais les formules de Baum-Welch permettent une réestimation itérative des paramètres a_{ij} et b_j du modèle [Baum, 1972; Liporace, 1982]. A la suite de la $(n+1)^{ième}$ réestimation des paramètres du modèle $\tilde{\lambda}_n$, le nouveau modèle $\tilde{\lambda}_{n+1}$ vérifie :

$$P(O/\tilde{\lambda}_{n+1}) \geq P(O/\tilde{\lambda}_n) \qquad (1\text{-}21)$$

La réestimation itérative est basée sur le principe de l'algorithme EM qui sera détaillé dans §3.1.2. Il peut être montré que sous certaines conditions l'estimateur de maximum de vraisemblance converge asymptotiquement sans biais vers les paramètres optimaux du modèle. Dans ce cas, il est l'estimateur de variance minimale [Nadàs, 1983]. Toutefois, les hypothèses permettant d'aboutir à ce résultat ne sont pas vérifiées en pratique : les échantillons n'ont pas été générés par la distribution prévue et leur nombre n'est pas toujours suffisamment grand.

L'apprentissage étant itératif, les valeurs initiales des paramètres du modèle sont cruciales pour assurer une convergence rapide vers le jeu de paramètres optimal. Les choix concernant la définition des modèles vont donc avoir une influence sur la qualité de l'apprentissage.

1.2. Définition des modèles

Les principes fondamentaux de la modélisation markovienne, introduits dans la section précédente, laissent une grande marge de manœuvre pour leur mise en œuvre concrète au sein d'un système de reconnaissance.

En premier lieu, il convient de définir le modèle utilisé dans le système. Pour cela, trois éléments doivent être précisés : l'unité de modélisation, la topologie des modèles et le type de fdp associées aux états. Les choix concernant ces éléments seront le plus souvent dictés au concepteur du système par sa connaissance experte du domaine. Quelques procédures automatiques ont toutefois été proposées pour s'affranchir de l'approche heuristique sur certains points.

1.2.1. Unité de modélisation

L'unité de modélisation définit le symbole qui sera représentée par un HMM. Trois critères permettent de guider le choix entre les alternatives. Les allophones contextuels constituent un des meilleurs compromis entre ces critères dès lors que des techniques adaptées permettent de résoudre la difficulté de leur apprentissage.

1.2.1.1. Critères de choix

De façon générale, l'unité de modélisation est imposée par la tâche qui fixe l'unité symbolique à reconnaître. Toutefois, lorsque les symboles peuvent être exprimés à l'aide de symboles plus élémentaires, il peut se révéler préférable de définir ces sous-symboles comme unités de modélisation. Lorsque le choix de la décomposition n'est pas unique, trois critères doivent être pris en compte pour déterminer la meilleure unité de modélisation :

- *la consistance* : des occurrences différentes d'une même unité ont des réalisations physiques proches ;
- *la capacité d'apprentissage* : on dispose pour chaque unité d'un nombre suffisant d'exemples pour garantir une estimation robuste des paramètres du modèle ;
- *la capacité d'expression* : la relation entre les unités de modélisation et les unités symboliques peut être facilement obtenue ; par exemple au moyen d'un lexique phonétisé.

Le dernier point est rarement abordé du fait que les relations entre les unités symboliques et les unités de modélisation sont issues de la connaissance experte du domaine, exception faite de l'approche par fénones (cf. paragraphe suivant). La recherche du compromis entre consistance et capacité d'apprentissage a conduit, en reconnaissance de la parole, à la coexistence de différentes approches.

1.2.1.2. Choix de l'unité

Les mots constituent l'unité symbolique à reconnaître *a priori* dans le cadre de la RAP. Ils sont donc des candidats de choix pour l'unité de modélisation. Lorsque la reconnaissance de la parole se limitait à la reconnaissance de mots isolés ou de petits vocabulaires, la modélisation par mots se révélait très efficace (voir, par exemple, [Rabiner, Wilpon *et al.*, 1989]). Ces unités sont consistantes car elles modélisent de façon implicite les co-articulations intra-mots.

Pour la RAP continue à grand vocabulaire, il devient complexe d'apprendre un modèle pour chaque mot. De plus, chaque nouveau mot introduit dans le système nécessitera un apprentissage particulier. Le recours à une modélisation sub-lexicale doit donc être envisagée pour satisfaire le critère lié à la capacité d'apprentissage. Ces unités peuvent être des phonèmes, des phonèmes contextuels, des concaténations de phonèmes (*e.g.* les diphones [Schwartz, Klovstad *et al.*, 1980]), des syllabes [Hunt, Lenning *et al.*, 1980] ou encore des unités acoustiques automatiques sans signification phonologique comme les fénones [Bahl, Brown *et al.*, 1993].

L'unité sub-lexicale la plus immédiate est le phonème : il représente le son le plus bref qui permet de distinguer tous les mots d'une langue donnée. Cette définition est d'ordre phonologique et ne tient pas compte de l'acoustique. Les réalisations physiques d'un phonème peuvent varier considérablement en fonction de facteurs tels que le contexte (co-articulation), la vitesse d'élocution, le dialecte, le style et le locuteur. Les occurrences acoustiques particulières d'un phonème sont appelées allophones.

Les langues, que nous sommes amenés à étudier, comportent un petit nombre de phonèmes : quelques dizaines pour l'anglais comme pour le français. Cette approche est donc satisfaisante du point de vue de sa capacité d'apprentissage. Elle l'est beaucoup moins du point de vue de sa consistance. Parmi les facteurs de variation cités plus haut, l'influence de la co-articulation peut être diminuée par la prise en compte du contexte lors de la modélisation. On distingue :

- *les multi-phonèmes* formés d'une concaténation de phonèmes. Le plus courant est le diphone qui représente deux phonèmes successifs ;
- *les syllabes* dont le statut linguistique est assez bien défini mais dont la localisation est difficile ;
- *les allophones contextuels* qui modélisent les phonèmes suivant leur contexte phonétique. Les unités allophoniques se distinguent des multi-phonèmes car leur support temporel est celui d'un seul phonème.

Dans chacun de ces cas, les unités sont consistantes mais difficiles à apprendre en raison de leur nombre élevé. L'approche par allophones contextuels est plus utilisée car elle présente des caractéristiques favorables à l'application de techniques d'apprentissage visant à compenser leur nombre élevé, telle que la mise en commun entre plusieurs allophones de la partie centrale du phonème.

Toutes les unités de modélisation précédentes ont en commun d'être issues de connaissances expertes du domaine de la parole. Une approche non-paramétrique au problème de la modélisation sub-lexicale a été introduite par les chercheurs d'IBM [Bahl, Brown *et al.*, 1993]. Leur proposition repose sur un découpage de l'espace de représentation acoustique. Chaque zone est associée à une unité élémentaire, le fénone.

La consistance et la capacité d'apprentissage de ces unités peuvent être réglées au travers des choix opérés lors de la partition de l'espace. Le problème est posé par leur capacité d'expression. Un certain nombre d'exemples sont utilisés pour inférer une décomposition de base dont les paramètres sont appris à l'aide du reste des exemples disponibles (les détails de la méthode pourront être trouvés dans [Jelinek, 1997]).

1.2.1.3. Apprentissage des modèles d'allophones contextuels

De nombreuses techniques ont été développées pour améliorer la capacité d'apprentissage des modèles d'allophones contextuels. En effet, pour n phonèmes différents, on obtient n^3 allophones contextuels gauche et droit (appelés des triphones bien qu'il ne s'agisse pas de multi-phonèmes). Pour le français comme pour l'anglais, une cinquantaine de phonèmes sont généralement considérés ; ce qui conduit à environ 125.000 triphones. La taille de la base de données acoustiques d'apprentissage peut alors devenir insuffisante pour apprendre correctement chacun des modèles. Il faut alors soit diminuer le nombre de modèles soit diminuer le nombre de paramètres indépendants dans le système jusqu'à ce que chacun dispose de suffisamment de données d'apprentissage. De plus, un certain nombre d'allophones contextuels peut ne pas être rencontré du tout dans la base.

Parmi les techniques proposées, les allophones généralisés consistent à regrouper dans une même classe les allophones présentant des réalisations acoustiques proches. Le regroupement peut-être obtenu de plusieurs manières : par une classification automatique basée sur une mesure de similarité entre modèles [Lee, 1990], par la construction d'un arbre de décision [Bahl, de Souza *et al.*, 1991] ou par l'utilisation de connaissances phonologiques [Ljolje, 1994]. Il a aussi été proposé de modéliser d'abord séparément l'influence des contextes gauches et droits ; puis de les interpoler au sein d'un nouveau modèle appelé quasi-triphone [Ljolje, 1994]. Cette idée est généralisée dans la modélisation par sénones [Hwang, Huang *et al.*, 1996] qui propose de regrouper les états des triphones en classes (les sénones) partageant des caractéristiques acoustiques communes (évaluées par la mesure de leur divergence de Kullback [Hwang & Huang, 1993]). Les techniques présentées ci-dessus, qui permettent la mise en œuvre de la modélisation contextuelle, peuvent être finalement vues comme des occurrences de systèmes à paramètres partagés (voir §1.3.2) qui ne se différencient que par la règle de décision conduisant au regroupement.

1.2.2. Topologie des modèles

La topologie d'un modèle regroupe le choix du nombre d'états ainsi que la définition d'une matrice de transition initiale. Ce choix est fortement dépendant de l'unité de modélisation choisie.

Les états des HMM peuvent être interprétés comme des zones stationnaires du signal. Les début et fin des réalisations phonétiques présentant des caractéristiques assez différentes de leur centre, plusieurs états sont nécessaires. La topologie gauche-droite à 3 états s'est révélée bien adaptée à la RAP continue. Les techniques d'inférence automatique peuvent être mises en œuvre pour améliorer l'adéquation entre les données et la topologie des modèles.

1.2.2.1. Modèle gauche-droit

Le modèle gauche-droit traduit la causalité du processus de production de la parole : il n'existe pas de cycle dans le graphe orienté engendré par les transitions entre états. Les modèles gauche-droit particuliers, autorisant le bouclage à l'état courant, le passage à l'état suivant ou le saut d'un état sont le plus couramment utilisés pour représenter les unités phonétiques. Trois états sont généralement utilisés (Figure 1-2). [Barras, 1996] confirme ce choix en testant sur la base de données TIMIT des systèmes utilisant des topologies phonétiques de 1 à 4 états ; les meilleurs taux de reconnaissance en DAP sont obtenus par les modèles à 3 états.

Des structures plus complexes ont été proposées : la topologie à 7 états de l'équipe d'IBM [Derouault, 1987] qui est utilisée dans le système SPHINX du CMU [Lee, 1988] ou encore le modèle allophonique du CNET [Jouvet, 1995].

1.2.2.2. Inférence de topologie

Quelques tentatives ont été menées pour déterminer automatiquement les topologies des modèles. Parmi celles-ci, un certain nombre sont fondées sur les techniques d'inférence grammaticale appliquées aux chaînes de symboles obtenues après quantification vectorielle des vecteurs de paramètres acoustiques [Casacuberta, Vidal *et al.*, 1990; Lockwood & Blanchet, 1993]. Par ailleurs, une approche heuristique a été proposée consistant à optimiser une structure *a priori* par un recuit simulé guidé par la réduction du taux d'erreur [de Mori, Galler *et al.*, 1995].

Jusqu'à présent, les topologies obtenues par ces techniques n'ont pas été jugées suffisamment pertinentes pour être retenues ; principalement car le coût de calcul pour les obtenir n'était pas justifié par leur apport en terme de performance. Nos expériences, concernant l'inférence automatique de topologie, sont présentées dans le second chapitre. Elles tentent de tirer profit de l'estimation par K-ppv.

La topologie du modèle fixe les valeurs initiales des éléments de la matrice de transition A ainsi que leur nombre. Il reste pour définir entièrement un modèle à définir le vecteur B en associant à chaque état une fdp.

1.2.3. Estimateur de probabilité

Deux groupes d'estimateur de probabilités peuvent être distingués : les estimateurs paramétriques et non-paramétriques. Les seconds, qui seront étudiés plus précisément dans le second chapitre, utilisent les données de références (échantillons de la distribution) pour interpoler localement les valeurs des fdp. Les premiers prédisent un modèle pour les fdp puis ajustent les paramètres de ces modèles à l'aide des données de référence.

En RAP, la distinction entre les estimateurs est couramment faite sur une autre de leur caractéristique liée à la continuité de l'espace d'évaluation des fdp. On distingue les estimateurs continus et l'estimateur discret par quantification vectorielle.

1.2.3.1. Estimateurs continus

Les estimateurs continus sont fondés sur le principe que la fdp à estimer appartient à une famille de fonctions. Ils définissent explicitement la famille de fonctions (par exemple les gaussiennes pour les GMM) et associent une somme pondérée (*mixture*) de fonctions à chaque classe. La somme pondérée associée à une classe est obtenue par l'apprentissage de ses paramètres sur un ensemble de données de référence appartenant à cette classe. Des variantes existent comme les estimateurs semi-continus qui évaluent un ensemble de fonctions *a priori* représentant au mieux l'ensemble des données de référence, seuls les coefficients de la somme pondérée étant particuliers aux états.

Principes

Les estimateurs de probabilité continus sont définis par une somme pondérée de fonctions élémentaires. La plus communément utilisée est la somme pondérée de fdp gaussiennes multidimensionnelles (GMM) :

$$b_i(x_0) = \sum_{g=1}^{G} c_{ig} \mathrm{N}(x_0; \mu_{ig}, \Sigma_{ig}) \qquad (1\text{-}22)$$

où μ_{ig} *et* Σ_{ig} sont respectivement le vecteur moyenne et la matrice de covariance de la $g^{\text{ième}}$ gaussienne de l'état i et c_{ig} le coefficient de pondération qui lui est affecté. L'hypothèse d'une indépendance entre les d dimensions de l'espace de représentation autorise l'utilisation de matrices de covariance diagonales ; ce qui limite le nombre de paramètres à estimer et simplifie les calculs.

Figure 1-3 Somme de gaussiennes.

L'approche paramétrique de l'estimation de fdp est validée par la théorie des noyaux [Parzen, 1962; Cacoullos, 1966]. Elle montre qu'il est possible d'estimer n'importe quelle fdp en tout point x_0 d'un espace multidimensionnel par une combinaison de fonctions :

$$\frac{1}{nh_n^d}\sum_{j=1}^{n} K\left(\frac{x_0 - x_j}{h_n}\right) \xrightarrow{n\to\infty} f(x_0) \tag{1-23}$$

avec $x_1 \ldots x_n$ n observations de la variable aléatoire X de distribution f. La suite $\{h_n\}$ converge vers 0 et K est une fonction noyau (*i.e.* elle est bornée, d'intégrale unitaire et vérifie $\lim\limits_{|y|\to\infty} |y|^d K(y) = 0$).

La théorie des noyaux fonde les principes de la construction d'une fonction à partir d'échantillons de ses valeurs. La distinction entre les approches paramétriques et non-paramétriques (§2.2.3) tient dans le choix des échantillons pris en compte : l'approche paramétrique considère des points moyens représentatifs des modes de l'ensemble des échantillons initiaux, l'approche non-paramétrique, comme nous le montrerons plus en détail dans §2.2.3, utilise l'ensemble des échantillons disponibles lors de l'estimation.

La loi normale respecte les contraintes des fonctions noyaux et représente donc un candidat de choix pour (1-23). D'autres fonctions ont été utilisées dans les systèmes de RAP : certaines sont des raffinements de la fonction gaussienne comme les sommes pondérées de gaussiennes auto-régressives [Juang & Rabiner, 1985], d'autres sont basées sur des fonctions très proches de la gaussienne comme les sommes pondérées de densités de Richter [Richter, 1986] ou les sommes pondérées de densités de Laplace [Ney & Noll, 1988]. Quelle que soit la fonction choisie, elle doit satisfaire à deux types de contraintes : les contraintes statistiques qui déterminent la qualité de l'estimation obtenue et les contraintes de modélisation permettant d'intégrer l'estimateur choisi dans le formalisme des HMM. Ainsi, afin de montrer la convergence des formules de réestimation utilisées dans l'algorithme de Baum-Welch dans le cas des estimateurs continus, la fonction élémentaire retenue doit posséder une symétrie elliptique [Liporace, 1982]. Cette hypothèse s'est substituée à celle de log-concavité, plus contraignante. L'ensemble des fonctions citées précédemment respecte cette contrainte.

Variantes

Les variantes de l'estimateur continu, les plus répandues, reposent sur le principe du partage de paramètres (§1.3.2) appliqué aux fdp des états. Elles interviennent dans les systèmes semi-continus [Huang & Jack, 1989; Huang, 1992] et à sommes pondérées liées (*tied-mixture*) [Bellagarda & Nahamoo, 1990]. Dans ces systèmes, toutes les fdp sont regroupées et partagées par toutes les classes. La fdp associée à un état est définie par un vecteur de G coefficients de pondérations, G étant le nombre total de fdp du système. Cette approche est finalement peu différente du cas général : tous les états possèdent une somme pondérée de G gaussiennes mais celles-ci sont partagées par tous les états, seules les pondérations sont associées aux états. La technique du partage des paramètres (voir §1.3.2) permet de montrer l'équivalence entre les deux estimateurs. Elle a aussi permis d'appliquer ce procédé à tous les niveaux d'un système markovien (état, modèle). [Duchateau, 1998] propose le terme de modèles réduits (*reduced models*) lorsque les estimateurs de fdp sont partagés par des sous-ensembles d'états.

Les systèmes à estimateur semi-continu offrent *a priori* une moins grande précision de modélisation que les systèmes à estimateur de probabilité continu. La contrepartie est qu'ils nécessitent moins de calculs dès lors qu'un moins grand nombre de gaussiennes doit être réévalué pour chaque trame (la somme pondérée de gaussiennes est un calcul peu coûteux comparé au calcul de la gaussienne pour une vecteur donné). Cette approche peut donc être envisagée pour des considérations de temps de calcul ou lorsque le nombre de données disponibles pour estimer les fonctions pour chaque état est insuffisant ; notamment lorsque l'on fait croître de façon importante le nombre d'états indépendants du système (dans le cas des modèles contextuels par exemple).

Le développement des techniques de décodage par passes multiples (§1.4.5) a conduit à des systèmes intégrant des ensembles de modèles dont la précision d'estimation varie. Certains systèmes de l'état de l'art gèrent l'augmentation de la précision par le passage de l'estimateur semi-continu à l'estimateur continu.

Mise en œuvre

L'estimateur de fdp gaussien est l'estimateur le plus utilisé de l'état de l'art. Lors de la campagne d'évaluation DARPA *Broadcast News* 1997, un seul système sur dix engagés n'utilisait pas un estimateur continu et parmi les 9 autres, seul le système de Philips utilisait un estimateur à base de laplaciennes [DARPA, 1998].

L'estimateur continu permet d'ajuster la précision d'estimation à un coût de calcul donné ; notamment par le nombre de fonctions élémentaires utilisées par mixture. De plus, la compression d'information introduite par la modélisation des données au travers des paramètres de la fonction choisie permet un gain important en stockage. Toutefois, cette observation doit être nuancée si on observe les systèmes de l'état de l'art à base de GMM qui peuvent utiliser jusqu'à plusieurs centaines de milliers de gaussiennes (environ 300000 pour le système du LIMSI de l'évaluation DARPA *Broadcast News* 97 [DARPA, 1998]). Cette inflation du nombre de gaussiennes réduit le gain en compression d'information et augmente le coût de calcul.

Des techniques d'optimisation ont été développées. Elles visent soit à améliorer l'apprentissage des paramètres des estimateurs, soit à diminuer l'effort de calcul. Une procédure automatique basée sur la mesure du χ^2 a été proposée pour la détermination du nombre optimal de gaussiennes par somme pondérée [Barras, 1996]. Cette procédure permet une réduction importante du nombre de paramètres sans baisse de performance. Divers procédés permettent l'accélération du calcul des mixtures de gaussiennes parmi lesquels la tesselation de l'espace de représentation [Ortmanns, Firzlaff *et al.*, 1997] ou encore la technique de sélection des fonctions actives (*Fast Removal of Gaussians*, FRoG) [Duchateau, 1998].

Le principal inconvénient de l'estimateur continu est qu'il repose sur l'hypothèse que la fonction élémentaire utilisée est cohérente avec la loi réelle des données. Or, les vecteurs d'analyse n'ont pas, en général, de distribution gaussienne [Barras, 1996; Montacié, Caraty *et al.*, 1996]. L'existence d'une somme particulière de fonctions qui converge vers la loi réelle est montrée théoriquement mais pas la manière de l'obtenir. Ainsi, le choix du nombre de fonctions élémentaires dans la somme pondérée, tout comme l'apprentissage leurs paramètres, est guidé par des heuristiques et ne permet pas, en général, d'assurer une convergence vers la loi réelle.

L'estimateur continu à base de mixtures de gaussiennes (GMM) constituera l'estimateur de l'état de l'art auquel nous nous référerons tout au long de cette étude pour évaluer les approches proposées. Une alternative à l'estimateur continu est l'estimateur discret par quantification vectorielle.

1.2.3.2. Estimateur discret par quantification vectorielle

Le principe de l'estimateur de fdp discret par quantification vectorielle est de partitionner l'espace de représentation en associant à chaque zone un représentant. Celui-ci est soit le centre de gravité de la partition (*centroïde*), soit le vecteur le plus proche de tous les autres (*prototype*). L'habitude veut que l'on parle dans les deux cas de prototypes.

L'ensemble des prototypes constitue le dictionnaire (*codebook*) de l'espace de représentation discretisé. Divers algorithmes ont été proposés pour la construction du dictionnaire des prototypes. Le plus courant est l'algorithme des k-moyennes utilisé dans sa version LBG (*Linde-Buzo-Gray*) [Linde, Buzo *et al.*, 1980]. Chaque vecteur de paramètres est ensuite associé au prototype du dictionnaire dont il est le plus proche au sens d'une certaine distance (classiquement une distance euclidienne pondérée comme la distance de Mahalanobis). La probabilité du vecteur sur un état est alors la probabilité du prototype pour cet état. Les probabilités des prototypes pour chaque état sont déterminées lors de l'apprentissage.

La taille du dictionnaire permet un contrôle de la distorsion introduite par la substitution d'un vecteur par son prototype. La contrepartie à cette perte de précision dans la modélisation des données est un coût de calcul faible : la distance d'un vecteur a tous les prototypes. Ce dernier point justifie son utilisation lorsque de fortes contraintes temporelles (par exemple un fonctionnement en temps réel) sont imposées au système. Le cas limite de l'estimation discrète, au prix d'un augmentation importante de son coût de calcul, consiste à conserver dans le dictionnaire tous les vecteurs de référence. L'estimateur discret devient alors l'estimateur des 1-ppv, cas particulier de l'estimateur non-paramétrique des K-ppv qui sera présenté ultérieurement.

Une fois les modèles définis, une procédure automatique permet l'apprentissage des paramètres restés libres : les probabilités de transition et les paramètres des fdp des états.

1.3. Techniques d'apprentissage

L'algorithme de Baum-Welch (§1.1.3.3) est communément utilisé pour l'apprentissage initial des modèles. A partir des modèles initiaux, plusieurs techniques permettent d'améliorer la qualité de l'estimation des paramètres. La réestimation connectée est une variante de l'algorithme de Baum-Welch qui permet l'apprentissage simultané de l'ensemble des modèles. La technique du partage de paramètres permet d'augmenter artificiellement les données d'apprentissage associées à chacun des paramètres. Finalement, une adaptation des modèles peut être réalisée afin de les spécialiser sur un locuteur particulier.

1.3.1. Réestimation connectée

Dans l'apprentissage par l'algorithme de Baum-Welch, tel qu'il a été introduit dans (§1.1.3.3), chaque modèle phonétique est appris séparément. Il est donc nécessaire de pouvoir associer les segments de données aux classes phonétiques modélisées par les HMM. Ceci suppose qu'il existe une segmentation *a priori* de la base de données. Cette opération, réalisée par des experts humains, est longue et complexe. C'est pourquoi peu de bases de données de parole disposent d'une segmentation phonétique experte. La méthode généralement retenue pour l'apprentissage des HMM consiste alors à générer des modèles initiaux à partir des données segmentées dont on dispose. Puis, ces modèles sont utilisés soit pour corriger automatiquement les erreurs de segmentation soit pour segmenter de nouvelles phrases à l'aide de leur transcription phonétique (ou lexicale).

La réestimation connectée permet la mise en œuvre d'une telle approche. Connaissant la transcription phonétique (à l'aide d'un lexique phonétisé si l'on ne dispose que de la transcription lexicale) d'une séquence acoustique d'apprentissage, un réseau est constitué par la concaténation des modèles phonétiques correspondants. Ce réseau est ensuite traité par l'algorithme de Baum-Welch comme un modèle unique. La procédure d'apprentissage complète est reprise dans la Figure 1-4.

```
•  Apprentissages séparés des modèles initiaux ;
•  Réestimation connectée :
      Pour chaque phrase :
            Construction du réseau de modèles phonétiques ;
            Estimation du réseau par l'algorithme de Baum-
            Welch : accumulation des statistiques ;
      Mise à jour des paramètres pour tous les modèles ;
```

Figure 1-4 Déroulement de l'apprentissage d'un système markovien avec la réestimation connectée.

La réestimation connectée permet donc la prise en compte de données non-segmentées lors de l'apprentissage des modèles. Elle conduit à augmenter considérablement le nombre de données utilisables en minimisant le coût humain et financier. Malgré tout, lorsque le nombre de paramètres du système augmente mais que le nombre de données ne peux plus être augmenté, la technique de partage de paramètres doit être envisagée.

1.3.2. Partage de paramètres

La technique du partage de paramètres entre HMM permet s'assurer une meilleure généralisation des modèles avec des données en quantité limitée. Tous les paramètres (ou groupe de paramètres) d'un HMM peuvent être partagés, à commencer par le HMM en entier dans le cadre des modèles contextuels généralisés (voir §4.2.3.1). Une description générale en a été donnée qui permet d'appliquer la technique de façon uniforme à l'ensemble des paramètres des HMM [Young, 1992]. Les formules de réestimation des procédures d'apprentissage ne sont pas modifiées par la technique du partage de paramètres [Bellagarda & Nahamoo, 1990].

Le partage de paramètres a deux avantages principaux :

- favoriser une *estimation robuste* des paramètres lorsque les données sont limitées. Ainsi, il est possible de concevoir des procédures automatiques qui groupent un certain type de paramètres jusqu'à ce que la quantité de donnée associée à chacun d'entre eux soit suffisante.
- réduire la *taille de stockage* des modèles et accélérer la vitesse de calcul. Lorsque le mécanisme de partage est correctement implémenté, une seule occurrence des paramètres doit exister pour un grand nombre de pointeurs vers celle-ci. Ce qui a pour effet de diminuer la taille des modèles en mémoire. De plus, lors de l'utilisation de ces paramètres, un système de cache permet de ne procéder qu'à un seul calcul.

Le partage des paramètres suppose de regrouper les paramètres en classes partageant des caractéristiques acoustiques communes. De telles classes peuvent être obtenues à partir des différents critères qui seront détaillés dans le cas de la modélisation contextuelle (*cf.* §4.2.3.1).

Le réestimation connectée et la technique du partage de paramètres permettent de s'assurer d'un nombre suffisant de données pour l'apprentissage des modèles. La première augmente le nombre de données disponibles, la seconde adapte le nombre de paramètres indépendants des modèles le nombre de données disponibles. La précision de la modélisation peut être aussi améliorée par la prise en compte du locuteur.

1.3.3. Adaptation au locuteur

L'adaptation au locuteur peut être obtenue soit par une transformation de l'espace de représentation soit par une transformation des modèles. La principale différence en pratique réside dans la possibilité, pour les transformations de modèles, de définir un ensemble de transformations associées soit à des groupes de modèles soit à des classes de paramètres des modèles.

1.3.3.1. Modes d'adaptation

L'adaptation au locuteur d'un système de reconnaissance peut être réalisée de différentes façons. On distingue 5 modes d'adaptation qui peuvent être combinés :

- *enrôlement* : l'adaptation est faite en préalable à l'utilisation du système de reconnaissance ;
- *incrémental* : chaque nouvelle reconnaissance est utilisée pour améliorer l'adaptation ;

- *batch* : mode non causal – les sorties du système de reconnaissance sont sauvegardées, utilisées pour l'adaptation puis décodées à nouveau ;
- *supervisé* : une transcription précise des données d'adaptation est disponible ;
- *non-supervisé* : la sortie du système de reconnaissance est utilisée comme transcription des données d'adaptation.

Clairement, le choix et les combinaisons de ces modes dépendent de l'application et de choix ergonomiques. Par exemple, un système de dictée vocale personnel utilisera typiquement un enrôlement supervisé (augmenté, éventuellement, d'une adaptation incrémentale non-supervisée) tandis qu'un système d'indexation de documents audiovisuels aura recours à un mode batch non-supervisé.

1.3.3.2. Adaptation de l'espace de représentation

Différentes techniques d'adaptation de l'espace de représentation ont été proposées : normalisation par intervalle formantique, normalisation par la fréquence fondamentale [Schwartz & Kubala, 1992] et le VTLN (*Vocal Tract Length Normalisation*) [Lee & Rose, 1996].

L'usage du VTLN s'est répandu dans les systèmes de l'état de l'art [DARPA, 1998]. Son principe est de définir une transformation de l'échelle des fréquences Mel afin de minimiser l'influence de la longueur du conduit vocal. Cette technique permet de s'affranchir de la modélisation par genre puisque c'est entre les hommes et les femmes qu'existe la plus grande variabilité du conduit vocal (si l'on excepte le cas des enfants).

1.3.3.3. Adaptation des modèles

Deux approches sont devenues communes pour l'adaptation au locuteur par ré-apprentissage des modèles initiaux : le MAP (*Maximum A Posteriori*) [Gauvain & Lee, 1994] et le MLLR (*Maximum Likelihood Linear Regression*) [Leggetter & Woodland, 1995].

<u>MAP</u>

Le critère MAP maximise la probabilité *a posteriori* des paramètres :

$$\tilde{\lambda}_{MAP} = \arg\max_{\lambda} P(\lambda/O) = \arg\max_{\lambda} P(O|\lambda)P(\lambda) \tag{1-24}$$

Dans cette approche, le jeu de paramètres est considéré comme une variable aléatoire dont il est possible d'estimer la probabilité *a priori* $P(\lambda)$. Le problème est alors de choisir la loi de probabilité qui sera associée aux différents paramètres de λ. Dans leur proposition, Gauvain et Lee optent pour une loi de Dirichlet pour les pondérations des mixtures de gaussiennes et une loi de Wishart pour les matrices de covariance de chaque gaussienne. Le critère du MLE correspond à une loi uniforme pour tous les paramètres.

Le critère MAP est essentiellement utilisé dans le cadre de l'adaptation dynamique au locuteur. Les modèles sont entraînés indépendamment du locuteur à l'aide du critère MLE. Lors de la reconnaissance, un premier décodage permet d'obtenir des données propres au locuteur. Celles-ci sont alors utilisées pour adapter les modèles à l'aide du critère MAP. Les probabilités *a priori* des paramètres, apprises sur les données indépendantes du locuteur, permettent un ré-apprentissage plus rapide et limite le risque de dérive du à l'imprécision des données d'adaptation.

Toutefois, la quantité de données d'adaptation étant généralement limitée, la méthode ne peut s'appliquer que sur un certain nombre des paramètres des modèles. Le choix de ceux-ci et la taille critique des données d'adaptation sont évalués lors du développement des systèmes. Si l'on se réfère aux systèmes engagés dans l'évaluation DARPA 1997, l'adaptation par MAP est utilisée de façon préférentielle pour l'adaptation par genre, ce qui suppose finalement un nombre de données d'adaptation assez important. Dans les cas où les données d'adaptation sont trop peu nombreuses, comme pour une adaptation non-supervisée au locuteur, le MLLR est préféré.

<u>MLLR</u>

La technique d'adaptation par MLLR a été développée pour les HMM à estimateurs de probabilité continus. Dans un premier temps, les données disponibles pour le locuteur permettent de calculer une transformation de type régression linéaire sur les moyennes des modèles. La transformation est évaluée pour maximiser la vraisemblance de l'adaptation et peut être estimée à partir de l'algorithme de Baum-Welch. Le partage de paramètres entre distributions permet de réaliser l'adaptation de distributions non utilisées dans les données d'adaptation. De même lorsque la quantité de données d'adaptation est très faible, une seule transformation peut être partagée par toutes les distributions du système. Avec l'augmentation de la taille des données d'adaptation, les paramètres des distributions peuvent être groupés en classes ayant chacune leur transformation propre.

Le nombre de transformations, en fonction des données d'adaptation disponibles, est déterminé par un arbre de régression. Chaque nœud représente une classe de paramètres. Le parcours de l'arbre permet de déterminer les classes qui vont pouvoir être adaptées. La construction de l'arbre peut être opérée suivant deux critères :

- *les macro-classes phonétiques* : les paramètres des distributions appartenant à la même macro-classe sont regroupés dans le même nœud ;

- *la proximité des distributions* : comme pour le partage de paramètres, les paramètres des distributions proches sont regroupés. Cette approche est plus appropriée à un nombre de classes élevé.

Les techniques d'adaptation permettent des gains non négligeables, bien que très variables en performances. Un décodage des données de test est généralement nécessaire en préalable à l'adaptation des modèles afin d'obtenir des données propres au locuteur.

1.4. Techniques de décodage

La reconnaissance segmentale continue suppose que les symboles que l'on cherche à reconnaître s'enchaînent sans que l'on sache pour un segment donné ni leur nombre ni leur localisation. Il n'existe pas de réponse simple à ce problème si on exclut la présence de segments facilement détectables entre les unités de base[1]. Dans le cadre de la RAP continue, les opérateurs de reconnaissance réalisant la localisation et l'identification des unités de base dans un même processus se sont révélés les plus efficaces.

Cette approche du décodage de modèles connectés contraint à envisager le décodage de tous les messages possibles. Une double problématique doit être alors résolue : la génération d'une structure de recherche représentant tous les messages possibles (définis préalablement par un réseau de décodage) et le décodage à l'aide de cette structure d'une suite d'unités correspondant au message prononcé. Les connaissances linguistiques formulées au travers d'un modèle de langage permettent de guider le décodage dans la structure de recherche.

1.4.1. Modèles de langage

Pour la reconnaissance de la parole continue, un modèle de langage qui participe au choix des candidats lors du décodage d'une phrase présente un grand intérêt. L'approche à syntaxe fixe décrit explicitement les messages possibles. Ils sont donc équiprobables mais l'explosion combinatoire est enrayée. Les approches probabilistes permettent d'évaluer la probabilité $P(M)$ pour n'importe quelle suite de mots M. L'explosion combinatoire est maîtrisée en ne s'intéressant qu'aux messages les plus probables.

[1] Des silences dans le cas de la parole.

1.4.1.1. Approche à syntaxe fixe

Les approches à syntaxe fixe consistent en un ensemble de règles exprimées à l'aide de CFG (*Context-Free Grammar*). Leur traduction en graphes détermine les séquences de mots valides. La limitation la plus évidente de ces approches est l'impossibilité de décrire de façon exhaustive une langue au travers d'une grammaire à base de règles [Black, 1993]. Toutefois, dans le cadre d'application sur des sous-langages simples, elles permettent de générer des graphes tout à fait complet (numéros de téléphone, heures…).

La seconde limitation importante des approches à syntaxe fixe réside dans leur incapacité à reconnaître des messages grammaticalement approximatifs ou même faux. Or, contrairement au langage écrit, le langage oral comporte assez facilement des constructions non grammaticales. Le cas extrême est la parole spontanée avec ses faux départs, hésitations… Il convenait donc de trouver une approche permettant de modéliser le langage d'un point de vue strictement probabiliste plus conforme à la nature imprécise du discours humain.

1.4.1.2. Approche probabiliste

Le principe d'un modèle de langage probabiliste est de capturer les régularités parmi les suites de mots. Ces régularités dans le langage de l'application résultent des contraintes syntaxiques, sémantiques et pragmatiques.

Strictement, la probabilité qu'un certain mot soit prononcé dépend de tous les mots l'ayant précédé dans la phrase. Si nous considérons comme événements les phrases (suites de mots), la probabilité d'une phrase $m_1m_2...m_n$ s'écrit comme le produit des probabilités conditionnelles :

$$P(m_1 \ldots m_n) = P(m_1)\prod_{i=2}^{n} P(m_i | m_1 \ldots m_{i-1}) \qquad (1\text{-}25)$$

Les probabilités (1-25) sont difficiles (voire impossibles) à estimer, et de toutes manières impossibles à stocker. Une façon de réduire le nombre de cas pris en compte consiste à réaliser une classification des chaînes de mots ou à définir des classes d'équivalence entre elles.

Deux approches principales sont alors possibles :

- *n-grammes* : deux chaînes $m_1 \ldots m_k$ et $l_1 \ldots l_h$ sont déclarées équivalentes si les n derniers mots sont identiques dans chacune. Ainsi, seules les probabilités $P(m|m_{k-n} \ldots m_k)$ doivent être estimées. En pratique, il s'agira d'uni- jusqu'au quintigrammes.
- *n-classes* : chaque mot est associé à une classe (grammaticale ou autre) qui est prise en compte pour réaliser la classification des chaînes de mots.

Dans les deux cas, il faudra estimer les probabilités $P(m|C_n)$ (avec C_n une classe d'équivalence) en relevant dans le corpus d'apprentissage les fréquences $f(m,C_n)$ d'apparition de m après un ensemble de mots de classe C_n, *i.e.* :

$$P(m|C_n) = f(m,C_n) = \frac{card(m,C_n)}{card(C_n)} \tag{1-26}$$

Les fréquences sont estimées sur de grands corpus de texte. En principe, les n-grammes peuvent être estimés par les fréquences d'apparition. Le problème est que le nombre potentiel de n-grammes est la puissance n de la taille du vocabulaire (soit pour seulement 1.000 mots, 1 milliard de trigrammes distincts). Beaucoup de ceux-ci vont être absents des corpus de textes ou n'apparaître que de rares fois (1 ou 2).

Deux approches ont été proposées pour améliorer la robustesse de l'estimation des n-grammes : le *deleted interpolation* [Jelinek, 1989] et le *back-off* [Katz, 1987]. Le principe du back-off est de retenir sur les n-grammes existant dans l'ensemble d'entraînement une part de leur probabilité (*discounting*) qui est redistribuée sur les n-grammes non rencontrés. Ces derniers sont alors estimés à partir d'un n-1-gramme étalonné (*back-off*). Le *deleted interpolation* utilise directement une combinaison linéaire des n-grammes des niveaux inférieurs (*n-1* jusqu'à 1). Les paramètres de la combinaison linéaire sont appris sur le corpus d'entraînement. Les modèles utilisés dans cette étude seront tous du type back-off, plus simple à mettre en œuvre.

1.4.1.3. Approches plus évoluées

Bien que l'estimation robuste des n-grammes reste un problème complexe (malgré les techniques introduites ci-dessus), l'intégration des modèles de langage lors du décodage conduit à des gains significatifs de performance. Toutefois, leur incapacité à exploiter les dépendances à long terme (les accords sujet-verbe, par exemple) a motivé l'étude de solutions plus efficaces.

Un certain nombre d'approches de modèles de langage évolués ont été envisagées : *tree-based models* [Bahl, Brown *et al.*, 1989], *cache models* [Kuhn & de Mori, 1990], *treillis models* [Waegner & Young, 1992], *trigger models* [Lau, Rosenfeld *et al.*, 1993], *history models* [Black, Jelinek *et al.*, 1992] ou encore les *multi-grammes* [Deligne, 1996]. Les principales limitations à la généralisation de leur usage dans les systèmes de reconnaissance de l'état de l'art sont la difficulté posée par leur intégration dans les algorithmes de décodage classiques et un coût de calcul généralement plus élevé que les n-grammes. La performance de ces modèles de langages doit donc être évaluée *a priori*.

1.4.1.4. Evaluation des modèles de langage

Le calcul de la perplexité [Jelinek, 1976] s'est très largement imposé en tant que critère d'évaluation des modèles de langage. La perplexité permet une mesure de l'information introduite par le modèle de langage dans le système. Sans information particulière lors du décodage, tous les mots du lexique sont envisageables à la suite d'un autre (équiprobabilité). L'introduction du modèle de langage permet de déterminer quels sont les mots les plus probables à la suite d'un autre. En ce sens, il réduit l'information liée à un mot quant à son possible successeur. Pour mesurer cette réduction, la perplexité nous donne le facteur de branchement moyen. Plus la perplexité est basse, moins le système est « perplexe » quant au choix des solutions.

La diminution de la perplexité n'entraîne pas systématiquement l'augmentation des performances du système de reconnaissance intégrant le modèle de langage [Clarkson & Robinson, 1998]. De plus, son influence sur le processus de reconnaissance n'est pas figée et peut être réglée dans une certaine mesure.

1.4.1.5. Facteur d'échelle

Dans l'approche statistique de la RAP, l'objectif est de trouver le message de plus grande probabilité *a posteriori*. L'équation (1-3) représente le cadre théorique dans lequel les distributions de probabilités, aussi bien acoustiques *P(O/M)* que linguistiques *P(M)*, sont connues de façon exacte. Or, la nécessité d'estimer ces distributions à partir de modèles (HMM pour l'acoustique, modèles de langage pour la linguistique) modifient notamment la dynamique des valeurs des probabilités. Afin de compenser cet effet, il est d'usage courant d'introduire un facteur d'échelle γ entre les deux types de probabilités lors du décodage :

$$\tilde{M} = \arg\max_{M} P(O|M).P(M)^{\gamma} \qquad (1\text{-}27)$$

Le facteur d'échelle est introduit dans le processus de décodage au travers de la structure de recherche. Il a une influence, lors du décodage, sur la segmentation [Ljolje, 1994]. Une valeur forte diminue le nombre d'insertions en pénalisant les transitions entre mots. A l'inverse, une valeur faible augmente les omissions en pénalisant les transitions entre mots. La meilleure valeur est donc celle qui équilibre les deux effets. Elle est généralement obtenue par quelques essais sur l'ensemble de développement.

Dans nos expériences, le facteur d'échelle pour les systèmes HMM/GMM a été fixé à 3 en accord avec l'étude menée dans [Barras, 1996]. Pour les systèmes HMM/K-ppv, les expériences de reconnaissance ont été réalisées pour des valeurs variant de 0 à 5 par pas de 0,5. Les valeurs offrant les meilleurs scores sur l'ensemble d'apprentissage ont été utilisées pour l'ensemble de test. Quelques expériences nous ont permis de vérifier que ces valeurs correspondaient effectivement aux meilleurs scores de reconnaissance sur le test. Les valeurs généralement retenues se situent autour de 1,5 (la valeur exacte sera précisée pour l'expérience concernée).

1.4.2. Réseau de décodage, structures de recherche linéaire et en arbre

Le réseau de décodage représente la syntaxe des entités symboliques que l'on cherche à reconnaître (mots ou phonèmes). Il définit l'ensemble des messages autorisés pour la tâche par un ensemble de règles de production de type BNF. Si un modèle de langage à syntaxe fixe est utilisé, le réseau de décodage est la reproduction des règles syntaxiques donnée par le modèle. Lorsque l'emploi de modèles de langage probabilistes est envisagé, le réseau de décodage se résume en une boucle sur l'ensemble des mots du vocabulaire. L'intervention de la syntaxe se fera lors du décodage au travers de la structure de recherche.

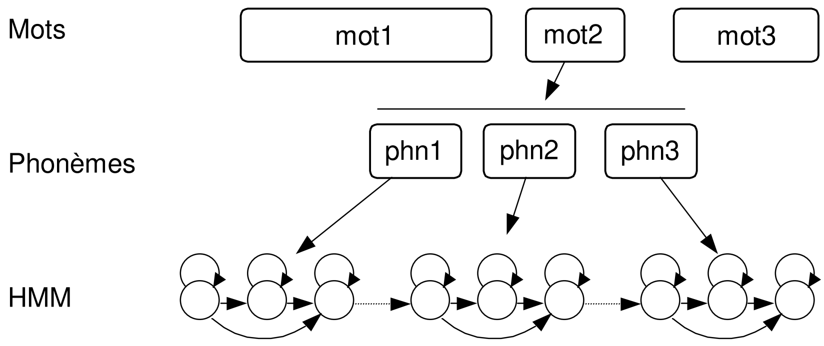


Figure 1-5 Illustration de la décomposition en sous-unités, des mots d'abord en phonèmes puis des phonèmes en HMM.

La structure de recherche est le graphe, composé uniquement de HMM, à partir duquel sera réellement effectué le décodage. Elle est obtenue par la compilation du lexique phonétisé et du réseau de décodage. Le lexique phonétisé fourni pour chaque entité symbolique sa description par un réseau d'unités élémentaires correspondant aux HMM. Ainsi, lors d'un décodage phonétique, les entités symboliques sont les phonèmes, décrits par le HMM qui leur est associé. Lors d'un décodage lexical, les entités symboliques sont les mots décrits par un réseau de HMM conforme à la décomposition phonétique du mot (illustrée Figure 1-5). Les descriptions du lexique phonétisé peuvent comprendre des alternatives et autoriser des élisions. Deux types de compilation peuvent être réalisés pour engendrer la structure de recherche.

La compilation la plus générale consiste à remplacer directement chaque occurrence d'un mot dans le réseau de décodage par sa décomposition phonétique donnée par le lexique phonétisé. On obtient alors une structure de recherche linéaire. Dans le cas de l'emploi de modèles de langage probabilistes, il existe une alternative qui consiste à factoriser les décompositions des unités élémentaires sous forme d'un arbre lexical [Ney, Haeb-Umbach *et al.*, 1992]. Tous les mots débutant par la même suite d'unités phonétiques sont représentés par le même sous-arbre et se distinguent par les feuilles associées aux terminaisons.

Des variantes, pour l'instant moins répandues, opèrent la factorisation aux deux extrémités des mots [Duchateau, 1998]. La représentation en arbre du lexique ne présente d'intérêt que lorsque le réseau de décodage est composé de l'ensemble des mots du vocabulaire mis en parallèle. Les structures de recherche en arbre évitent la duplication inutile d'unités de modélisations. L'intérêt est de ne pas multiplier les mêmes traitements sur les mêmes unités lors du décodage ; notamment les coupures sont partagées par les mots d'un sous-arbre. Le principal inconvénient des structures en arbre est de rendre complexe l'emploi d'un modèle de langage.

Avec une structure de recherche linéaire, lors du décodage, il est simple d'associer la probabilité du bigramme à un chemin lors de la transition entre deux mots. Avec une structure en arbre, il n'est pas possible de déterminer *a priori* à quel mot sera associé un chemin puisque plusieurs mots se partagent le même sous-arbre. Différentes solutions ont été apportées à ce problème (voir, par exemple, [de Mori, 1998] p. 231). L'enjeu était, évidemment, de réduire le moins possible le gain dû à la méthode. Une solution consiste à dupliquer les arbres lexicaux de sorte à associer chaque arbre soit à un prédécesseur (*word-conditioned copy*) soit à un instant de démarrage (*time-conditioned copy*) et donc de retarder la prise en compte du bigramme [Ney, Haeb-Umbach *et al.*, 1992 ; Ortmanns, Eiden *et al.*, 1998]. Les deux options se révèlent comparables en terme d'efficacité [Ortmanns, Ney *et al.*, 1996].

Le choix d'une structure lexicale linéaire ou en arbre conditionne l'algorithme de décodage. Les récents développements ont montré que la structure en arbre permet des gains en temps de calcul appréciables et donc constitue une opportunité soit pour évoluer vers des systèmes temps-réel soit pour introduire une plus grande précision à d'autres niveaux du processus de reconnaissance [Ravishankar, 1996].

1.4.3. Principaux Algorithmes de décodage

Deux types d'algorithmes de décodage de modèles connectés peuvent être distingués : les approches de recherche en arbre basées sur l'algorithme A* (ou *décodages à pile*) et les approches de recherche en treillis basées sur la *programmation dynamique*. Pour chaque catégorie, les différentes variantes de l'état de l'art sont présentées.

1.4.3.1. Décodages à pile (Stack Decoders)

Les algorithmes de décodage à pile sont fondés sur l'algorithme A* [Nilsson, 1971]. Le principe est de maintenir une pile d'hypothèses partielles (*théories*). Ces hypothèses correspondent à des chemins partiels qui ne sont pas forcément de même longueur. Elles sont classées par ordre décroissant de vraisemblance acoustique. Lors d'une itération, la meilleure hypothèse est prolongée par tous les mots possibles. Les nouvelles hypothèses engendrées sont insérées dans la pile. Ce processus est répété jusqu'à ce qu'une des hypothèses atteigne la fin de la phrase (émission de la dernière trame).

L'admissibilité de A* (*i.e.* la garantie de trouver le chemin optimal) est montrée lorsque le coût total d'un chemin est évalué à l'aide de la vraisemblance du chemin partiel et d'une fonction de coût heuristique qui sous-estime la vraisemblance jusqu'à la fin. Pour le décodage à pile, la fonction de coût heuristique influe beaucoup sur la capacité de l'algorithme à ne pas parcourir trop de chemins inutiles. Plusieurs solutions peuvent être envisagées pour l'estimation de cette fonction allant de l'évaluation de la variation moyenne de la vraisemblance sur l'ensemble d'apprentissage à l'utilisation parallèle de plusieurs piles [Jelinek, 1997].

Un certain nombre de techniques d'optimisation ont été proposées pour l'algorithme de décodage à pile. La plus utilisée est le *fast-match* [Waast, Bahl *et al.*, 1995]. Cette technique réduit le coût de l'algorithme en permettant de ne choisir de prolonger que les hypothèses les plus prometteuses. Dans sa version la plus répandue, elle consiste à prédéterminer les phonèmes suivants (de 1 à 3) de sorte à ne prolonger le chemin considéré que par les mots débutant par ces phonèmes. L'algorithme de décodage n'est alors plus admissible. Le *fast-match* présente l'intérêt de ralentir l'expansion de la pile des solutions partielles.

1.4.3.2. Programmation dynamique

Les principaux algorithmes basés sur la programmation dynamique pour la reconnaissance de mots connectés sont : le *One-Pass* (OP) [Vintsyuk, 1971; Bridle, Brown *et al.*, 1982], le *Two-Levels* (TL) [Sakoe, 1979] et le *Level-Building* (LB) [Myers,1981]. Ils ont été énoncés dans le cadre de la reconnaissance par DTW puis étendus au cas des HMM en s'appuyant sur l'algorithme de Viterbi. Globalement, ces trois algorithmes, et leurs variantes développées au cours du temps, résolvent le même problème d'optimisation et donc parviennent à la même solution optimale. Leurs différences résident dans leur organisation, leur complexité et les contraintes qui peuvent être introduites.

La technique de reconnaissance de mots connectés à deux niveaux (*Two-Levels*) fut la première application pratique du DTW aux mots connectés. Son principe est de trouver pour toutes les portions du message les formes de référence optimales puis de trouver la meilleure manière de les connecter. Deux processus d'alignement sont nécessaires, un au niveau des unités de bases, l'autre au niveau du message. L'algorithme LB fut développé comme une amélioration du TL ; il conduit surtout à un coût de calcul moins élevé.

Le LB consiste à rechercher d'abord le meilleur mot pour le début de la phrase sur toutes les trames. Il détermine ensuite les meilleurs second mots à partir des meilleurs premiers mots. Ce processus est répété jusqu'aux meilleurs $L^{\text{ièmes}}$ mots. L doit être fixé au préalable, ce qui est une contrainte puisqu'on ne connaît pas le nombre de mots de la phrase. Toutefois, la solution optimale au rang L restera optimale pour les rangs supérieurs. Il suffit donc de choisir L suffisamment grand ce qui, en contrepartie, fait perdre beaucoup d'intérêt à la méthode. Les décodages par TL et LB nécessitent des aller-retour sur la suite de vecteurs acoustiques à décoder ; ils ne peuvent donc être implantés en temps réel. La complexité du TL $o(T^2)$, avec T la longueur en trames du message à décoder, est élevée. Celle du LB est plus raisonnable $o(LT)$, L étant le nombre maximal d'unité présente dans le message. Mais dans les deux cas, la modélisation des niveaux (*levels*) interdit des structures de recherche en boucle.

Le *One-Pass* n'a aucun des défauts précédents : le décodage est obtenu après un simple parcours du message, sa complexité est en $o(T)$ et des structures de recherche en boucle peuvent être envisagées. Le OP traite toutes les références en parallèle et progresse de manière synchrone trame par trame. Son principe est de réintroduire les sous-chemins optimaux atteignant la fin d'un modèle vers le début de tous les autres modèles. Cet algorithme est à la base des moteurs de reconnaissance de la parole actuels. L'intégration de techniques de recherche en faisceaux (*beam-search*) a pu être réalisée [Bridle, Brown *et al.*, 1982]. Le principe de la recherche en faisceaux consiste à ne pas propager toute solution partielle dont la probabilité est inférieure de plus d'un certain seuil par rapport à la probabilité de la meilleure solution.

[Godin & Lockwood, 1989] propose un schéma des algorithmes de reconnaissance de mots connectés basés sur la DTW au sein duquel les trois algorithmes TL, LB et OP sont équivalents. Les TL et LB sont notamment traduits de façon à pouvoir être exécutés de manière synchrone (sans aller-retour). Parallèlement, [Young, Russel *et al.*, 1989] propose l'algorithme de propagation de jetons (*Token Passing*, TP). Cet algorithme, implémenté dans HTK (voir Annexe C), repose sur une conception objet de la programmation dynamique qui permet d'introduire facilement et à plusieurs niveaux des contraintes sur l'enchaînement des modèles connectés. Ce dernier point a favorisé l'introduction des modèles de langage de type n-gramme et n-classe lors du décodage.

1.4.4. Formalisme du passage de jeton

L'algorithme du passage de jeton propose une nouvelle formulation de l'algorithme de Viterbi compatible avec les contraintes de la reconnaissance de mots connectés. Il peut être considéré comme un algorithme de parcours de graphe fini quelconque et permet ainsi d'unifier les problématiques de l'algorithme de Viterbi et de l'algorithme OP.

1.4.4.1. Structure de recherche

Dans le modèle de propagation de jetons, l'ensemble des contraintes sur l'enchaînement des modèles acoustiques représenté par la structure de recherche, est compilé en un ensemble de nœuds reliés par des arcs valués. Ce sont les états des HMM ou un vecteur de référence pour le DTW. Un nœud encapsule chaque état de tous les modèles acoustiques. Chaque nœud contient un jeton qui représente une solution partielle du décodage de la phrase-test à un instant t. En terme de chemin, cette solution partielle est constituée du sous-chemin décrivant la suite de mots déjà décodée et du chemin (dans le sous-réseau de modèles acoustiques) arrivant à l'état markovien (du modèle acoustique) associé au nœud considéré. La valeur de l'arc représente une probabilité d'émission $d_j(t)$ (mise à jour au décodage de chaque trame de test) à laquelle on peut ajouter une valeur fixe dépendant des contraintes linguistiques p_{ij} (les transitions inter-états ou le modèle de langage des bigrammes, par exemple). Parmi l'ensemble des nœuds, on étiquette ceux qui correspondent aux fins de mots.

Il est à noter que cette approche n'est pas en contradiction avec la définition générale des structure de recherche (cf. §1.4.2). Son intérêt réside dans une formulation plus générale.

1.4.4.2. Algorithme du passage de jeton

Au décodage de chaque trame de test : pour tout nœud, on propage son jeton vers les nœuds suivants en incrémentant sa valeur de la valuation de l'arc emprunté. A partir de l'ensemble des jetons arrivant sur un nœud, on attribue un nouveau jeton au nœud (le meilleur jeton au sens probabiliste). Si le nœud est étiqueté comme fin de mot, on ajoute celui-ci au sous-chemin optimal associé au jeton. La formulation détaillée de l'algorithme est donnée dans la Figure 1-6.

Une itération de l'algorithme est illustrée dans la Figure 1-7. Trois jetons sont transmis à l'état étudié (3) après ajout à leur vraisemblance de la valeur de l'arc emprunté (p_{13}, p_{23} ou p_{33}) et de la probabilité d'émission $d_3(t)$. Parmi les trois, le jeton de plus grande vraisemblance est associé à l'état.

La complexité de cet algorithme est donc égale au coût de traitement d'une trame multiplié par le nombre de trames. Le coût de traitement d'une trame se décompose en une partie acoustique et une partie linguistique. La complexité de la partie acoustique est en $o(N_{mod})$ avec N_{mod} le nombre de modèles acoustiques.

```
•  Initialisation
     Un jeton de valeur nulle est attribué à tous les états
     initiaux. Les autres reçoivent un jeton de valeur infinie ;
•  Traitement
     Pour t=1 à T faire :
           Pour tous les états i faire :
                 Passer une copie du jeton de l'état i vers tous
                 les états connectés j, en incrémentant sa valeur
                 de pij+dj(t);
           Fin de faire
           Pour tous les états i faire :
                 Trouver le jeton de plus petite valeur dans l'état
                 i, éliminer les autres;
           Fin de faire
     Fin de faire
•  Arrêt
     Examiner tous les états finaux, le jeton avec la plus petite
     valeur correspond au meilleur score d'alignement ;
```

Figure 1-6 Algorithme du passage de jeton.

La complexité de la partie linguistique dépend du type de modèle de langage utilisé. Dans l'algorithme, les contraintes linguistiques interviennent entre deux mots (sur un arc). Le modèle le plus naturel à utiliser est donc le bigramme. On peut toutefois envisager le recours au modèle des trigrammes. Or, l'utilisation optimale d'un modèle des trigrammes (ou des triclasses) nécessiterait la remise en cause du mot précédent les deux considérés. Cela est envisageable par une nouvelle formulation de l'algorithme mais aboutirait à un coût de calcul en $o(N_{moy}^3)$ avec N_{moy} nombre moyen de jetons en fin de mot à chaque trame. Or, en général, N_{moy} est égal au nombre de mot du lexique et le coût de calcul en $o(N_{moy}^3)$ devient rapidement prohibitif.

Les techniques de décodage présentées dans ce paragraphe permettent de résoudre le problème de la reconnaissance de mots connectés. Toutefois, afin de diminuer le coût de calcul global du décodage, il peut être réalisé en plusieurs étapes.

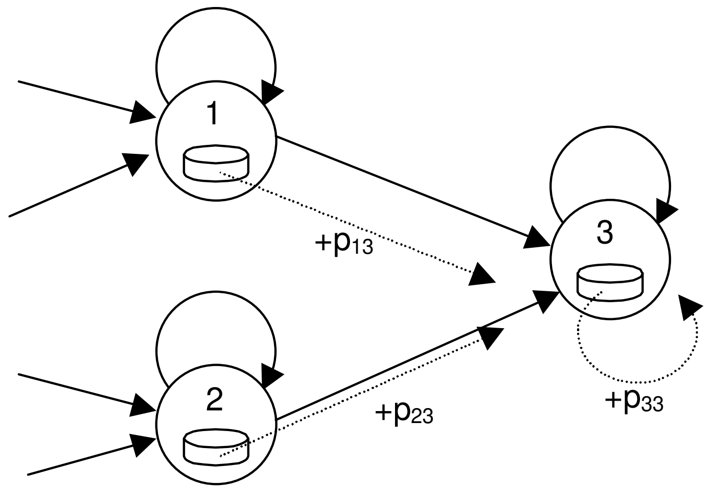


Figure 1-7 Illustration de l'algorithme du passage de jeton.

1.4.5. Décodage en passes multiples

De plus en plus, les systèmes de RAP continue font appel à plusieurs passes de décodage. Le principe est d'utiliser les informations fournies par les résultats des passes précédentes pour diminuer la complexité ou améliorer la précision des passes suivantes. Les informations transmises entre les passes peuvent être de plusieurs types : la transcription de la phrase, les N meilleures transcriptions ou un graphe d'hypothèses. Un graphe d'hypothèses est constitué d'un ensemble de mots reconnus par l'algorithme de décodage avec lesquels sont mémorisés le segment temporel auquel ils sont associés et leur score acoustique. Un graphe d'hypothèses conserve en sus l'identité du mot précédent. Lorsque cette information n'est pas conservée, il s'agit d'un treillis d'hypothèses. La manière la plus simple d'obtenir un graphe d'hypothèses consiste à retenir lors du décodage les N meilleurs chemins en fin de mots pour chaque trame. Une présentation plus détaillée des graphes d'hypothèses est donnée dans l'Annexe D.

Différents types de passes peuvent être distingués en fonction de l'utilisation qui est faite de l'information extraite :

- *diminution de l'espace des solutions* : dans ce cas, le principe est de réduire progressivement l'espace des solutions de sorte à appliquer des connaissances de plus en plus précises et coûteuses en temps de calcul. L'espace des solutions initialement représenté par la structure de recherche est transmis entre les passes sous forme des N meilleures solutions ou d'un graphe d'hypothèses.

  Classiquement, l'augmentation de la complexité des systèmes entre les passes se concrétise au travers des modèles phonétiques (le principal critère, dans le cas de l'emploi d'estimateurs de type GMM, étant le nombre de gaussiennes total dans le système) et au travers des modèles de langage qui passent classiquement des bigrammes aux trigrammes.

L'algorithme de recherche avant-arrière (*forward-backward search*) est un exemple de décodage en passes multiples [Austin, Schwartz *et al.*, 1991]. Un premier passage est effectué avec l'algorithme de Viterbi (*forward search*). Les scores partiels, mémorisés sur chaque trame pour l'ensemble des mots actifs, sont utilisés par un second passage rétrograde (*backward search*). Cette technique permet de générer une liste des N meilleures solutions en ne stoppant l'algorithme qu'après que N solutions partielles aient abouties [Schwartz & Chow, 1990; Schwartz & Austin, 1991]. L'ordre des passes peut être inversé : arrière d'abord et avant ensuite [Valtchev, Odell *et al.*, 1994]. Mais, dans ce cas, le système ne peut fonctionner en temps réel.

Une approche originale consiste à réaliser une première passe de décodage phonétique afin de générer un treillis phonétique qui permet de sélectionner le vocabulaire actif utilisé lors du décodage lexical ultérieur [Montacié, Caraty *et al.*, 1997]. Cette passe peut-être considérée comme un *fast-match* préalable au décodage qui présente l'intérêt d'être global à la phrase. Un autre intérêt de cette technique est de permettre une adaptation des modèles de langage pour les passes suivantes.

- *adaptation des modèles acoustiques* : l'architecture multi-passes peut aussi servir à adapter dynamiquement les modèles acoustiques au locuteur de test. Dans ce cas, une transcription automatique des données disponibles du locuteur est réalisée. Ces données sont utilisées pour adapter les modèles à l'aide d'un ré-apprentissage de type MAP ou MLLR (voir §1.3.3.3). Puis, une nouvelle transcription est obtenue à l'aide des modèles adaptés lors des passes suivantes.

Le type, le nombre et l'ordre des passes sont soumis à de grandes fluctuations dans les systèmes de l'état de l'art. On peut s'en convaincre en passant en revue les systèmes engagés dans les évaluations DARPA [DARPA, 1997; DARPA, 1998] et AUPELF-UREF (Annexe D). Le choix de ces différents paramètres est issu de l'expérience des concepteurs des systèmes et des besoins particuliers de la tâche. Un exemple de système de dictée vocale à quatre passes est donné dans l'Annexe D : détection du genre du locuteur, extraction d'un lexique dynamique, décodage par l'algorithme TP et réorganisation du graphe d'hypothèses. La dernière passe est constituée de l'algorithme des probabilités conjointes qui extrait par programmation dynamique la meilleure phrase d'un graphe d'hypothèses à l'aide des modèles des trigrammes et triclasses [Caraty, Barras *et al.*, 1996].

Le décodage par passes multiples n'est évidemment pas un cadre idéal car il nécessite de fractionner les connaissances. Il serait préférable de pouvoir faire cohabiter l'ensemble des connaissances lors d'une unique passe de décodage basée sur un algorithme de recherche de type A* afin notamment d'optimiser les influences réciproques des différents niveaux de connaissances (acoustiques, lexicaux et syntaxiques). D'ailleurs, les concepteurs des systèmes tentent de diminuer le nombre de passes de leur système afin de se rapprocher de ce cadre. Le système de reconnaissance de BBN est ainsi passé de 4 à 2 passes de décodage sans perte notable de performance et avec, de surcroît, une diminution de l'effort de calcul de l'ordre d'un tiers [Nguyen & Schwartz, 1997].

Pour achever cette présentation des HMM en RAP, un dernier point, qui n'est généralement pas pris en compte dans les systèmes de l'état de l'art, est la mise en œuvre d'un mécanisme de contrôle temporel lors du décodage.

1.4.6. Contrôle temporel

La modélisation implicite de la loi de durée d'un état constitue l'une des faiblesses majeures des HMM. On peut montrer que sous certaines hypothèses la loi de durée implicite d'un modèle à un état suit une loi exponentielle décroissante. Or, il semble que les unités phonétiques de la parole suivent plutôt une loi Gamma [Crystal & House, 1982]. Certains auteurs ont introduit des modèles de durée qui corrigent les probabilités des modèles phonétiques, par exemple avec une loi de Poisson ou une loi Gamma. Le défaut de ces approches est de proposer une modélisation statistique *a priori* des durées des phonèmes sans référence au signal à décoder. Une approche du contrôle temporel obtenu à partir du signal à décoder a été proposée dans [Barras, 1996].

Le principe est de guider l'enchaînement des modèles phonétiques au cours du décodage par la détection des changements acoustiques. Les discontinuités dans le signal sont trouvées à l'aide, par exemple, de la méthode des ruptures de modèles [André-Obrecht, 1988]. L'étude statistique de la répartition des ruptures détectées sur une base de données de référence montre une forte corrélation entre leur position et les instants de transitions entre phonèmes donnés par l'étiquetage fin. La détection d'une rupture peut donc être interprétée lors du décodage comme un instant favorable pour une transition entre les modèles représentant les unités phonétiques. Ceci est concrétisé dans le processus de reconnaissance par l'introduction d'une pénalité dans la vraisemblance des chemins passant d'un modèle à un autre en dehors du voisinage d'une rupture. La taille du voisinage et la valeur de la pénalité sont optimisées sur l'ensemble d'apprentissage. Cette technique présente l'avantage de ne pas être dépendante de la modélisation phonétique utilisée. Une amélioration de 2% a été relevée sur le DAP de la base TIMIT [Barras, 1996].

1.5. Conclusion

Au cours de ce premier chapitre, les principales problématiques qui concourent à l'application de la modélisation markovienne à la RAP continue ont été développées et les solutions de l'état de l'art présentées. La construction d'un système de RAP continue, grand vocabulaire et indépendant du locuteur, tel que le système de dictée vocale D-DAL que nous avons développé au LIP6 (voir Annexe D), découle de la combinaison de ces techniques.

Bien qu'un système de reconnaissance soit construit sur un certain nombre de composants identifiables, l'influence de chacun d'entre eux dans la performance globale du système est souvent difficile à dégager. Lors des campagnes d'évaluation, les tests contrastifs visent à mesurer plus particulièrement l'apport de certains composants en figeant les autres. Cette démarche nécessite de déterminer à chaque étape de la construction d'un système de reconnaissance la configuration optimale avant l'intégration des composants suivants. Afin de ne pas fractionner excessivement ce processus, deux étapes principales peuvent être distinguées. Tout d'abord, un système de base est construit qui intègre les composants minimaux permettant la mise en œuvre du système sur la tâche concernée. Puis, l'intégration des techniques plus évoluées permet la migration vers un système de référence dont les performances sont représentatives de l'état de l'art.

Le système de base, qui sera considéré dans cette étude, est fondé sur une modélisation phonétique indépendante du contexte, avec une topologie gauche-droite à 3 états, un estimateur de probabilité basé sur les GMM et des vecteurs de paramètres composés de coefficients MFCC et de l'énergie. Les modèles sont d'abord appris à l'aide de l'algorithme de Baum-Welch, puis affiner par une réestimation connectée. Le décodage des données est réalisé au travers d'un structure de recherche linéaire par l'algorithme TP. Le modèle de langage est un bigramme phonétique de type back-off dont le seuil est fixé à 1. A partir de cette configuration, l'évolution vers un système de référence sera obtenue par l'intégration des coefficients différentiels du premier ordre et un passage à une modélisation contextuelle avec partage de paramètres. Des expériences de décodage acoustico-phonétique (DAP) doivent permettre d'évaluer à chaque étape les performances des systèmes.

La modélisation markovienne dans les systèmes de RAP continue a permis d'obtenir des résultats qui n'avait jamais été atteint jusqu'alors. Ces performances doivent toujours être améliorées. Pour y parvenir, il est nécessaire de s'affranchir pas à pas des hypothèses non réalistes sur lesquelles reposent la mise en œuvre des HMM. L'hypothèse d'indépendance conditionnelle entre les trames, par exemple, a déjà fait l'objet de recherches qui ont conduit à des solutions variées comme les HMM du second ordre [Mari, Haton *et al.*, 1997] ou les modèles segmentaux [Ostendorf, Digalakis *et al.*, 1996]. Les voies d'amélioration des modèles markoviens appliqués à la RAP continue restent nombreuses. Parmi celles-ci, l'augmentation de la discrimination entre modèles nous semble de nature à apporter un gain substantiel à l'ensemble du processus de reconnaissance.

Pour atteindre cet objectif, notre proposition consiste en l'introduction de l'estimateur de probabilité discriminant des K-ppv dans le formalisme markovien. La validation de notre approche se fera en quatre étapes :

- justification de l'introduction de la discrimination au niveau local par rapport aux méthodes discriminantes globales (chapitre 2) ;
- validation de la qualité de l'estimation locale des K-ppv par rapport à l'estimateur de l'état de l'art GMM (chapitre 2) ;
- introduction de l'estimateur des K-ppv dans un système HMM de base et évaluation comparée de ce système avec un système de base HMM/GMM (chapitre 3) ;
- évolution du système de base vers un système de référence par l'intégration de techniques performantes (coefficients différentiels et modélisation contextuelle avec partage de paramètres) (chapitre 4).

Notons, finalement, que de notre point de vue, l'intérêt de l'introduction de l'estimateur des K-ppv dans un système markovien ne se limite pas à l'amélioration de performances que nous pouvons en espérer. Il doit aussi permettre de mieux comprendre le fonctionnement des systèmes markoviens. Cette analyse sera, en effet, facilitée par la nature non-paramétrique de l'estimateur (l'absence de modèles augmente la visibilité du rôle joué par les données dans les HMM). De même, l'opportunité qu'il offre de comparer son comportement à celui des GMM devra être mise à profit pour isoler les éléments déterminants de l'estimation probabiliste des données dans les HMM.

2. DISCRIMINATION DANS LES SYSTEMES MARKOVIENS

La discrimination dans un opérateur de reconnaissance caractérise sa capacité à séparer correctement les classes de données. Dans l'approche statistique de la reconnaissance, elle est donc liée à ses performances de reconnaissance au travers du principe de décision bayesien du maximum *a posteriori*.

Le problème posé dans le cas d'un opérateur de reconnaissance basé sur les modèles markoviens est que le critère utilisé pour son apprentissage n'est pas lui-même fondé sur le principe de maximum *a posteriori* mais du maximum de vraisemblance. Ce critère est généralement montré comme optimal du point de vue de la décision bayesienne pourvu qu'un nombre suffisant de données d'apprentissage soient disponibles. Or, plus fondamentalement, son optimalité repose sur l'hypothèse qu'il existe un jeu de paramètres avec lequel le modèle puisse représenter exactement les données utilisées. Cette hypothèse n'étant pas vérifiée en pratique, l'apprentissage par maximum de vraisemblance présente le défaut de ne pas chercher explicitement à introduire de la discrimination dans l'opérateur de reconnaissance.

Une plus grande discrimination dans les modèles de Markov peut donc, en premier lieu, être recherchée dans la remise en cause des techniques d'apprentissage des HMM. Les méthodes généralement utilisées pour obtenir des HMM plus discriminants consistent à poursuivre l'apprentissage des modèles standards en introduisant un critère discriminant. Deux types de solutions existent :

- les méthodes reposant sur la maximisation d'un critère d'*information mutuelle* entre les données et les modèles ;
- les méthodes, plus directes, reposant sur un critère de *réduction des erreurs*.

Ces approches ont en commun d'agir globalement sur les modèles. L'approche que nous proposons repose sur l'hypothèse que la discrimination globale (au niveau du modèle) peut être obtenue à partir d'une discrimination locale (au niveau des états markoviens). Pour cela, un certain nombre de méthodes classiques en reconnaissance des formes peuvent être envisagées :

- *la transformation linéaire de l'espace de représentation* par une analyse en composantes principales (ACP) ou par une analyse linéaire discriminante (ALD) ;
- *l'interpolation non-linéaire des frontières* des classes par réseaux de neurones ;
- *les estimateurs de probabilité non-paramétriques* : noyaux de Parzen ou K-ppv.

En l'absence d'hypothèses valables sur la forme réelle des données, l'estimateur non-paramétrique des K-ppv paraît un choix intéressant. Les performances de discrimination locale de l'estimateur K-ppv sont évaluées sur une tâche d'identification locale au travers de la règle de décision des K-ppv. Ces résultats sont comparés à ceux obtenus dans les mêmes conditions par l'estimateur de l'état de l'art à base de GMM.

2.1. Méthodes globales d'apprentissage discriminant

Différentes méthodes d'apprentissage ont été proposées pour améliorer les capacités de discrimination des HMM. Elles peuvent être séparées en deux grands groupes : celle issues de l'application de la théorie de l'information à l'estimation des paramètres ; celles fondées sur le point de vue plus pragmatique de la réduction des erreurs.

2.1.1. Critères de mesure de l'information

L'ensemble des techniques d'apprentissage issues de la théorie de l'information est fondé sur un critère maximisant l'entropie entre les données et les modèles. Les différences résident dans le type d'entropie retenue. Les formules permettant le calcul de l'entropie et de l'information mutuelle sont rappelées.

2.1.1.1. Entropie et information mutuelle

Dans le cadre de la théorie de l'information [Oswald, 1986], une source d'information est un système capable d'émettre des symboles pris parmi un ensemble fini. Ces symboles sont émis avec une certaine probabilité. Lorsqu'une source émet un symbole, elle diminue l'incertitude de l'observateur : elle lui fournit de l'information. Ainsi, la quantité d'information d'une source mesure l'incertitude qu'elle résoudra par l'émission d'un de ses symboles. Si nous prenons, par exemple, un ensemble de L éléments équiprobables et émis de façon indépendante, l'information contenue par la source est :

$$I = \log(L) \tag{2-1}$$

Le choix d'une mesure logarithmique provient de sa conformité avec quelques exigences intuitives. On voudrait, par exemple, que la quantité d'information d'une source dont les symboles sont regroupés par N soit N fois supérieure à celle de la source de base. Or, la nouvelle source dispose de L^N symboles. Sa quantité d'information est donc :

$$I' = \log(L^N) = N \times \log L = N \times I \tag{2-2}$$

Ce qui est conforme à notre intuition. La base de logarithme utilisée définit l'unité de l'information. En prenant une base 2, on obtient 1 pour une expérience ayant 2 issues équiprobables. L'unité est alors le bit.

Lorsque les symboles sont émis avec une loi de probabilité *P(X)* (avec *X* variable aléatoire de la sortie de la source dont *x* est une réalisation), on mesure la quantité d'information de la source par son entropie donnée par :

$$H(X) = -\sum_{x} P(x) \times \log P(x) = -E\left[\log(P(X)\right] \qquad (2\text{-}3)$$

Si la source est uniforme (*i.e.* *P(X)=1/L*), on retrouve bien *I* (2-1).

L'entropie mesure l'incertitude liée à la source. Et il semble naturel de mesurer la quantité d'information d'un système par la diminution de son entropie. Ainsi, la quantité d'information totale d'un système est obtenue lorsqu'il a été complètement déterminé. Son entropie est alors nulle :

$$I(X) = H(X) - 0 = H(X) \qquad (2\text{-}4)$$

Considérons maintenant 2 sources, représentées par les variables aléatoires *X* et *Y* de réalisations respectives *x* et *y*, l'entropie permet de mesurer le niveau de relation entre les événements liés à chacune des sources. Une mesure de la quantité moyenne d'incertitude de la variable *X*, *Y* étant connue, est définie par l'entropie conditionnelle :

$$H(X/Y) = -\sum_{x,y} P(x,y) \log P(x/y) = -E[\log P(X/Y)] \qquad (2\text{-}5)$$

Il est possible d'estimer la quantité d'information fournie par *Y* sur *X* comme :

$$I(X;Y) = H(X) - H(X/Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \qquad (2\text{-}6)$$

On peut remarquer que *I(X ; Y) = I(Y ; X)*. Cette quantité est appelée l'information mutuelle moyenne entre *X* et *Y*. Les relations entre les différentes entropies (marginales, conditionnelles et conjointes) et l'information mutuelle de *X* et *Y* sont représentées dans la Figure 2-1.

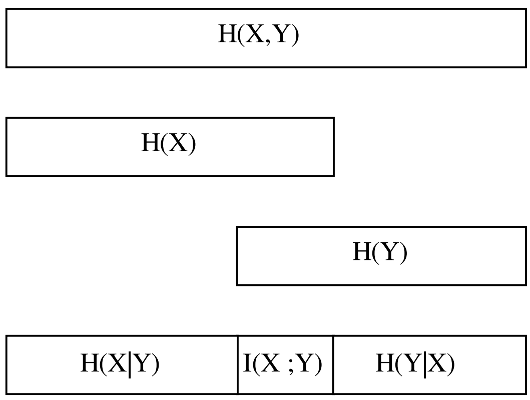


Figure 2-1 Relation entre les entropies marginales *H(X)* et *H(Y)*, conditionnelles *H(X/Y)* et *H(Y/X)*, conjointe *H(X,Y)* et l'information mutuelle *I(X;Y).*

2.1.1.2. *Théorie de l'information et reconnaissance de la parole*

Du point de vue de la théorie de l'information, le problème de la reconnaissance de la parole revient à considérer les séquences d'observations acoustiques comme le résultat du codage de séquences de mots par un canal acoustique probabiliste. Ainsi, M est la variable aléatoire représentant les séquences de mots prononcées et O la variable aléatoire représentant les séquences d'observations acoustiques obtenues. L'incertitude sur une phrase m considérant la séquence acoustique observée o est donnée par l'entropie conditionnelle $H(M/O)$ de M par rapport à O. Ce terme peut être interprété comme le nombre minimal de bits nécessaire pour spécifier M connaissant O. Il exprime l'incertitude qui perdure sur le message malgré la connaissance de l'acoustique. L'entropie conditionnelle est liée à la probabilité *a posteriori* du message sur l'acoustique par :

$$H(M/O) = -E[\log P(M/O)] \qquad (2\text{-}7)$$

Comme nous l'avons introduit dans le paragraphe précédent, les quantités impliquées dans (2-7) sont évaluées à partir de la connaissance des lois de distribution de M et O et des distributions conjointes de M et O. Or ces dernières n'étant pas connues en pratique, elles doivent être estimées à partir de modèles. On peut montrer que l'entropie (2-7) est minimale pour la vraie loi de probabilité et donc :

$$H_\Lambda(M/O) \geq H(M/O) \qquad (2\text{-}8)$$

où Λ représente l'ensemble des modèles sur lesquels sont évaluées les distributions conjointes et marginales de M et O. Le but de l'apprentissage des modèles est donc de diminuer le plus possible la valeur de l'entropie conditionnelle mesurées au travers des modèles de façon à approcher au mieux la valeur correspondant aux lois de probabilités réelles. L'équation (2-7) nous assure de la conformité de cet objectif au principe du décodage par maximum *a posteriori* puisqu'alors les probabilités *a posteriori* des messages sur leur acoustique seront maximisées.

S'appuyant sur la règle de Bayes, le critère du MLE atteint cet objectif par l'augmentation de la probabilité conditionnelle $P_\lambda(O/M)$ de M et O (cf. §1.1.3.3). Dans le cas de l'apprentissage segmental, cette probabilité peut se ramener à :

$$P_\lambda(O/M) = P(O/\lambda)P(M) \qquad (2\text{-}9)$$

avec λ le HMM modélisant M. $P(M)$ ne dépend par de λ, cette probabilité n'est pas prise en compte dans la maximisation (1-20). L'entropie conjointe et la probabilité conditionnelle sont liées entre elles par :

$$H_\lambda(M,O) = -E\left[\log(P(O/\lambda)P(M))\right] \qquad (2\text{-}10)$$

et l'entropie conditionnelle à l'entropie conjointe par :

$$H_\lambda(M/O) = H_\lambda(M,O) - H(O) \qquad (2\text{-}11)$$

Le critère du MLE conduit donc bien à la minimisation de l'entropie conditionnelle de M et O sous les hypothèses qu'aucune des probabilités marginales ne dépend du modèle λ. D'autres critères peuvent être utilisés qui varient essentiellement par le type de l'entropie qu'ils optimisent.

2.1.1.3. Estimation par information mutuelle maximale

L'apprentissage par maximisation de l'information mutuelle (*Maximal Mutual Information Estimation*, MMIE), introduit par l'équipe d'IBM [Bahl, Brown *et al.*, 1987], est le paradigme des méthodes discriminantes fondées sur la théorie de l'information.

Présentation

Dans l'approche de l'apprentissage des HMM par MMIE, les paramètres du modèle λ sont estimés au travers de la maximisation de l'information mutuelle maximale entre les modèles et les données d'apprentissage :

$$\hat{\lambda}_{MMIE}(O,M) = \arg\max_{\lambda} I_\lambda(O\,;M) = \arg\max_{\lambda} H(M) - H_\lambda(M/O) \qquad (2\text{-}12)$$

La relation la plus évidente entre les critères MLE et MMIE est qu'ils sont tous les deux conformes au décodage par maximum *a posteriori*. Le premier propose de maximiser une approximation de la probabilité conditionnelle de M et O qui conduit à minimiser l'entropie conditionnelle sous les hypothèses que les probabilités marginales n'ont pas d'influence lors du processus d'apprentissage. Le second propose de maximiser directement une approximation de la probabilité *a posteriori* qui ne fait plus comme hypothèse que l'indépendance de $P(M)$.

Les deux points de vue ne diffèrent donc finalement que par la prise en compte de la probabilité marginale de l'acoustique $P(O)$. Le MLE considère que cette probabilité n'est pas fonction des modèles appris et n'intervient donc pas dans l'apprentissage des modèles. Le MMIE pour sa part cherche en conservant $P_\lambda(O)$ a fournir une meilleure approximation de la probabilité *a posteriori*. Ce faisant il fait intervenir un terme qui prend en compte l'ensemble des modèles : il est donc discriminant.

Variante

L'estimation par MMIE est à rapprocher de l'estimation par maximum de vraisemblance conditionnelle (*Conditional Maximum Likelihood Estimation*, CMLE) [Nadàs, 1983] :

$$\hat{\lambda}_{CMLE}(O,M) = \arg\max_{\lambda} - H_\lambda(M/O) = \arg\max_{\lambda} E\left[\log P_\lambda(M/O)\right] \qquad (2\text{-}13)$$

Les estimateurs CMLE et MMIE ne différent que par le terme représentant l'entropie du modèle de langage (liée à la probabilité *a priori P(M)*). Or, les paramètres du modèle de langage sont évalués préalablement à l'estimation des paramètres des modèles acoustiques. La minimisation de *H(M)* est obtenue en choisissant un modèle de langage ayant la plus petite perplexité sur les données d'apprentissage. L'hypothèse, émise pour l'apprentissage selon le critère du MMIE, que *H(M)* est constante lors de l'apprentissage des paramètres des HMM étant vérifiée en pratique, les estimateurs MMIE et CMLE sont alors équivalents.

<u>Mise en œuvre</u>

Le MMIE ne conduit pas à des formules de réestimation analytiques comme le MLE. Sa mise en œuvre repose sur des techniques d'optimisation de type descente de gradient [Bahl, Brown *et al.*, 1987]. Plus récemment, une extension de l'algorithme de Baum-Welch (*Extended Baum-Welch*) a permis de proposer une procédure d'apprentissage itérative de type EM [Gopalakrishnan, Kanevsky *et al.*, 1991]. Sa mise en œuvre reste complexe du fait de la présence d'une constante dans les formules de réestimation dont la valeur doit être déterminée finement car elle assure le compromis entre la garantie de convergence et la rapidité avec laquelle elle est atteinte.

De façon générale, l'apprentissage par MMIE est plus complexe à mettre en œuvre que le conventionnel MLE. Il est aussi plus exigeant en coût de calcul : lorsque que le MLE ne considère qu'un seul modèle, le MMIE s'intéresse à tous simultanément, et les méthodes mises en œuvre sont par nature plus coûteuses. [Normandin, Cardin *et al.*, 1994] relève que l'apprentissage par MMIE peut s'avérer 35 fois plus coûteux que le MLE sur la même tâche.

Un certain nombre de chercheurs ont rapporté des améliorations grâce au MMIE. Une des premières applications notables du critère du MMIE a été rapportée par [Normandin, Cardin *et al.*, 1994] qui diminue de 40% les taux d'erreur sur une tâche de reconnaissance de chiffres connectés. Plus récemment, des améliorations ont été rapportées sur un ensemble de bases significatives. Sur TIMIT (Annexe E), l'apprentissage d'un système HMM/24-GMM par le critère du MMIE permet une réduction de 1,75% du taux d'erreur [Valtchev, 1995]. Sur la base *Wall Street Journal* (voir Annexe D), des réductions relatives de 5 à 15 % ont été rapportées [Valtchev, Odell *et al.*, 1996].

L'amélioration des performances due au MMIE est plus importante sur les tâches de reconnaissance de mots connectés que pour la reconnaissance de parole continue [Valtchev, 1995]. De même, les résultats d'auto-cohérence sont, en général, très supérieurs aux résultats sur les tests indépendants. Il est à craindre que le critère du MMIE ne conduise à un sur-apprentissage des modèles. Finalement, les résultats expérimentaux montrent que l'apport du critère du MMIE évolue en proportion inverse de la précision de modélisation. Ainsi, sur la base TIMIT, un système de monophones à une gaussienne par état voit son taux de reconnaissance progresser de 5,5% sur le test après un ré-apprentissage par le MMIE quand, dans les mêmes conditions, un système à 24 gaussiennes par état ne progresse que de 1,8% [Valtchev, 1995].

Le critère du MMIE constitue actuellement le paradigme des critères discriminants. Il existe toutefois d'autres critères qui présentent des propriétés comparables.

2.1.1.4. Autres critères : MDI, H-critères

L'apprentissage basé sur l'information de discrimination minimum (*Minimum Discrimination Information*, MDI) fut proposé par les chercheurs d'AT&T [Ephraïm, Dembo *et al.*, 1989]. L'information de discrimination est une mesure de proximité entre deux densités de probabilité. Le critère MDI tente d'optimiser les paramètres par minimisation de l'information discriminante de l'acoustique et des modèles. Pour cela, la distribution de la source (*i.e.* de l'acoustique) est caractérisée par ses moments au travers des matrices de covariance partielles évaluées sur les séquences de données. Puis par des projections contraintes itératives, les paramètres des modèles sont évalués de sorte à rapprocher au mieux leur distribution de celle de la source.

Le MLE et le MMIE peuvent être reformulés en terme de MDI [Ephraïm & Rabiner, 1990]. Ce critère a été étudié théoriquement dans le cadre de HMM utilisant des densités de probabilités estimées par des gaussiennes auto-regressives. A notre connaissance, aucune implémentation de ce critère n'a permis de mesurer son efficacité pratique. De plus, son extension à d'autres estimateurs de probabilité ne paraît pas évidente.

La théorie des H-critères [Gopalakrishnan, Kanevsky *et al.*, 1988] fournit un cadre général de description des estimateurs MLE, CMLE et MMIE. Il s'agit de la famille d'estimateurs générée par l'estimateur parametré :

$$H_\lambda(h_1, h_2, h_3) = h_1 H_\lambda(M, O) + h_2 H(M) + h_3 H_\lambda(O) \tag{2-14}$$

Les différents critères sont construits comme une somme pondérée des entropies conjointes et marginales du langage M et de l'acoustique O. Le MLE est le minimum d'entropie conjointe donné par $H_\lambda(1,0,0)$. Le CMLE est le minimum d'entropie conjointe conditionnelle donné par $H_\lambda(1,0,-1)$ et enfin le MMIE est le maximum d'information mutuelle donné par $H_\lambda(1,-1,-1)$. Rappelons que dans les trois cas l'optimisation de $H(M)$ n'est pas prise en compte et donc les entropies réellement optimisées lors du processus d'apprentissage sont données dans le Tableau 2-1.

Critère	Entropie optimisée
MLE	$H(M,O)$ Conjointe
CMLE	$H(M/O)$ Conditionnelle
MMIE	$H(M;O)$ Mutuelle

Tableau 2-1 Critères d'apprentissage et entropies maximisées.

Les H-critères permettent d'exprimer une large classe de critères dès lors que la participation de chacune des entropies à la maximisation peut être pondérée. Les pondérations doivent être fixées *a priori,* ce qui limite la portée de la méthode. Les expériences menées en faisant varier h_3 de 0 à -1 par pas de 0,1 avec $h_1=1$ et $h_2=-1$ ont permis toutefois de constater une amélioration sur un ré-apprentissage par MMIE de base ($h_3=-1$) [Gopalakrishnan, Kanevsky *et al.*, 1988].

Nous devons retenir que les méthodes d'apprentissage discriminant issues de considérations liées à la mesure de l'information reciproque entre les messages linguistiques et leur acoustique diffèrent essentiellement par le choix de la mesure d'entropie maximisée (cf. Tableau 2-1). Il est notable, dans tous les cas, que les écarts de performance constatés après un apprentissage discriminant sont d'autant plus grands que les modèles concernés sont peu précis. Inversement, plus la précision de la modélisation augmente, moins le recours à ces méthodes est justifié.

2.1.2. Critères de réduction des erreurs

Une alternative à l'approche par mesure de l'information consiste à baser l'apprentissage des modèles directement sur leur capacité à classer correctement les données de l'ensemble d'apprentissage. Alors que les méthodes précédentes subordonnent cet objectif à la qualité de la modélisation au travers d'une mesure d'entropie, les méthodes fondées sur la minimisation des erreurs s'attachent directement et uniquement à la finalité des modèles : reconnaître.

2.1.2.1. Apprentissage correctif

L'apprentissage "correctif" (*corrective training*) fût la première tentative de stratégie d'apprentissage fondée sur la minimisation des erreurs [Bahl, Brown *et al.*, 1988 ; Bahl, Brown *et al.*, 1993]. Son principe est de « corriger » l'apprentissage MLE lorsque celui-ci ne classe pas correctement les données et de renforcer les modèles mis en forte concurrence. Son application sur une tâche de reconnaissance de mots isolés dépendante du locuteur a permis une réduction de 16% des erreurs [Bahl, Brown *et al.*, 1993].

La méthode a été adaptée à la RAP continue [Lee & Mahajan, 1990]. Après avoir résolu le problème posé par la prise en compte des insertions et des délétions dans la procédure de réestimation, la méthode, introduite dans le système SPHINX du CMU, a permis un gain de l'ordre de 4% sur le test de la campagne DARPA 1987 (*Ressource Management*).

Malgré des résultats encourageants, il reste que les paramètres des modèles en compétition sont « corrigés » de façon heuristique. Ainsi, même si de gains non négligeables ont pu être obtenus à l'aide de cette méthode, l'absence de formulation théorique rigoureuse reste un problème fondamental qui limite son emploi.

2.1.2.2. Erreur minimale de classification

La technique d'apprentissage discriminante du MCE/GPD consiste à minimiser la fonction objective représentant l'erreur de classification minimale (*Minimum Classification Error*, MCE) à l'aide de la technique discriminante de descente de gradient généralisée (*Generalized Probabilistic Descent*, GPD) [Katagiri, Lee *et al.*, 1991; Juang & Katagiri, 1992].

L'erreur de classification d'une séquence d'observations $o_1 .. o_T$ est donnée par :

$$err(o_1 \ldots o_T) = \begin{cases} 1 \text{ si } o_1 \ldots o_T \in m \text{ et } m \neq D(o_1 \ldots o_T) \\ 0 \text{ sinon} \end{cases} \tag{2-15}$$

L'erreur $err(o_1 .. o_T)$ vaut donc 1 si la règle de décision $D(o_1 .. o_T)$ ne reconnaît pas la phrase m alors que celle-ci est effectivement la phrase prononcée pour $o_1 .. o_T$.

L'erreur théorique globale est évaluée par l'espérance de l'erreur de classification de l'ensemble des séquences d'observations d'apprentissage :

$$err(O) = E[err] \tag{2-16}$$

En pratique, cette équation est remplacée par le calcul de l'erreur empirique sur l'ensemble des exemples :

$$err(O) = \sum err(o_1 \ldots o_T) \tag{2-17}$$

L'optimisation de cette représentation de l'erreur de classification n'est pas possible par les techniques de gradient classiques car elle n'est pas dérivable. Un des apports majeurs de la technique MCE a été l'introduction d'une fonction de l'erreur de classification dérivable.

Le critère de classification minimale s'exprime alors :

$$\hat{\lambda}_{MCE}(O,M) = \arg\min_{\lambda} err_{\lambda}(O) \tag{2-18}$$

avec

$$err_{\lambda}(O) = \frac{1}{1 + e^{-d_{\lambda}(O)}} \tag{2-19}$$

et

$$d_\lambda(O) = -P(O/\lambda) + \max_{\lambda \neq \lambda'} P(O/\lambda') \tag{2-20}$$

avec λ' représentant le modèle de plus grande vraisemblance autre que λ. Cette version de la distance est une version simplifiée mais fonctionnelle de la distance généralement présentée. Cette distance permet une mesure continue de l'écart entre la vraisemblance de la solution attendue et de la solution incorrecte la plus proche (il est possible de prendre en compte les N plus proches solutions incorrectes). Ainsi, $d_\lambda >> 0$ indique une erreur de classification et $d_\lambda << 0$ une classification correcte. L'intérêt de cette approche est donc de fournir une représentation dérivable de l'erreur de classification. Elle peut alors être optimisée par un algorithme de type descente de gradient, en l'occurrence le GPD.

La forme générale de la réestimation des paramètres pour l'itération $n+1$ est :

$$\lambda_{n+1} = \lambda_n - \varepsilon_n U_n \nabla err_{\lambda_n}(O) \tag{2-21}$$

où ∇err représente le gradient de l'erreur, U_n une matrice définie positive (en général la matrice diagonale des variances des paramètres) et ε_n une valeur positive représentant le pas d'apprentissage qui permet d'affiner le compromis entre précision et vitesse de la convergence.

Cette méthode d'apprentissage est en définitive très proche de l'apprentissage par MMIE. La fonction objective qui guide les deux procédures d'optimisation utilise les mêmes statistiques basées sur une mesure de l'écart entre les vraisemblances de la solution attendue et des solutions proches. Les résultats de l'application du MCE/GPD, comme le laissent supposer les remarques précédentes, sont assez proches de ceux obtenus avec le critère du MMIE. Nous retiendrons particulièrement que ces travaux conduisent à une conclusion identique : les méthodes d'apprentissage discriminant voient leur intérêt diminuer avec l'accroissement de la précision de la modélisation des observations acoustiques sur les états. L'introduction de la discrimination dans les HMM directement au niveau de la modélisation des observations acoustiques paraît donc une alternative intéressante.

2.2. Méthodes locales discriminantes

Dans les méthodes décrites précédemment, l'amélioration de la discrimination est obtenue par la recherche globale de meilleurs paramètres pour les HMM. Une alternative consiste à introduire explicitement la discrimination localement au niveau de la modélisation des observations sur les états des HMM. Plusieurs approches sont alors possibles : une transformation linéaire de l'espace de représentation, l'interpolation non-linéaire par des réseaux de neurones des frontières des classes ou l'estimation non-paramétrique des probabilités d'émission des états.

2.2.1. Transformations linéaires de l'espace de représentation

Plusieurs approches sont envisageables pour transformer linéairement l'espace de représentation de façon à le rendre plus discriminant. Elles sont issues de méthodes d'analyse de données multidimensionnelles comme l'analyse en composantes principales (ACP) ou l'analyse linéaire discriminante (ALD) (voir, par exemple, [Fukunaga, 1972]).

2.2.1.1. Analyse en composantes principales

Le critère utilisé pour l'ACP est la préservation de l'information. L'hypothèse pratique est que l'information sur les données apportée par une dimension de l'espace de représentation peut être mesurée par la variance des données sur cette dimension. Plus une dimension présente une variance forte, plus l'information qu'elle fournit est conséquente.

Les vecteurs propres de la matrice de covariance de l'ensemble des vecteurs d'apprentissage sont les axes comportant le plus d'information. Leur importance relative est déduite de la valeur propre associée au vecteur propre considéré. L'espace réduit est donc obtenu par la projection de l'espace initial sur les vecteurs propres de la matrice de covariance dans l'ordre décroissant de leur valeur propre.

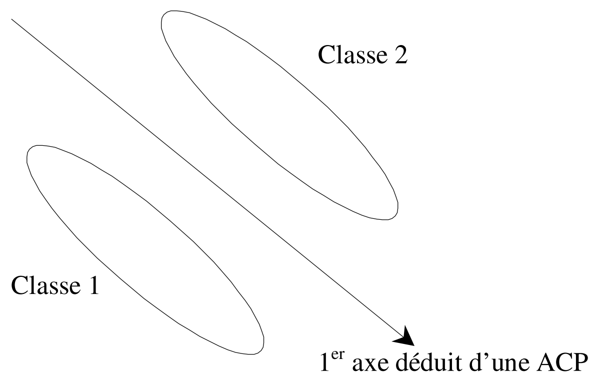


Figure 2-2 Cas d'opposition entre information et discrimination des données.

Le principal inconvénient de cette méthode est que la notion d'information n'est pas liée à la discrimination entre les classes. Comme l'illustre la Figure 2-2, dans certains cas, elles peuvent même être opposées. Il semble donc préférable de recourir à une méthode directement fondée sur un critère discriminant.

2.2.1.2. *Analyse linéaire discriminante*

L'analyse linéaire discriminante est une technique très utilisée en reconnaissance des formes statistiques pour extraire de l'espace de représentation un sous-espace qui favorise la discrimination entre les classes des données. Le critère dérivé peut se résumer par : après projection sur l'hyperplan obtenu, chaque classe doit être la plus regroupée possible, et l'ensemble des classes le plus dispersé possible.

L'ALD est la transformation linéaire qui maximise le *Fisher-ratio* sur l'ensemble des dimensions. Le *Fisher-ratio* est une mesure permettant d'évaluer la séparabilité des classes dans une dimension donnée [Duda & Hart, 1973]. Plusieurs mesures permettent de généraliser la notion de *Fisher-ratio* sur l'ensemble des dimensions de l'espace. Elles sont basées sur le rapport entre les matrices de covariance intra-classe et inter-classe.

L'ALD se résume finalement, comme l'ACP, à la projection de l'espace initial sur les vecteurs propres (dont le nombre retenu fixe la dimension du nouvel espace) d'une matrice. Dans le cas de l'ACP, cette matrice est la matrice de covariance des données. Dans le cas de l'ALD, la matrice est le produit de l'inverse de la matrice de covariance intra-classe par la matrice de covariance inter-classe. Il a été montré que l'ALD est plus efficace dans le cas de classes dont la distribution est unimodale [Fukunaga, 1972].

2.2.1.3. *Applications à la RAP*

Une méthode permettant d'introduire la discrimination localement par des transformations linéaires liées aux états des HMM a été étudiée [Doddington, 1989; Woodland & Cole, 1991]. Le sous-espace obtenu après transformation des paramètres maximise pour chaque état la discrimination entre les données correctes et les données erronées proches.

Pour définir la transformation, deux ensembles de données sont définis : les données correctement classées et les données de confusion. Le premier ensemble regroupe les données associées à un état par les procédures d'apprentissage. Le second contient les données associées à un état de façon incorrecte par un décodage. L'ALD est utilisée pour définir le sous-espace le plus discriminant entre les deux groupes de données. La matrice des covariances inter-classe est calculée après avoir centré les données de confusion sur la moyenne des données correctes.

Cette approche a recours aux mêmes statistiques que la méthode du MMIE. Les données correctes sont obtenues classiquement par alignement des séquences acoustiques sur les modèles correspondant à leur transcription de référence. Les données de confusion sont dérivées d'une reconnaissance (sans grammaire) des séquences acoustiques. Woodland rapporte un gain de 11% par l'application de la discrimination linéaire d’état sur la reconnaissance de la base E-set. Les 24 MFCC+E+D sont projetés sur un sous-espace réduit à 16 coefficients.

Les espaces de représentation en reconnaissance de la parole ont des dimensions élevées et très complexes. Dans ce cadre, une approche linéaire se révèle vite limitée, les surfaces de décision des données étant extrêmement complexes. Les réseaux de neurones offrent une solution non-linéaire au problème de la séparation des classes.

2.2.2. Approches non-linéaires par réseaux de neurones

Les méthodes connexionnistes sont des méthodes paramétriques sans hypothèses sur la forme réelle des densités de probabilité des données. Plutôt que la répartition des données dans leur espace, les méthodes connexionnistes tentent d’estimer au mieux les surfaces séparatrices des classes. Ces surfaces sont modélisées au travers d’une suite de fonctions non-linéaires dépendante de la structure du réseau. La topologie du réseau conditionne la famille de fonctions qui pourra être estimée.

2.2.2.1. Principes

Trois types de réseaux de neurones sont maintenant classiques en reconnaissance de la parole : les perceptrons multicouches (*Multi-Layer Perceptron*, MLP), les réseaux de neurones à délai temporel (*Time-Delay Neural Networks*, TDNN) et les réseaux récurrents (*Recurrent Networks*, RN). L'architecture de base de ces trois types de réseaux, illustrée Figure 2-3, est celle du MLP [Rumelhart, Hinton *et al.*, 1986].

Un réseau de neurone de type MLP consiste en un ensemble d'unités (les neurones) connectées entre elles par des poids. Les unités sont organisées en couches. Chaque unité d'une couche est liée à toutes les unités de la couche suivante par des poids. L'état d'une unité est défini par son niveau d'activation, somme des activations des unités auxquelles elle est reliée pondérée par les poids des connections.

Ce type de réseaux peut estimer n'importe quelle fonction non-linéaire pourvu qu'un nombre suffisant de couches soit prévu [Lippmann, 1987]. Ce résultat est évidemment théorique et ne tient pas compte des limitations inhérentes à la détermination optimale du nombre de couches et du nombre d'unités les composant ainsi qu'aux procédures d'apprentissage de ces modèles. Ce dernier point reste un des éléments critiques de l'utilisation des réseaux de neurones.

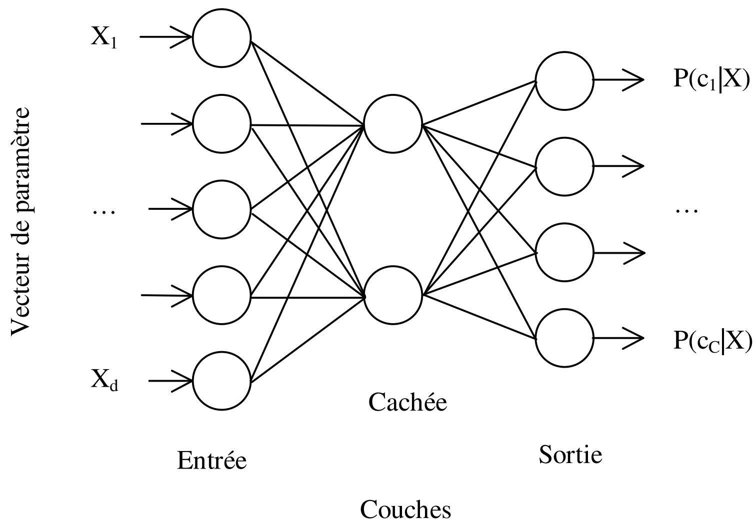


Figure 2-3 Réseaux de neurones de type MLP.

On retiendra qu'au moins 3 couches sont nécessaires[1] pour résoudre des problèmes non linéairement séparables : une couche d'entrée dont le nombre d'unités est égal à la taille du vecteur de paramètres, une couche cachée avec un nombre d'unités fixé empiriquement et une couche de sortie dont le nombre d'unités est donné par le nombre de classes à séparer (voir Figure 2-3). L'apprentissage des poids d'un réseau à plusieurs couches est obtenu par l'algorithme de rétro-propagation du gradient de l'erreur [Rumelhart, Hinton *et al.*, 1986]. Après l'apprentissage, les niveaux d'activation de la couche de sortie peuvent être interprétés comme les probabilités *a posteriori* des classes auxquelles elles ont été associées.

2.2.2.2. Variantes

Les réseaux de neurones de type MLP appliqués à la reconnaissance de la parole présentent un inconvénient majeur qui est le manque de mécanisme de gestion des distorsions temporelles. L'introduction des réseaux de neurones dans le formalisme markovien a permis d'apporter une réponse à ce problème. Avant cela, deux types de réseaux permettaient la reconnaissance de segments d'observations.

Les réseaux de type TDNN possèdent des unités de délai qui permettent de retarder la prise en compte d'un coefficient du vecteur d'entrée. Ainsi, la mise en cascade de telles unités permet de tenir compte des trames adjacentes et modélise donc le contexte. Les réseaux de type RN proposent une approche comparable mais positionnent leurs

[1] Et vraisemblablement suffisantes dans la mesure où il semble qu'un réseau à plusieurs couches cachées est équivalent à un réseau à une couche cachée de taille plus importante [Deroo, 1998].

unités de délai en boucle sur la couche de sortie du réseau. Elles proposent ainsi une partie des unités de sortie des trames précédentes en entrée d'une nouvelle trame (en plus du vecteur de paramètres de celle-ci). Il est à noter que la prise en compte du contexte est aussi réalisée avec les MLP classiques en présentant en entrée du réseau une suite de trames adjacentes (de 5 à 9 généralement).

Bien qu'ils soient théoriquement plus performants que les TDNN, il semble que les RN sont plus difficiles à utiliser d'un point de vue pratique à cause des problèmes de convergence ([Bengio, de Mori *et al.*, 1991] soulève l'instabilité de l'apprentissage et le risque élevé de tomber sur des minima locaux) et du temps d'apprentissage ([Franzini & Waibel, 1991] montre que l'augmentation des performances par rapport aux TDNN est faible, comparée à l'effort de calcul supplémentaire). Ces remarques sont à nuancer depuis les travaux de Robinson [Robinson, 1994].

Les limitations des TDNN et des RN ont conduit à considérer l'utilisation conjointe des réseaux de neurones et des HMM dans le cadre de modèles hybrides.

2.2.2.3. Mise en œuvre des modèles hybrides HMM/réseaux de neurones

Plusieurs approches distinctes ont été tentées pour la modélisation hybride HMM/réseaux de neurones :

- Réseaux de neurones en *pré-traitement* des HMM
  Deux approches existent pour l'emploi des réseaux de neurones en pré-traitement des HMM : l'estimation des probabilités d'émission des états ou la quantification vectorielle des trames de signal pour les HMM discrets.
  Dans le premier cas, les réseaux de neurones produisent des approximations des probabilités *a posteriori* pour chaque classe représentée par un état markovien $P(C|x)$ [Bourlard & Wellekens, 1989]. L'introduction de fonctions d'activation non-linéaires (classiquement une sigmoïde symétrique comme la tangente hyperbolique) dans le réseau ne permet pas d'assurer que la somme des sorties sera normalisée sur l'ensemble des classes. Dans la pratique, cette propriété est généralement imposée par l'utilisation de la fonction *softmax* sur la couche de sortie qui normalise explicitement les sorties du réseau [Morgan, Bourlard *et al.*, 1993]. Elles peuvent alors être interprétées comme une estimation de la probabilité $P(C_i|x)$. Ces probabilités sont ensuite divisées par la fréquence relative des classes et fournies aux HMM comme estimation de la probabilité d'émission des états. Le but poursuivi est d'intégrer un mécanisme d'alignement temporel dans les réseaux. Les deux modèles sont optimisés séparément.
  Les systèmes qui reposent sur cette approche sont : le système hybride HMM/MLP [Bourlard & Wellekens, 1990; Bourlard, Morgan *et al.*, 1992], le système hybride HMM/RN [Robinson, 1994], le système hybride HMM/TDNN [Devillers, 1992].

Dans le second cas, les réseaux de neurones sont utilisés pour réaliser une quantification vectorielle des trames acoustiques. Les sorties d'un tel réseau sont alors les index des prototypes retenus pour chacun des états des HMM. Cette approche n'a pas connu autant de succès que la précédente. Son attrait réside essentiellement dans l'utilisation de HMM discrets moins coûteux en temps de calcul.

Le système qui repose sur cette approche est le système hybride HMM/LVQ [Iwamida, Katagiri *et al.*, 1991; Rigoll, 1994].

- *Apprentissage conjoint* des deux modèles

  Cette approche est similaire à la précédente mais les deux modèles ont un apprentissage coopératif : rétro-propagation de l'erreur d'alignement temporel dans le réseau de neurone ou ré-apprentissage du réseau de neurone avec un nouvel alignement temporel (segmentation).

  Les systèmes qui reposent sur cette approche sont : les systèmes fondés sur l'apprentissage de Viterbi connexionniste [Franzini & Waibel, 1991], les réseaux Alphanets [Bridle, 1990], les systèmes fondés sur l'optimisation globale [Bengio, de Mori *et al.*, 1992].

- Réseaux de neurones en *post-traitement* des HMM

  Dans cette approche, les réseaux de neurones sont utilisés pour corriger ou affiner la classification obtenue à partir des HMM seuls. Ainsi, une fois une segmentation proposée par les HMM, les réseaux de neurones sont introduits pour réestimer les zones confuses (identifiées à l'aide d'une technique de N-best) ou les phonèmes les plus difficiles à identifier (déterminés sur l'ensemble d'apprentissage). Les systèmes qui reposent sur cette approche sont : les systèmes fondés sur la réestimation des zones confuses [Austin, Zavaliagkos *et al.*, 1992], les systèmes fondés sur la réestimation des phonèmes les moins bien reconnus [Niles, 1991].

- *Coopération* lors de la reconnaissance

  Deux techniques de coopération ont été étudiées. Tout d'abord la combinaison de deux systèmes HMM en parallèle : un système hybride et un système standard. Puis, la combinaison linéaire des probabilités d'émissions obtenues par un estimateur standard et un MLP.

  Les systèmes qui reposent sur cette approche sont : les systèmes parallèles HMM/TDNN et HMM/GMM [Dugast, Devillers *et al.*, 1994], les systèmes intégrant une combinaison des probabilités GMM/MLP [Renals, Morgan *et al.*, 1992].

Parmi, les quatre approches décrites ci-dessus, l'approche par pré-traitement est celle qui a connu le plus de succès et conduit à des systèmes compétitifs. On peut citer, parmi d'autres, deux systèmes fondés sur l'approche pré-traitement : le système hybride

HMM/RN ABBOT de reconnaissance de parole continue grand vocabulaire de l'anglais qui participe à l'évaluation DARPA [DARPA, 1997; DARPA, 1998] et offre maintenant une version temps réel [Cook, Robinson *et al.*, 1998]; et le système de la Faculté de Mons de reconnaissance du français [Deroo, 1998], basé sur le système hybride HMM/MLP NOWAY [Renals & Hochberg, 1995].

Le principal avantage de l'utilisation des réseaux de neurones pour la reconnaissance de la parole continue est l'introduction de la discrimination au niveau des trames acoustiques. Ses inconvénients sont une phase d'apprentissage extrêmement longue au moyen de procédures complexes dont la convergence n'est pas toujours assurée.

Les réseaux de neurones, comme les estimateurs continus et semi-continus, constituent des estimateurs performants permettant d'obtenir de bons résultats de reconnaissance. Toutefois, la nature paramétrique de ces estimateurs ne leur permet d'estimer correctement les probabilités réelles qu'au prix d'une augmentation toujours plus grande du nombre de leurs paramètres. A cet égard, l'utilisation d'estimateurs de probabilité non paramétriques pourrait se révéler profitable.

2.2.3. Estimateurs de probabilité discriminants non-paramétriques

Les estimateurs non paramétriques ne font aucune hypothèse sur la forme des distributions à estimer. Ils n'ont donc pas de paramètres à ajuster. Leur principe est une estimation locale à partir des données de référence.

L'estimation par histogramme constitue le paradigme des techniques d'estimation non-paramétrique. L'espace est divisé en « pavés » sur lesquels la distribution est estimée par le nombre d'échantillons contenus dans le pavé. Cette opération doit être répétée pour chaque classe. La complexité de cette technique est en variation exponentielle de la dimension de l'espace de représentation. Ainsi, une fois déterminée la précision Pr requise par dimension, il faudra considérer Pr^d pavés par classe. Dans le cadre de la reconnaissance de la parole ($10<d<40$), cette technique n'est pas utilisable dès lors que l'essentiel de ces pavés ne contient aucun échantillon ou un nombre trop faible pour une estimation fiable. Il faut donc avoir recours à des techniques plus évoluées.

Les détails sur la présentation des estimateurs non-paramétriques pourront être trouvés dans [Fukunaga, 1972], [Duda & Hart, 1973] et [Dubuisson, 1990]. Les chapitres suivants s'inspirent de ces ouvrages.

2.2.3.1. Principe des estimateurs non-paramétriques

Considérons n observations $x_1,x_2,\ldots,x_n$, indépendantes et identiquement distribuées, d'une variable aléatoire X, la probabilité qu'un point x_0 quelconque appartienne à une région R de l'espace peut être estimée par :

$$\hat{P}_n(x_0 \in R) = \frac{k}{n} \tag{2-22}$$

où k représente le nombre d'observations de X présentes dans la région R. Le terme $\hat{P}_n(x)$ est une variable aléatoire binomiale de paramètres P et n :

$$\mathrm{Prob}\left(\hat{P}_n(x_0 \in R) = k/n\right) = C_n^k P(x)^k (1 - P(x))^{n-k} \tag{2-23}$$

où $P(x)$ est la distribution réelle de X. $\hat{P}_n(x)$ est l'estimateur de maximum de vraisemblance de $P(x)$. L'espérance et la variance de $\hat{P}_n(x)$ sont données par :

$$E\left[\hat{P}_n(x)\right] = \sum_{k=0}^{n} (k/n) \times C_n^k P(x)^k (1 - P(x))^{n-k} = P(x) \tag{2-24}$$

$$\begin{aligned} \mathrm{var}\left[\hat{P}_n(x)\right] &= \sum_{k=0}^{n} \left((k/n) - P(x)\right)^2 \times C_k^n P(x)^k (1 - P(x))^{n-k} \\ &= \frac{P(x)(1 - P(x))}{n} \end{aligned} \tag{2-25}$$

On en déduit que $\hat{P}_n(x)$ est un estimateur sans biais de $P(x)$. De là, une estimation de la densité de probabilité $\hat{p}_n$ en x_0 peut être obtenue par :

$$\hat{p}_n(x_0) = \frac{\hat{P}_n(x_0)}{Vol_n} = \frac{k_n}{n \times Vol_n} \tag{2-26}$$

où Vol_n est une mesure du volume de la région R_n, fonction de n (nombre de points disponibles) dans laquelle sont comprises les k_n observations les plus proches de x_0.

La convergence de $\hat{p}_n(x_0)$ vers $p(x_0)$ est assurée dès lors que les trois conditions suivantes sont remplies :

[1] $\lim_{n\to\infty} V_n = 0$

[2] $\lim_{n\to\infty} k_n = 0$

[3] $\lim_{n\to\infty} k_n / V_n = 0$

La première condition assure la convergence vers $p(x_0)$ en imposant la continuité de $\hat{p}_n(x_0)$ autour de x_0. La seconde condition assure la convergence en probabilité de k_n/n vers P. Enfin, la troisième condition assure que le rapport entre les échantillons compris dans la région R_n et le nombre total d'échantillons reste faible.

L'estimateur ainsi défini, il reste à préciser les valeurs de k_n et Vol_n. Ces deux variables ne sont pas indépendantes, il suffit donc d'en fixer une. Deux méthodes sont alors envisageables :

- fixer *a priori* les régions R_n en vérifiant que leur volume Vol_n est conforme aux conditions [1] et [3]. Par exemple :

$$Vol_n = \frac{1}{\sqrt{n}}$$

  Cette méthode constitue l'estimation par noyaux [Parzen, 1962].

- fixer *a priori* les k_n fonction de n et faire croître le volume Vol_n de R_n jusqu'à ce qu'il contienne k_n échantillons. Cette méthode constitue l'estimation par les k plus proches voisins [Fix & Hodges, 1951].

  La dépendance des estimateurs non-paramétriques au nombre n d'échantillons de l'ensemble de référence étant établie, nous ne la préciserons plus dorénavant.

2.2.3.2. Estimation par noyau

L'estimation par noyau nécessite de définir *a priori* le volume Vol des régions R. Une fois ce volume défini, l'estimation de $p(x_0)$ revient à compter le nombre d'échantillons présents dans un volume Vol autour de x_0 :

$$\hat{p}(x_0) = \frac{k}{n \times Vol} \tag{2-27}$$

où k représente le nombre d'échantillons appartenant à R. Le dénombrement des échantillons est réalisé au travers d'une fonction indicatrice, aussi appelée noyau pour certaines de ses propriétés. La fonction la plus simple, dite noyau de Parzen, est une fonction par morceaux qui vaut 1 dans un hypercube de côte h et 0 partout ailleurs :

$$K\left\{\frac{x_0 - x_i}{h}\right\} = \begin{cases} 1 \ si \ x_i \in R \\ 0 \text{ sinon} \end{cases} \tag{2-28}$$

Ainsi, k est obtenu par :

$$k = \sum_{i=1}^{n} K\left\{\frac{x_0 - x_i}{h}\right\} \tag{2-29}$$

et l'estimateur prend la forme :

$$\hat{p}(x_0) = \frac{1}{n \times h^d} \sum_{i=1}^{n} K\left\{\frac{x_0 - x_i}{h}\right\} \tag{2-30}$$

La fonction K n'est pas limitée au noyau de Parzen et peut prendre une infinité de forme. Les conditions à respecter sont celles qui définissent un noyau et permettent que $\hat{p}(x_0)$ reste homogène à une densité de probabilité :

[4] $K(x) \geq 0$ pour tout x

[5] $\int K(x)dx = 1$

Un certain nombre de noyaux a été proposé : le noyau gaussien (ou d'Epanichnikov), le noyau sinus cardinal… Ils ont en commun d'être maximum en 0 et de décroître plus ou moins vite pour atteindre 0 en *h*. De cette manière, une pondération est introduite dans le dénombrement des plus proches voisins en fonction de leur distance au point étudié. Dans le cas d'espaces multidimensionnels, les noyaux couramment utilisés sont définis comme produit des noyaux monodimensionnel, ce qui a pour effet de conserver les propriétés [4] et [5]:

$$K(x_1,\ldots,x_d) = \prod_{i=1}^{d} K(x_i) \qquad \text{(2-31)}$$

Les deux points clés de l'estimation par noyaux sont donc : le choix du noyau et la valeur de *h* donnant le volume *Vol*.

Des expériences de reconnaissance à l'aide de noyaux de Parzen ont été réalisées [Soudoplatoff, 1986]. L'analyse de ces travaux est rendue complexe du fait de l'introduction d'un mécanisme supplémentaire aux HMM : les générateurs de vecteurs acoustiques. Plutôt que d'associer directement les points aux états des HMM, Soudoplatoff propose de les associer à des générateurs de vecteurs acoustiques qui sont liés aux états des HMM par une probabilité. On ne peut donc vraiment évaluer l'apport seul de l'estimateur par noyau de Parzen. Cette étude préliminaire n'a pas donné lieu à des développements ultérieurs qui auraient pu confirmer et expliciter l'intérêt de l'estimateur par noyau de Parzen.

L'estimation par noyau de Parzen est coûteuse du fait de l'évaluation du noyau sur chaque échantillon de l'ensemble de référence. Dans la section suivante, une variante de l'estimateur par noyau moins coûteuse et plus simple à mettre en œuvre est présentée.

2.2.3.3. Estimateur des K-ppv

L'estimation par K-ppv est basée sur la deuxième possibilité d'estimation de $p(x_0)$: on fixe *a priori* k en fonction de n et on fait croître le volume *Vol* de *R* jusqu'à ce qu'il contienne k échantillons. Ainsi, c'est le volume *Vol* qui s'ajuste à la valeur de la densité en x_0 dans l'équation (2-27).

L'estimation ainsi obtenue peut-être montrée asymptotiquement sans biais sous les conditions [1] et [3] [Loftsgaarden & Quesenbury, 1965]. On notera aussi, qu'ainsi calculée, elle ne peut-être jamais être nulle, ce qui n'est pas vrai dans le cas de l'estimation par noyau.

Dans le cadre de la classification, les valeurs des densités de probabilité fournie par l'estimateur des K-ppv pour chaque classes sont comparées. Ainsi la précision de l'estimation doit-elle être mesurée comme sa capacité à associer correctement un point à sa classe dans le cadre de la règle du maximum de probabilité *a posteriori*. Au préalable à son évaluation expérimentale, l'erreur d'estimation des K-ppv peut être majorée en fonction de K ou du nombre de classes considérées.

2.2.3.4. Majoration de l'erreur de l'estimateur des K-ppv

Les principaux éléments de l'estimation de l'erreur des K-ppv ont été énoncés dans [Cover & Hart, 1967]. A titre d'exemple, nous reprenons ici la démonstration de la majoration de l'erreur de classification du 1-ppv dans le cas de C classes.

Il y a erreur de classification si une observation x appartient à c_i et qu'elle est affectée à c_j. Dans le cas du 1-ppv, cela signifie que le ppv x_1 de x est un point associé à la classe c_j. L'erreur de classification conditionnelle, considérant x et x_1, peut s'exprimer :

$$err(x,x_1) = P(c_1/x_1)\sum_{\substack{j=1\\ j\neq 1}}^{C} P(c_j/x) + \ldots + P(c_C/x_1)\sum_{\substack{j=1\\ j\neq C}}^{C} P(c_j/x) \qquad (2\text{-}32)$$

Lorsque l'ensemble de référence est suffisamment large, x et x_1 sont très proches et leurs probabilités *a posteriori* peuvent être considérées égales. Alors, l'équation (2-32) peut se reformuler :

$$err(x) = P(c_1/x)\sum_{\substack{j=1\\ j\neq 1}}^{C} P(c_j/x) + \ldots + P(c_C/x)\sum_{\substack{j=1\\ j\neq C}}^{C} P(c_j/x) \qquad (2\text{-}33)$$

Le premier terme de cette équation correspond à l'erreur de classification lorsque x appartient à c_1 et ainsi de suite. En utilisant $P(c_1)+..+P(c_C)=1$, on peut reformuler l'équation (2-33) :

$$err(x) = 1 - \sum_{j=1}^{C} P^2(c_j/x) \qquad (2\text{-}34)$$

Par ailleurs, l'erreur de classification de Bayes, considérant x, est donnée par :

$$err^*(x) = 1 - \max_j \{P(c_j/x)\} = 1 - P(c_{\max}/x) \qquad (2\text{-}35)$$

L'inégalité de Schwartz permet de majorer l'erreur (2-34) :

$$(C-1)\sum_{\substack{j=1\\ j\neq \max}}^{C} P^2(c_j/x) \geq \left\{\sum_{\substack{j=1\\ j\neq \max}}^{C} P(c_j/x)\right\}^2 \qquad (2\text{-}36)$$

Or, le membre droit de cette inégalité est égal au carré de l'erreur minimale bayesienne :

$$\left\{\sum_{j=1;\, j\neq \max}^{C} P(c_j/x)\right\}^2 = \{1 - P(c_{\max}/x)\}^2 = err^*(x)^2 \qquad (2\text{-}37)$$

On peut alors ajouter $(C-1)P^2(c_{max}/x)$ à chaque membre de l'inégalité (2-36) :

$$(C-1)\sum_{j=1}^{C} P^2(c_j/x) \geq err^*(x)^2 + (C-1)\{1-err^*(x)\}^2 \tag{2-38}$$

Finalement, le membre gauche de cette inégalité peut s'exprimer à l'aide de l'expression (2-34). Ainsi, l'erreur de classification est majorée par une fonction de l'erreur bayesienne :

$$err(x) \leq 2err^*(x) - \frac{C}{C-1}err^*(x)^2 \tag{2-39}$$

L'inégalité de Jensen nous permet de généraliser cette majoration à l'erreur moyenne de classification ε :

$$\varepsilon = E[err] \leq E\left[2err^* - \frac{C}{C-1}err^{*2}\right] = 2\varepsilon^* - \frac{C}{C-1}\varepsilon^{*2} \tag{2-40}$$

soit

$$\varepsilon \leq 2\varepsilon^* - \frac{C}{C-1}\varepsilon^{*2} \tag{2-41}$$

Dans le cadre de la RAP, *C* est généralement de l'ordre de quelques dizaines, donc :

$$\varepsilon \leq 2\varepsilon^* - \varepsilon^{*2} \tag{2-42}$$

La probabilité d'erreur pour l'estimateur des K-ppv est comprise entre une fois et deux fois l'erreur optimale de Bayes. D'autre part, il est possible de montrer que la probabilité d'erreur diminue lorsque K augmente (voir par exemple [Fukunaga, 1972]). [Devijver & Kittler, 1982] propose une majoration de l'erreur de l'estimateur des K-ppv, fonction de K, dont la démonstration n'est pas reprise ici :

$$\varepsilon_K \leq \left(1 + \frac{2}{\sqrt{\pi K/2}}\right)\varepsilon^* \tag{2-43}$$

soit, dans le cas de l'estimateur des 50-ppv :

$$\varepsilon_K \leq 1{,}2\ \varepsilon^* \tag{2-44}$$

Ainsi, l'erreur des 50-ppv est comprise entre 100% et 120% de l'erreur bayesienne. La règle de décision déduite de l'estimateur des K-ppv a donc une erreur proche de l'optimum de Bayes qui peut être réduite par l'augmentation de K. Cette propriété constitue un des avantages majeurs de la méthode par rapport aux autres estimateurs : il existe un facteur dont on est sûr qu'il fera diminuer l'erreur de classification. Dans le cas des GMM, l'augmentation du nombre de gaussiennes par somme pondérée permet en général d'atteindre ce but mais cette propriété repose en grande partie sur la qualité des techniques d'apprentissage mises en œuvre. Malgré de bonnes propriétés théoriques, certains éléments ont limité l'application de l'estimateur des K-ppv dans le cadre de la RAP.

2.2.3.5. Limitations à l'emploi de l'estimateur des K-ppv

La principale limitation à l'emploi de l'estimateur des K-ppv dans des tâches de RAP représentatives est la nécessité d'un grand nombre d'échantillons de référence qui implique un coût de calcul très important.

Deux points permettent de réduire le coût de calcul des K-ppv d'un vecteur : la diminution du nombre de distances à calculer et un accès rapide aux vecteurs de référence. Le premier point sera résolu par le développement d'algorithmes rapides (voir §2.3.1). Quant au second, les tailles des mémoires à accès rapide actuelles permettent qu'il ne soit plus un réel obstacle.

Dans le cas de la base TIMIT, l'ensemble de référence comprend 1.124.823 vecteurs. Un vecteur de paramètres (12 MFCC+E) nécessite en représentation en virgule flottante 13x4=52 octets. L'utilisation de l'estimateur des K-ppv nécessitera donc environ 52 Méga-octets (52x1.124.823 octets) pour représenter en mémoire centrale l'ensemble des vecteurs de référence. Le standard de la taille de la mémoire centrale des machines grand public devenant 128 Méga-octets, le problème de l'occupation mémoire n'est plus crucial.

2.3. Discrimination locale par la règle des K-ppv

Avant d'introduire l'estimateur des K-ppv dans le formalisme markovien, ses performances en identification locale sont évaluées. Dans ce cadre, la prise de décision est immédiate et ne requiert pas de traitement global sur l'ensemble des trames. Ainsi, l'estimateur des K-ppv sera transformé en opérateur de reconnaissance à travers la règle de décision des K-ppv. Auparavant, la mise en œuvre de l'estimateur des K-ppv sur des bases de données de grandes dimensions a nécessité le développement d'un algorithme de calcul rapide.

2.3.1. Algorithme rapide de calcul des K-ppv

De nombreuses approches ont été proposées afin de lever la difficulté soulevée par le temps de calcul important de l'estimateur des K-ppv [Miclet & Dabouz, 1983; Kim & Park, 1986; Neimann & Goppert, 1988; Broder, 1990; Vidal, 1994]. Dans notre cas, le problème a été résolue par le développement d'un algorithme très rapide [Montacié, Caraty *et al.*, 1997]. Le programme développé par C. Montacié et M.-J. Caraty utilise trois méthodes pour obtenir un algorithme de calcul des K-ppv très rapide : les méthodes de Kittler-Richetin, de Friedman et de Fukunaga. Les détails concernant l'algorithme pourront être trouvé dans [Montacié, 1999]. Nous introduisons le principe des méthodes sur lesquelles il est fondé ainsi que le gain en temps de calcul (par rapport au calcul de base) constaté lors de nos expérimentations sur la base TIMIT :

- la méthode de Kittler-Richetin [Kittler, 1978; Richetin, Rives *et al.*, 1980] estime la distance euclidienne (d_E) en utilisant des *distances moins coûteuses* comme la distance de Manhattan (d_M) ou la distance *Max* (d_∞). Cette méthode repose sur les inégalités (2-45) et (2-46) entre les distances d_M, d_E et d_∞ :

$$d_E(x_1,x_2)\times\sqrt{d} \geq d_M(x_1,x_2) \geq d_E(x_1,x_2) \qquad (2\text{-}45)$$

$$d_E(x_1,x_2) \geq d_\infty(x_1,x_2) \geq \frac{d_E(x_1,x_2)}{\sqrt{d}} \qquad (2\text{-}46)$$

  où x_1 et x_2 sont des vecteurs de dimension d. Le gain observé en temps de calcul est de 80%.

- la méthode de Friedman [Friedman, Bentley *et al.*, 1975] classe les vecteurs d'analyse suivant la *composante de l'espace de représentation de plus forte dispersion*. Lorsque, pour un vecteur, la distance euclidienne projetée sur cette composante devient supérieure à la distance euclidienne totale du point étudié à son k$^{\text{ème}}$ plus proche voisin parmi les vecteurs déjà examinés, la recherche est interrompue. Le gain en temps de calcul est de 75%.

- la méthode de Fukunaga [Fukunaga & Narendra, 1975] découpe l'espace des vecteurs d'analyse de l'ensemble d'apprentissage en *clusters hiérarchiques*. Le calcul des K-ppv d'un vecteur x_1 est réalisé par une exploration en profondeur d'abord de la structure hiérarchique. Chaque cluster *Sp* est étudié, son centroïde *Cp* et son rayon *Rp* sont calculés. Deux inégalités (2-47) et (2-48) sont utilisées pour éviter des calculs inutiles. Si l'inégalité (2-47) est vraie, aucun vecteur de *Sp* n'est un des *K* plus proches voisins de x_1. Si l'inégalité (2-48) est vraie le vecteur x_2 n'appartient pas à l'ensemble des *K* plus proches voisins du vecteur x_1.

$$d_E(x_1,C_p) \geq d_k + R_p \qquad (2\text{-}47)$$

$$d_E(x_1,C_p) \geq d_k + d_E(x_2,C_p) \qquad (2\text{-}48)$$

  où d_k est la distance entre le vecteur x_1 et son k$^{\text{ème}}$ plus proche voisin parmi les vecteurs déjà traités.

L'algorithme rapide de K-ppv, développé à partir de ces trois méthodes, a été testé sur la base de parole TIMIT (Annexe E). Nous rappelons que l'ensemble de référence est composé de 1.124.823 trames (3696 phrases) et le noyau de test de 57.919 trames (192 phrases). Une décomposition en 30.000 clusters hiérarchiques a été utilisée. Pour chaque vecteur de l'ensemble d'apprentissage et de test, l'estimateur 50-PPV est obtenu à partir de ses 50 plus proches voisins dans l'ensemble de référence. Le gain total en temps de calcul constaté est de 99,8%. Cette économie permet de calculer 16 50-PPV par seconde sur une machine de type PENTIUM PC à 600 MHz.

2.3.2. Règle des K-ppv

Pour réaliser l'identification locale de trames de parole, une règle de décision doit être déduite de l'estimateur des K-ppv. Considérons, tout d'abord, un problème de discrimination à 2 classes. La règle de décision offrant les meilleures garanties de classer correctement un vecteur x_0 entre 2 classes c_1 et c_2 est la règle du maximum *a posteriori* :

$$D(x_0) = 1 \text{ si } P(c_1 / x_0) \geq P(c_2 / x_0) \qquad (2\text{-}49)$$

que la formule de Bayes permet de transformer en :

$$D(x_0) = 1 \text{ si } P(x_0 / c_1)P(c_1) \geq P(x_0 / c_2)P(c_2) \qquad (2\text{-}50)$$

Dans l'expression de la probabilité *a posteriori* en fonction des vraisemblances, la probabilité *a priori* $P(x_0)$ est omise dans la mesure où ne dépendant pas de la classe elle n'intervient pas dans la classification. La maximisation de la probabilité *a posteriori* est alors équivalente à la maximisation de la probabilité conjointe des donnée et de leur classe.

Si l'estimation par K-ppv est utilisée pour déterminer la règle de décision D, on obtient :

$$D(x_0) = 1 \text{ si } \frac{k_1}{n_1} \times \frac{n_1}{n} \geq \frac{k_2}{n_2} \times \frac{n_2}{n} \qquad (2\text{-}51)$$

où n_1,k_1 et n_2,k_2 représentent le nombre d'échantillons pour chacune des classes parmi l'ensemble de référence et parmi les K plus proches voisins (n_1+n_2=n et k_1+k_2=K). Cette expression est alors équivalente à :

$$D(x_0) = 1 \text{ si } k_1 \geq k_2 \qquad (2\text{-}52)$$

On aboutit donc à une règle de décision extrêmement simple, illustrée Figure 2-4 : x_0 est affecté à la classe possédant le plus grand nombre de plus proches voisins parmi les K. La règle de décision des K-ppv peut être facilement généralisée au cas de C classes. Considérons k_i ppv parmi K pour chaque classe i, la règle de décision devient :

$$D(x_0) = \arg\max_i (k_1, \ldots, k_i, \ldots, k_C) \qquad (2\text{-}53)$$

L'opérateur de reconnaissance obtenu à partir de cette règle de décision va être évalué sur une tâche d'identification locale.

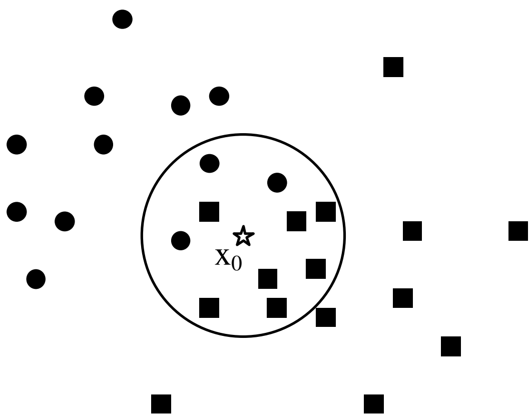


Figure 2-4 Exemple d'application de la règle des 10-ppv sur les classes ■ et ● x_0 est associé à la classe ■, majoritaire parmi ses 10-ppv.

2.3.3. Comparaison K-ppv/GMM en identification locale

Le principe de l'identification locale (ou trame-à-trame) est d'assigner chaque trame de l'ensemble d'évaluation à la classe phonétique la plus probable sans considérations segmentales (*i.e.* indépendamment des trames adjacentes). La règle de décision déduite de l'estimateur des K-ppv sera utilisée : la classe reconnue est la classe associée au plus grand nombre de ppv de la trame considérée.

La règle des K-ppv a déjà été appliquée à la RAP isolée [White & Fong, 1975] mais jamais à une telle échelle. Les expériences réalisées jusqu'à présent n'offraient pas de résultats concluants dans la mesure où la dualité nombre de données/calculabilité limitait son utilisation à des tâches réduites. Le développement de l'algorithme des K-ppv rapides a permis de résoudre ce conflit. L'estimateur des K-ppv peut donc être évalué sur une base représentative (TIMIT).

La règle des K-ppv va être comparée à un opérateur de reconnaissance fondé sur l'estimateur de l'état de l'art gaussien. Un GMM est initialisé pour chaque classe par l'algorithme des k-moyennes [Linde, Buzo *et al.*, 1980] puis ses paramètres sont affinés à l'aide de l'algorithme de Baum-Welch. L'algorithme des k-moyennes permet de diviser l'espace des données de façon itérative en un nombre donné de clusters. Pour l'apprentissage des GMM, ce nombre est fixé au préalable et correspond au nombre de gaussiennes dans le GMM. Il peut être aussi obtenu de manière non-supervisée à l'aide d'un critère de distorsion. Dans ce cas, la division de l'espace est stoppée lorsque plus aucun cluster n'a de distance moyenne à l'ensemble de ses éléments supérieure à un certain seuil. Chaque cluster est associé à une gaussienne dont le vecteur moyenne et la matrice de covariance sont évalués sur les données appartenant au cluster. A chaque itération, un nouveau cluster est créé en divisant le cluster de plus forte dispersion. Ce dernier est coupé en deux : une légère perturbation sépare son centre en deux nouveaux centres. Dans la version utilisée pour nos expériences, le coefficient de perturbation tient compte de la variance totale du cluster pour amplifier la perturbation selon les dimensions de plus fortes dispersions.

Les expériences sont menées sur les 1.124.823 trames de l'ensemble de référence (auto-cohérence) et sur les 57919 trames du noyau[1] de test de TIMIT (Annexe E). Le taux d'identification est calculé par une comparaison trame-à-trame entre l'alignement de référence [Seneff & Zue, 1988] et l'étiquetage obtenu par identification. Les scores étant compris entre 40 et 60%, les rayons des intervalles de confiance pour ces expériences sont d'environ 0,1% pour l'ensemble d'apprentissage et 0,4% pour le test (cf. Annexe B).

Il est habituel depuis [Lee, 1988] pour les expériences menées sur TIMIT d'autoriser certaines confusions entre phonèmes de sorte que les 48 classes utilisées lors du décodage sont ramenées à 39 lors du calcul des scores (cf. Annexe E). Des expériences préliminaires ont montré que si ce mode d'évaluation est favorable à l'estimateur gaussien, dans le cas de l'estimateur des K-ppv l'utilisation des 39 classes directement lors de l'identification conduit à de meilleurs résultats. Pour chaque estimateur, le mode le plus favorable a donc été utilisé lors de l'évaluation.

2.3.3.1. Résultats de l'estimateur des K-ppv

La Figure 2-5 présente les résultats, détaillés dans le Tableau 2-2, de l'identification locale pour l'estimateur des K-ppv pour K variant de 1 à 50.

[1] Dans la mesure où le noyau de test de TIMIT constituera le seul ensemble de test utilisé tout au long de notre étude, il sera dorénavant appelé indistinctement « ensemble de test ».

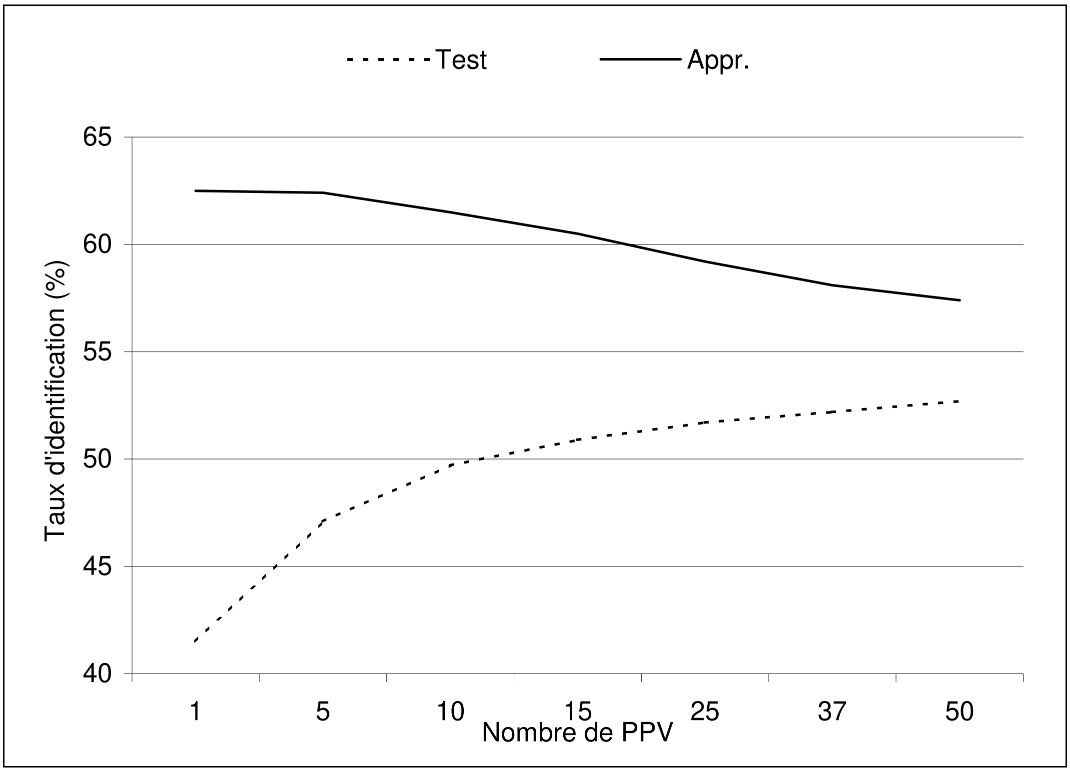


Figure 2-5 Evolution du taux d'identification locale de la règle des K-ppv fonction de K.

Nb de PPV	Taux d'identification	
	Appr.	Test
1	62,5	41,5
5	62,4	47,1
10	61,5	49,7
15	60,5	50,9
25	59,2	51,7
37	58,1	52,2
50	57,4	52,7

Tableau 2-2 Taux identification locale de la règle des K-ppv fonction de K.

Conformément aux calculs théoriques (cf. §2.2.3.4), le nombre d'erreurs de la règle des K-ppv diminue lorsque K croit. Un écart sur les taux de reconnaissance du test de 11,2% est observé entre les estimateurs 1-PPV et 50-PPV. Cette amélioration n'est toutefois vraie que pour l'ensemble de test. L'ensemble d'apprentissage voit son taux d'identification chuter d'environ 5% entre les estimateurs 1-PPV et 50-PPV. Ce phénomène peut s'expliquer par la présence, parmi les ppv des points de l'ensemble d'apprentissage, de points appartenant à la même séquence acoustique ou au moins à une séquence acoustique du même locuteur. Avec l'augmentation de K, ce phénomène s'estompe par la prise en compte de points issus de séquences acoustiques d'origines plus diverses. C'est ce que semble corroborer l'asymptote commune des ensembles de test et d'apprentissage. On peut en conclure que l'augmentation du nombre de ppv améliore les capacités de généralisation de l'estimateur des K-ppv.

2.3.3.2. Résultats de l'estimateur GMM

La Figure 2-6 présente les résultats, détaillés dans le Tableau 2-3, d'identification locale pour les GMM en fonction du nombre de gaussiennes par somme pondérée (de 1 à 100).

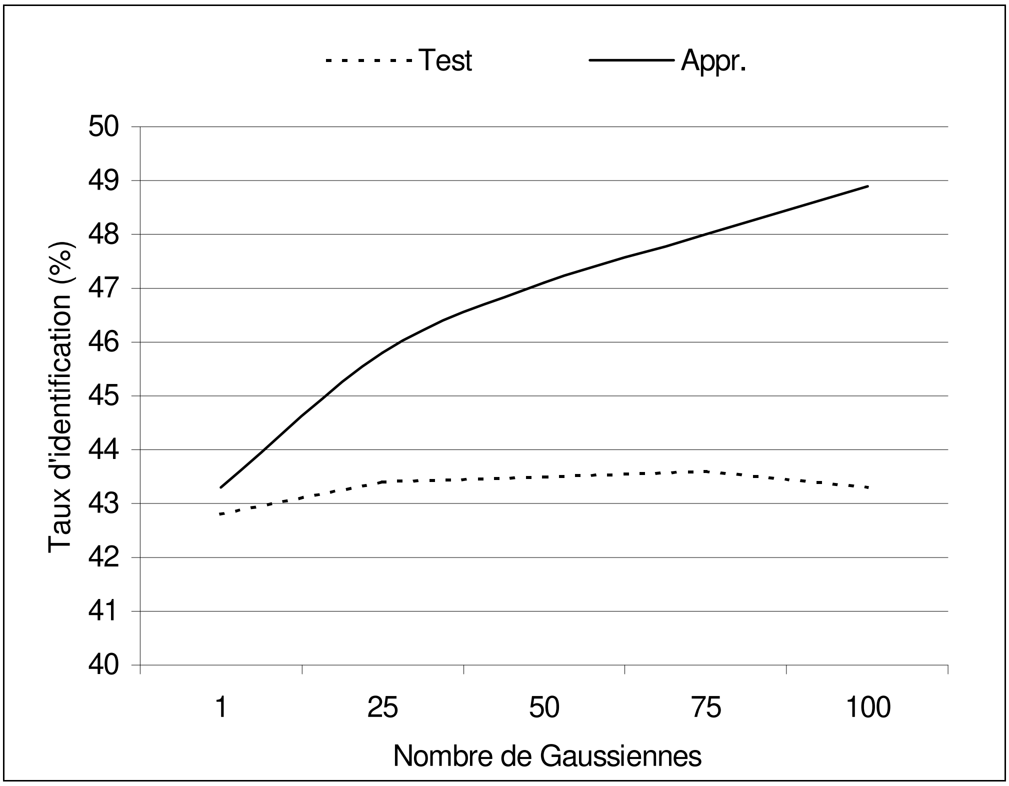


Figure 2-6 Evolution du taux d'identification locale des GMM en fonction du nombre de gaussiennes par mixture.

Nb de gaussiennes	Taux d'identification	
	Appr.	Test
1	43,3	42,8
25	45,8	43,4
50	47,1	43,5
75	48,0	43,6
100	48,9	43,3

Tableau 2-3 Taux d'identification locale des GMM en fonction du nombre de gaussiennes par mixture.

Le taux d'identification en auto-cohérence augmente avec le nombre de gaussiennes par somme pondérée (+5,2 % entre 1 et 100 gaussiennes). Pour le test, le meilleur score est obtenu pour 75 gaussiennes avec 43,6% d'identifications correctes. Il semble ensuite s'infléchir. Ces résultats tendent à montrer que l'estimateur GMM se spécialise (sur-apprentissage) avec l'augmentation du nombre de gaussiennes.

2.3.3.3. Mise en relation des résultats

Les taux d'identification locale des estimateurs K-ppv et GMM sont mis en rapport. Pour chaque ensemble d'évaluation, les deux meilleures configurations ont été considérées (Figure 2-7).

L'estimateur des 50-ppv obtient des résultats supérieurs à l'estimateur GMM de l'ordre de 9% sur le test et jusqu'à 14% en auto-cohérence. Les meilleurs taux d'identification en auto-cohérence sont obtenus avec les estimateurs 1-ppv et 100-GMM. Sur le test, les estimateurs 50-ppv et 75-GMM ont fourni les résultats les plus élevés.

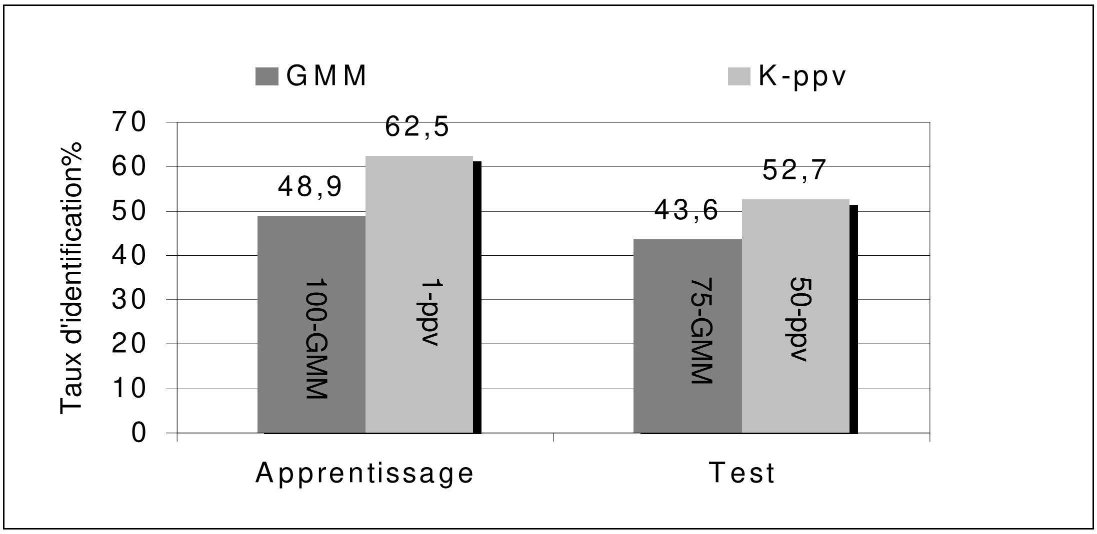


Figure 2-7 Meilleurs taux d'identification locale des estimateurs K-ppv et GMM sur l'apprentissage et le test de TIMIT.

Afin de mesurer plus finement les différences entre les estimateurs K-ppv et GMM, les résultats d'identification locale des 50-ppv et 75-GMM sont décomposés d'abord sur l'ensemble des classes phonétiques (Tableau 2-4) puis sur les macro-classes phonétiques (Tableau 2-5).

Phn	Nb Trames	Taux (%) d'identification	
		50-ppv	75-GMM
Voyelles			
iy	2127	62,6	32,0
ih	3461	56,2	20,4
eh	1696	18,5	17,2
ae	1458	37,0	31,1
aa	2898	57,0	36,4
ah	2016	29,0	19,5
uw	679	14,7	23,4
uh	220	0,0	14,1
er	2298	59,0	38,9
ey	1554	27,3	27,8
ay	1362	13,8	14,8
oy	294	0,0	5,8
aw	524	1,0	12,8
ow	1163	17,1	16,9
Liquides			
l	1941	53,1	43,2
r	1552	21,2	31,6
y	294	1,4	35,7
w	880	35,6	44,1
Silences			
sil	14035	91,0	76,3

Phn	Nb Trames	Taux (%) d'identification	
		50-ppv	75-GMM
Nasales			
m	1274	36,9	42,4
n	2016	51,8	40,9
ng	312	6,7	29,8
Fricatives			
z	1530	23,1	40,5
v	553	22,1	33,6
f	1371	64,0	53,4
th	294	0,3	32,0
s	3661	84,7	55,4
sh	918	67,5	54,8
hh	519	12,3	41,6
ch	359	1,9	26,5
jh	274	5,1	25,5
dh	467	5,1	22,7
Occlusives			
b	228	2,6	20,2
d	292	1,4	17,1
dx	247	1,2	20,6
g	192	1,0	30,2
p	602	16,8	34,6
t	835	21,6	29,6
k	1523	26,5	23,2

Tableau 2-4 Taux d'identification locale pour les estimateurs 50-ppv et 75-GMM (meilleur en grisé) sur le test de TIMIT détaillés par classes phonétiques.

Macro-classe	Taux d'identification	
	50-ppv	75-GMM
Voyelles	39,7	25,6
Liquides	35,9	39,0
Silences	91,0	76,3
Nasales	42,6	40,4
Fricatives	52,1	46,8
Occlusives	17,8	25,8

Tableau 2-5 Taux d'identification locale pour les estimateurs 50-ppv et 75-GMM (meilleur en grisé) sur le test de TIMIT détaillés par macro-classes phonétiques.

La différence de performance entre les estimateurs 50-ppv sur 75-GMM se repartit à peu près sur l'ensemble des macro-classes phonétiques avec un écart plus marqué pour les silences (+14,7%) et les voyelles (+14,1%) (Tableau 2-5). Seules les liquides (+3,1%) et les occlusives (+8%) sont mieux reconnues par l'estimateur 75-GMM. Par contre, lorsque que l'on détaille les résultats par classes phonétiques, les différences de comportements sont plus nuancées : 24 classes sur 39 sont mieux reconnues par l'estimateur 75-GMM. Toutefois, ces classes ont pour caractéristiques d'être les moins représentées dans l'ensemble de test. Or les proportions de l'ensemble de test respectent celles de l'ensemble de référence. Il semble donc que l'estimateur des 50-ppv soit sensible au nombre de représentant par classe. D'ailleurs les classes les plus faiblement représentées (quelques centaines de trames) sont presque intégralement non reconnues par l'estimateur des 50-ppv : /oy/, /aw/, /y/, /ch/, /d/, /dx/ et /g/. Cet effet est compensé globalement par une bonne reconnaissance des classes nombreuses : /iy/, /ih/, /aa/, /er/, /l/, /n/, /f/, /s/ et surtout /sil/.

L'absence de modélisation dans l'estimation par K-ppv permet d'expliquer ce résultat. Dans le cas de l'estimation par GMM, un modèle est associé à chaque classe. Lors de l'identification, seul ce modèle est pris en compte et le nombre de représentants qui a permis de l'estimer n'a plus d'influence directe. La modélisation permet dans une certaine mesure de compenser les défauts de l'ensemble d'apprentissage.

Par ailleurs, l'absence de modélisation rend l'estimation par K-ppv plus sensible à l'étiquetage de référence. Des améliorations de performance peuvent donc être espérées d'adaptations *ad hoc* des références à l'estimateur des K-ppv.

2.3.4. Adaptations des références

L'étiquetage de la base TIMIT présente de fortes garanties de fiabilité. Toutefois, de l'aveu même de ces concepteurs, certaines erreurs ou imprécisions peuvent subsistées. Il convient donc d'adapter l'alignement de référence à l'estimateur des K-ppv. Pour cela, nous proposons une méthode permettant la correction des frontières entre segments phonétiques basée sur les résultats d'identification locale par l'estimateur des K-ppv. Une technique permettant d'intégrer directement lors du décodage notre connaissance de l'imprécision de l'étiquetage de référence est introduite. Au préalable, l'analyse des résultats d'identification locale avec l'estimateur des K-ppv nous conduit à réaliser une séparation parole/silence sur la base de données.

2.3.4.1. Séparation parole/silence

La prise en compte d'une classe composée de segments de silence est communément acceptée dans le cadre de la reconnaissance de la base TIMIT [Lee, 1988; Gauvain, Lamel *et al.*, 1994; Robinson, 1994]. Dans le cas de l'utilisation de l'estimateur des K-ppv, cette situation doit être reconsidérée dès lors qu'elle conduit à avoir une classe représentant à elle seule 10% de l'ensemble de référence (157.500 trames sur 1.124.823).

La présence d'une classe prépondérante n'est pas souhaitable dans le cadre d'une estimation non-paramétrique. Tout particulièrement pour l'estimateur des K-ppv, il est conseillé que les probabilités *a priori* des classes dans l'ensemble de référence soient conformes à leur probabilité *a priori* dans la réalité [Davies, 1988]. Or, on peut s'interroger sur le sens de la probabilité *a priori* du silence. En effet, le silence n'étant pas un événement linguistique à proprement parler, il ne peut être considéré comme faisant partie du flux d'information acoustique si l'on excepte les courtes pauses entre mots et les tenues des occlusives. De plus, un nombre de représentants trop élevé conduit à une forte reconnaissance de la classe aux dépens des autres. Ainsi, lors de l'identification locale du test de TIMIT, on note que les taux de substitutions interclasses de la macro-classe « Silences » sont beaucoup plus élevé dans le cas de l'estimateur 50-ppv que dans celui du 75-GMM (Tableau 2-6).

Macro-classe	Fricatives	Liquides	Nasales	Occlusives	Silences	Voyelles
Fricatives	33,9 24,9	0,9 0,9	1,5 1,1	9,9 3,5	3,8 11,6	3,2 5,9
Liquides	3,9 1,7	7,1 6,6	3,3 4,1	5,5 0,4	0,9 2,0	40,2 49,2
Nasales	5,3 1,8	4,0 3,7	28,7 20,3	5,5 0,3	3,7 13,9	12,3 17,2
Occlusives	25,8 20,9	5,3 4,2	2,8 2,5	20,1 8,0	5,0 19,0	15,2 27,6
Silences	10,7 3,2	1 ,1 0,7	3,6 1,4	6,5 1,5	0,0 0,0	1,9 2,2
Voyelles	5,6 1,1	13,3 7,2	2,4 1,5	7,2 0,5	0,3 1,0	45,6 49,0

Tableau 2-6 Taux de substitutions inter macro-classes pour les estimateurs 75-GMM et 50-ppv (encerclé).

Afin de pallier cet inconvénient, nous proposons d'extraire de l'ensemble de référence les segments de silence. Cette approche présente en outre l'intérêt d'être en accord avec les préoccupations actuelles concernant la transcription de documents multimédia. En effet, pour cette tâche, et contrairement à la reconnaissance de la base TIMIT, le flux audio comporte des éléments non linguistiques qu'il convient d'isoler au préalable à la reconnaissance. Parmi ceux-ci, on peut distinguer principalement les silences, les bruits et la musique. Une première segmentation doit être réalisée afin d'isoler les segments de parole.

Cette séparation est communément obtenue à l'aide de GMM qui modélisent chacun des types de segment acoustique. Cette technique est employée par l'essentiel des systèmes participant à l'évaluation DARPA *Broadcast News* [DARPA, 1998]. Une autre méthode fondée sur un arbre de décision utilisant les histogrammes d'énergie, une détection voisée/non-voisée, un suivi de pitch, un DAP et une reconnaissance du locuteur a été également proposée [Montacié & Caraty, 1998]. Cette méthode, dont l'efficacité a été montrée, présente l'intérêt de ne pas nécessiter d'apprentissage (contrairement aux GMM) et donc ne requiert pas une segmentation préalable d'une partie des données.

L'existence d'un étiquetage de référence de la base TIMIT permet de ne pas avoir recours aux méthodes automatiques de segmentation parole/silence. Toutefois, afin de valider la faisabilité de l'approche, deux expériences simples de séparation automatique ont été réalisées. Dans un premier temps, la classification parole/silence est obtenue à l'aide de la règle des 50-ppv appliquée sur ces deux classes. Dans un second temps, seule l'énergie des observations est prise en compte. Les trames de parole ont leur énergie supérieure à un certain seuil. Le seuil utilisé (0,4) est déduit de l'histogramme des énergies à l'aide d'un critère de moindre erreur. L'erreur minimale est évaluée en attribuant un coût plus élevé à la mauvaise classification de la parole. Il paraît *a priori* moins dommageable de laisser du silence dans les zones étiquetées parole que d'éliminer à tort des zones de parole. Les résultats de ces deux expériences sont présentés dans le Tableau 2-7.

Règle de décision	Classes	Taux d'identification	
		Appr.	Test
50-ppv	Silence	80,7	80,9
	Parole	95,6	96,1
	Sil.+Par.	93,6	**93,9**
Energie	Silence	81,9	81,1
	Parole	93,1	93,7
	Sil.+Par.	91,5	91,8

Tableau 2-7 Taux d'identification locale de la séparation parole/silence en fonction de la règle de décision employée : 50-ppv ou seuil d'énergie.

La règle de décision dérivée des 50-ppv permet la meilleure séparation. Les résultats de l'identification locale en 39 classes par l'estimateur des 50-ppv (Tableau 2-2) laissaient présager ces bons résultats dans la mesure où la classe silence était déjà correctement reconnue à 91%. La méthode, pourtant très simple, du seuil d'énergie offre des résultats comparables à la méthode basée sur les 50-ppv. On peut en déduire que la plus grande part de l'information de discrimination entre le silence et la parole est contenue dans le coefficient d'énergie. Cette méthode peu coûteuse permet donc de réaliser automatiquement la séparation silence/parole. Les techniques utilisées sont relativement sommaires. Elles pourront être améliorées par exemple par la prise en compte de la densité de passage par zéro [Rabiner & Schafer, 1978].

Afin de situer cette étude dans un cadre idéal, la séparation silence/parole sera réalisée dans nos expériences à partir de l'étiquetage de référence de TIMIT. Les segments correspondant aux étiquettes *h#* (débuts et fins de phrases), *pau* (pauses) et *epi* (zones de faible énergie apparaissant typiquement lors des changements de voisement) sont supprimés. Les occlusions des occlusives (voisées et non-voisées) qui au préalable faisaient partie de la classe silence constituent maintenant une classe à elles seules. On a alors 39 classes finales et 46 classes intermédiaires (pour les systèmes HMM/GMM) : 48 classes moins *epi* et *sil* (qui regroupait *h#* et *pau*). Les résultats d'identification locale, après extraction des segments de silence, sont donnés dans le Tableau 2-8.

Estimateur	Taux d'identification	
	Appr.	Test
Standard		
75-GMM	48,0	43,6
50-ppv	57,4	52,6
Standard (sans silence dans le score)		
75-GMM	39,1	33,1
50-ppv	46,9	40,4
Après séparation silence/parole		
75-GMM	42,9	36,7
50-ppv	52,1	45,8

Tableau 2-8 Taux d'identification locale des estimateurs 75-GMM et 50-ppv
avec une séparation parole/silence.

Les expériences d'identification locale sur TIMIT sans silence font apparaître une baisse de résultats. Dans le cas standard, la classe silence, étant très bien reconnue et possédant un grand nombre de représentants, avait tendance à améliorer les performances globales d'identification. Si l'on considère les résultats d'identification standard sans intégrer la classe des silences lors du calcul du taux, celui-ci descend à 40,4% pour l'estimateur 50-ppv (soit un écart de 12,2%). Or, le taux d'identification finalement obtenu après la séparation silence/parole est de 45,8%. La séparation silence/parole a donc permis un gain de 5,4% (45,8 - 40,4%) sur l'identification des autres classes. Ce gain est dû à la disparition des confusions avec la classe des silences.

Ce résultat doit être relativisé car il correspond à une séparation silence/parole idéale. Par contre, il est remarquable que l'écart entre les performances des estimateurs 50-ppv et 75-GMM sur le test est toujours de 9,1% (contre 9,0% dans le cas standard) alors que la reconnaissance de la classe silence était plus favorable à l'estimateur 50-ppv (91% contre 76,3%, écart qui induisait un différence de 3,6% sur le score global).

Dans la suite de notre étude, une séparation idéale silence/parole sera considérée par défaut dans nos expériences. Cette première étape de l'adaptation des références à l'estimateur des K-ppv est suivie par une remise en cause de la position des frontières entre les segments phonétiques.

2.3.4.2. Corrections des erreurs d'alignement

Pour la réalisation de l'étiquetage de la base TIMIT, les transcriptions phonétiques ont d'abord été réalisées manuellement, puis la technique du bootstrap a été utilisée pour l'alignement [Seneff & Zue, 1988]. Un ensemble réduit a été aligné manuellement ; puis des modèles ont été appris permettant l'alignement automatique du reste de la base de données. L'alignement obtenu a finalement été vérifié par des experts. Il présente donc de fortes garanties quant à sa justesse.

Toutefois dans le cadre des systèmes de RAP continue, la parole n'est pas analysée de façon continue mais au travers de trames représentant des segments de signal de l'ordre de 25 ms. Certaines trames vont donc chevaucher des segments appartenant à deux classes phonétiques différentes. La règle appliquée pour ces trames a été de leur associer celle des deux classes qui couvre plus de la moitié de la durée de la trame. Cette règle a le mérite d'être simple mais n'est pas forcément cohérente avec la règle de décision utilisée pour l'identification. Ainsi, pour déterminer si les erreurs d'alignement conduisent à des erreurs d'identification par la règle des K-ppv, la corrélation entre la proximité des frontières et les erreurs d'identification a été étudiée.

Chaque trame de l'ensemble de test a été étiquetée par sa distance (en nombre de trames) à la trame frontière la plus proche. L'histogramme des distances a été calculé. On constate que 24% des trames de l'ensemble de test sont des trames frontières (distance valant 1). Cette opération a été renouvelée spécifiquement sur les trames mal identifiées par la règle des 50-ppv. On note alors que 29% des trames mal identifiées sont des trames frontières. Cet écart de 5% dans la distribution des distances révèle que la fréquence des trames frontières est plus élevée parmi les trames mal identifiées.

Nous proposons de modifier l'alignement de l'ensemble de référence afin d'adapter l'alignement au processus de décision des K-ppv. La règle des 50-ppv est appliquée sur l'ensemble de référence. Lorsqu'une trame mal reconnue est une trame frontière, son étiquette est modifiée si l'étiquette donnée par la règle des 50-ppv est cohérente avec l'étiquette voisine. Cette procédure est itérée jusqu'à ce plus aucune trame frontière ne soit modifiée. Le nombre de trames modifiées en fonction du nombre d'itération et le taux d'identification de l'ensemble de référence après modification des étiquettes évoluent rapidement lors de la première itération (43705 trames modifiées et +4% sur le taux). Puis le score se stabilise à 58% au bout de 4-5 itérations (soit une augmentation de 6%) et le nombre de frontières déplacées devient négligeable (Tableau 2-9). Il apparaît donc que les imprécisions d'étiquetage ont une influence importante sur les capacités d'identification de la règle des K-ppv.

Itérations	Etiquettes modifiées	Taux d'identification
0	0	52,1
1	43705	56,1
2	12535	57,1
3	6153	57,5
4	3597	57,8
5	2305	57,9
6	1571	58,0
7	1102	58,1
8	782	58,2
9	585	58,2
10	473	58,2

Tableau 2-9 Nombre d'étiquettes modifiées par itération et nouveau taux d'identification locale de l'ensemble de référence.

Le problème posé pour l'évaluation de la méthode est que si l'on suppose des erreurs d'alignement sur l'ensemble de référence, il faut aussi en supposer pour l'ensemble de test. Pour mesurer correctement l'efficacité de la méthode, il convient donc de l'appliquer aussi sur l'ensemble de test. Dans la mesure où les expériences d'identification locale ne constituent pas le but de notre étude mais un protocole d'évaluation intermédiaire, la modification de l'ensemble de test peut être tolérée. L'intérêt de la méthode sera validé lors des expériences de DAP qui ne font plus intervenir l'alignement temporel (du moins n'est-il pas pris en compte dans le protocole d'évaluation).

Les résultats en identification locale sont rapportés dans le Tableau 2-10 dans le cas standard (pas de segmentation), dans le cas d'un ajustement de la segmentation de l'ensemble de référence seul et dans le cas d'un ajustement de la segmentation des ensembles de référence et d'évaluation (apprentissage ou test).

Segmentation (Réf./Eval.)	Taux d'identification	
	Appr.	Test
Non/Non	52,1	45,8
Oui/Non	51,3	45,5
Oui/Oui	58,2	52,0

Tableau 2-10 Taux d'identification locale avec et sans re-segmentation des frontières des ensembles de référence et d'évaluation (apprentissage ou test).

Les résultats de l'identification locale confirment que la re-segmentation des frontières ne présente d'intérêt que lorsque l'ensemble d'évaluation est lui-même re-segmenté et pas seulement l'ensemble de référence (Tableau 2-10). On note alors que le simple déplacement des frontières entre segments phonétiques permet un gain de 6% en identification. Ce résultat est encourageant, il existe toutefois des trames (frontières ou non) pour lesquelles aucune affectation ne sera satisfaisante.

2.3.4.3. Pondérations des ppv

Certaines régions de l'espace de représentation sont très confuses. Les points qui y appartiennent ne sont pas fiables. Le chevauchement des trames sur des segments appartenant à des classes phonétiques différentes en est une des raisons principales. Pour tenir compte de ce problème, nous proposons lors de l'identification de pondérer la participation d'un point par un indice de confiance. Cette approche est différente de la pondération par la distance [Dudani, 1976] car elle ne porte pas sur une propriété générale des points (plus ils sont éloignés moins l'information qu'ils fournissent est fiable) mais sur une mesure de fiabilité attachée à chaque point en particulier. Ainsi, l'indice de confiance, évalué sur l'ensemble de référence, mesure le rapport des participations du point à des décisions correctes sur l'ensemble de ses participations.

La règle des 50-ppv est appliquée à chaque point de l'ensemble de référence. Lorsque la décision engendrée est en accord avec l'étiquetage de référence (sans correction des frontières), on incrémente le nombre de bonnes participations de chacun des ppv appartenant à la classe attendue. Dans tous les cas on incrémente leur nombre de participations. On obtient, ainsi, pour tous les points, un indice de confiance calculé comme le rapport entre le nombre de participations à des décisions correctes sur le nombre total de participations. La nouvelle règle des K-ppv est obtenue par comparaison des sommes des indices de bonne participation de ses K-ppv pour chaque classe.

Pondération	Taux d'identification	
	Appr	Test
Non	52,1	45,8
Oui	51,0	45,7

Tableau 2-11 Taux d'identification locale de l'estimateur des 50-ppv avec et sans pondération lors de la décision.

Cette méthode, dont les résultats sont fournis dans le Tableau 2-11 pour l'estimateur 50-ppv, donne lieu à une dégradation du taux d'identification en auto-cohérence et à une stagnation sur le test. Ces résultats peuvent s'expliquer par le fait que même si un point se trouve à proximité de points dont la confiance est faible, ils seront malgré tout les seuls à participer à la décision. Une solution à ce problème peut être d'éliminer les points de confiance faible de l'ensemble de référence. La méthode d'édition permet de mettre en œuvre une telle approche.

2.3.5. Edition et condensation de l'ensemble de référence

L'édition est une technique standard en RdF pour améliorer la reconnaissance de l'estimateur des K-ppv. Elle a pour effet d'éliminer les points non représentatifs de l'ensemble de référence. Cette technique est classiquement associée à la méthode de condensation qui élimine les points inutiles de l'ensemble d'apprentissage. Les deux méthodes permettent d'augmenter la rapidité de calcul de la règle des K-ppv mais seule l'édition est fondée pour améliorer la qualité de la reconnaissance.

2.3.5.1. Méthode d'édition

Le principe de la méthode d'édition introduite par Wilson [Wilson, 1972] est d'éliminer de l'ensemble de référence les points conduisant aux erreurs d'identification. On améliore ainsi d'une part la discrimination entre classes et de plus la rapidité du calcul en diminuant le nombre de distances à calculer.

Les mauvais points sont des points non représentatifs de la tâche ou trop « ambigus ». Deux approches peuvent être envisagées pour mettre en œuvre la méthode d'édition selon que l'on décide de l'élimination d'un point parce qu'il est mal reconnu ou parce qu'il est responsable d'erreurs de classification. Ainsi, l'approche classique de l'édition éliminé de l'ensemble de référence tous les points mal reconnus par la règle des K-ppv. L'algorithme de Wilson correspondant à l'approche classique procède en deux étapes qui sont décrites dans la Figure 2-8.

```
•  Traitement
      Application de la règle des K-ppv sur l'ensemble de référence
      lui-même ;
      Tous les points dont l'identification n'est pas en accord
      avec l'étiquetage de référence sont éliminés de l'ensemble de
      référence ;
•  Arrêt
      Le traitement est répété un nombre donné de fois ou jusqu'à
      ce qu'un nombre maximum de points soit éliminé ;
```

Figure 2-8 Algorithme d'édition de Wilson.

On peut montrer que la règle du 1-ppv appliquée sur l'ensemble de référence édité a une probabilité d'erreur inférieure à la règle de base du 1-ppv [Devijver & Kittler, 1982]. Ainsi, on diminue le temps de calcul et on améliore les performances. Ces dernières peuvent être d'avantage améliorées par des itérations de la procédure d'édition (*Multiedit*, [Devijver & Kittler, 1982]), pour autant que l'on dispose au préalable d'un ensemble de référence suffisamment grand.

L'intérêt de la méthode précédente est que la diminution de l'erreur d'estimation a été montrée. Toutefois, la démonstration ne s'applique que lorsque la règle du 1-ppv est utilisée avec l'ensemble édité. Son extension au cas des K-ppv ne paraît pas simple. Ainsi, nous avons été amenés à développer une nouvelle méthode plus adaptée à l'utilisation de la règle des K-ppv avec l'ensemble édité. L'approche que nous proposons, en relation avec la méthode de pondération développée dans §2.3.4.3, élimine de l'ensemble de référence tous les points qui participe plus souvent à des cas de mauvaises classifications que de bonnes.

Le principe retenu est d'éditer les points lorsqu'ils participent à trop d'erreurs d'identification sur l'ensemble de référence. Pour cela, on fait la moyenne de leurs interventions comme ppv d'autres points lors de l'identification de l'ensemble de référence : lorsqu'ils participent à une identification correcte ils ont un bonus, lorsqu'ils interviennent dans une erreur d'identification ils sont pénalisés. Les points dont le solde est positif en bonus sont conservés. Les autres sont éliminés de l'ensemble de référence. Le rapport entre la valeur unitaire de bonus et celle de pénalité permet de moduler l'importance accordée aux erreurs. Nous avons choisi un bonus de 2 et une pénalité de

1. Ainsi, seuls les points ayant participé à 2 fois plus d'erreurs que d'identifications correctes sont éliminés de l'ensemble de référence. L'algorithme se décompose en deux étapes qui sont détaillées dans la Figure 2-9.

La méthode proposée permet une élimination plus sure des points responsables des erreurs de classification que dans le cas de la méthode de Wilson. Toutefois, la diminution de l'erreur d'estimation après application de la méthode n'est plus assurée. En l'absence de démonstration théorique, elle devra être constatée empiriquement. Les deux méthodes présentées ont été testées sur TIMIT. Les résultats sont regroupés avec ceux de la méthode de condensation dans §2.3.5.3.

```
• Traitement
    Identification locale de l'ensemble de référence.
    Pour chaque point :
        si l'identification est correcte : bonus pour tous les ppv
        du point appartenant à la classe de référence ;
        sinon : pénalité ;
    Pour chaque point : si le solde bonus/pénalité est négatif,
    il est éliminé de l'ensemble de référence ;
• Arrêt
    Le traitement est répété un nombre donné de fois ou jusqu'à
    ce qu'un nombre maximum de points soit éliminé ;
```

Figure 2-9 Algorithme d'édition proposé.

2.3.5.2. Méthode de condensation

Le principe de la méthode de condensation introduite par Hart [Hart, 1968] est d'éliminer de l'ensemble de référence les points ne participant pas de manière pertinente à la reconnaissance. Pour ce faire, on recherche un sous-ensemble « consistant », non nécessairement optimal, de l'ensemble de référence. Par "consistant", on entend que ce sous-ensemble utilisé avec la règle des K-ppv classe de nouveau correctement les points qui l'étaient sur l'ensemble de référence intégral. Nous considérerons deux approches pour condenser l'ensemble de référence.

L'algorithme de condensation de Hart est subordonné à l'utilisation de la règle du 1-PPV. Son déroulement est repris dans la Figure 2-10. Le tableau T_1 contient l'ensemble recherché. Tous les points condensés ont été choisis s'ils ne peuvent être correctement classés avec un ensemble de référence réduit. Cette procédure favorise les points proches des frontières entre classes. L'obtention d'un sous-ensemble « consistant » suppose donc que l'on élimine de l'ensemble de référence les points au centre des classes. C'est sur cette propriété que nous proposons un algorithme de condensation compatible avec la règle des K-ppv.

```
• Initialisation
    Création de T₁(tableau des points de référence) et T₂ (tableau
    des points correctement classés); le premier point est placé
    dans T₁ tous les autres dans T₂ ;
• Traitement
    Pour tous les points de T₂ :
        On les place dans T1 s'ils sont mal classés par la règle du
        1-PPV appliquée sur l'ensemble de référence constitué des
        points du tableau T₂, sinon on les laisse dans T₂ ;
• Arrêt
    T₂ est vide ou il ne se produit plus de transfert entre T₂ et
    T₁ ;
```

Figure 2-10 Algorithme de condensation de Hart.

Les points au centre des classes ont, entre autres caractéristiques, d'être très bien reconnus. Ils se distinguent donc par un nombre élevé de ppv de la bonne classe. L'algorithme que nous proposons consiste en l'élimination progressive des points les mieux reconnus tant que le taux de reconnaissance global de l'ensemble de référence ne descend pas en deçà d'un certain seuil. Ce seuil peut être absolu ou relatif au taux d'identification avant condensation. L'algorithme de condensation obtenu, détaillé Figure 2-11, est donc fondé sur la règle des K-ppv et procède par itération sur le niveau de reconnaissance des points en cours d'élimination (matérialisé par le nombre de ppv corrects).

La condensation est favorisée par l'utilisation préalable de la méthode d'édition. L'avantage majeur de la méthode de condensation réside dans le gain en temps de calcul. Le taux d'erreur d'identification doit être faiblement affecté par cette procédure. De plus, il faut souligner que cette méthode est « destructrice » : c'est-à-dire qu'elle ne conserve pas toute l'information initiale. L'ensemble obtenu n'est pas équivalent en terme de quantité d'information à l'ensemble initial. Dans la pratique, ceci signifie que l'on doit être sûr avant de condenser l'ensemble de référence que seule la règle des K-ppv sera appliquée sur celui-ci ultérieurement.

```
•  Traitement
      Pour k=K jusqu'à 1
         Tous les points sont ordonnés en fonction du nombre de
         leurs ppv de la bonne classe parmi K ;
         Tous les points ayant k ppv de la bonne classe sont
         éliminés successivement (en décomptant des ppv les points
         déjà éliminés) ;
         Calcul du taux de reconnaissance après élimination des
         points condensés ;
•  Arrêt
      Le nouveau taux d'identification de l'ensemble de référence
      franchit un seuil minimal ;
```

Figure 2-11 Algorithme de condensation proposé.

2.3.5.3. Expériences d'édition/condensation

Les expériences d'édition et de condensation sont réalisées conjointement. Elles sont souvent utilisées successivement : l'édition d'abord, puis la condensation. La Figure 2-12 illustre les effets caractéristiques de l'une et l'autre méthode sur les ensembles de données associées à deux classes. L'édition élimine les points du mauvais coté de la surface de décision (donc mal classés). La condensation tend à conserver les points proches de la surface de décision.

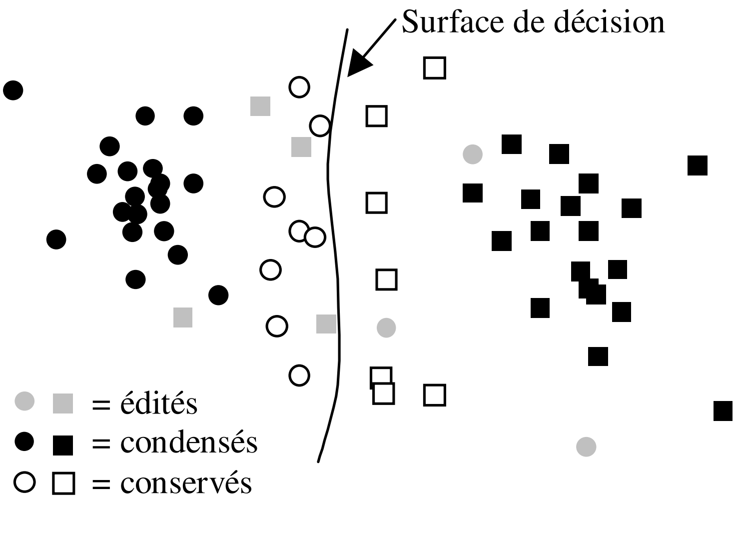


Figure 2-12 Effets de l'utilisation des méthodes d'édition et de condensation sur la séparation de 2 classes (● et ■).

Le protocole expérimental diffère des algorithmes introduits dans les sections précédentes. Pour des considérations de temps de calcul, les K ppv des points ne sont pas recalculés au sein d'une même itération. L'élimination d'un point de l'ensemble de référence conduit à ce qu'il ne soit plus pris en compte parmi les ppv des points suivants. Formellement, elle devrait conduire à la recherche d'un nouveau ppv.

Ensemble de référence	Nb Trames	Taux d'identification		
		Référence	Appr	Test
Initial	950800	52,1	52,1	45,8
Edition1	495409	**89,3**	49,2	45,7
Edition2	489435	49,7	50,2	46,0
Condensation	840602	45,9	51,7	45,3

Tableau 2-12 Taux d'identification locale de l'estimateur des 50-ppv pour des ensemble de référence édités et condensés sur trois ensembles d'évaluation : référence, apprentissage (=référence initial) et test.

Trois ensembles d'évaluation ont été considérés dans nos expériences : l'ensemble de référence, l'ensemble d'apprentissage (qui correspond dans tous les cas à l'ensemble de référence initial) et l'ensemble de test. Si l'on se réfère à l'évaluation sur l'ensemble de test, les expériences d'identification locale montrent des performances comparables avec et sans application des techniques d'édition/condensation (Tableau 2-12). Les évaluations en auto-cohérence montrent, par contre, une dégradation des performances. La seule exception notable concerne l'auto-cohérence de l'édition de Wilson (Edition1) qui avec 89,3% présente un taux d'identification très élevé. Ce bon comportement n'est pas confirmé sur le test où aucune amélioration n'est relevée.

Les résultats d'édition montrent qu'il est possible de diviser par deux la taille de l'ensemble de référence sans diminuer pour autant les taux d'identification locale. Cette constatation peut permettre de réduire encore le coût de calcul de l'estimateur des K-ppv. Les performances en temps de calcul de l'estimateur des K-ppv n'étant pas notre objectif, nous conserverons, dans la suite de notre exposé, l'ensemble de référence initial. Afin de mesurer complètement l'intérêt des techniques d'édition/condensation dans le cadre de la RAP continue, les ensembles de référence édités/condensés seront évalués dans le cadre de la reconnaissance segmentale dans le chapitre suivant.

2.4. Conclusion

Un effort important a été consacré ces dernières années à l'introduction d'une plus grande discrimination dans les modèles de Markov utilisés en RAP continue. Les techniques de l'état de l'art reposent principalement sur un ré-apprentissage global des modèles à l'aide d'un critère discriminant. Ces techniques ont en commun d'être coûteuses en temps de calcul. De plus, les critères utilisés nécessitent le recours aux techniques d'optimisation de type descente de gradient stochastique et conduisent à des mises en œuvre complexes dont les propriétés de convergence ne sont pas toujours bien assurées. Une alternative consiste à introduire la discrimination localement au niveau des états des modèles.

Un certain nombre de méthodes de discrimination locale, issues des recherches en RdF, est disponible. Parmi celles-ci, les méthodes linéaires sont peu adaptées aux espaces de représentation de grande dimension et aux tâches présentant des surfaces de décision de grande complexité. Parmi les approches non-linéaires, les réseaux de neurones ont été activement étudiés dans le cadre de la reconnaissance de la parole. Toutefois, ils se heurtent aux même problèmes de complexité de mise en œuvre et de limitation des preuves de convergence que les techniques d'apprentissage global à base de critères discriminants. Notre choix s'est donc porté vers un estimateur simple et ne nécessitant pas d’hypothèses simplificatrices : l’estimateur non-paramétrique des K-ppv.

Cet estimateur n’avait jusqu’alors pas été introduit dans le cadre de la RAP continue en raison d’un coût de calcul important. Ce coût ayant été fortement diminué, son usage a pu être envisagé. Les expériences d’identification locale à l'aide de la règle des K-ppv ont confirmé le meilleur comportement de l'estimateur des K-ppv par rapport à une méthode paramétrique telle que les GMM. Un écart de l’ordre de 10% est relevé sur l’identification des trames du noyau de test de TIMIT entre les estimateurs des 50-ppv et 75-GMM. L’introduction de l’estimateur des K-ppv dans un système markovien de RAP continue peut donc être envisagée avantageusement pour améliorer d’une part la qualité de la représentation probabiliste locale des données et d’autre part la discrimination globale des modèles.

3. APPRENTISSAGE D'UN SYSTEME MARKOVIEN K-PPV

Après avoir confirmé ses bonnes propriétés locales, l'estimateur des K-ppv est introduit dans le contexte de reconnaissance segmentale des modèles de Markov cachés. La mise en œuvre d'un système markovien K-ppv implique la re-formulation des probabilités d'émissions des état. Pour cela, il devient nécessaire de déterminer une information non préalablement disponible : l'association des données de référence aux états des modèles.

Un cadre formel est présenté pour la prise en compte de cette information : le degré d'association d'un vecteur de référence sur un état sera quantifié par un coefficient d'appartenance appris sur les données acoustiques d'entraînement. L'algorithme de Baum-Welch, qui sert habituellement pour l'apprentissage des paramètres des systèmes HMM/GMM, est adapté pour permettre l'apprentissage des coefficients d'appartenance. Cette adaptation nous amène à discuter différents choix. Deux formules de réestimation et trois procédures d'initialisation sont proposées et testées. Le problème, particulier à l'estimateur des K-ppv posé par les vecteurs spectraux ayant des probabilités nulles sur leur modèle, est aussi traité ainsi que le choix de la probabilité conditionnelle employée. Finalement, l'opportunité de l'apprentissage par la réestimation connectée est évaluée.

Cette étude conduit finalement au développement d'un système de base HMM/K-ppv dont les performances sont comparées sur deux tâches de reconnaissance phonétique (identification segmentale et DAP) à celles d'un système de base HMM/GMM.

3.1. Apprentissage EM adapté à l'estimateur des K-ppv

L'adaptation de la probabilité d'émission d'état des HMM à l'estimateur des K-ppv nécessite l'introduction d'un nouveau paramètre : les coefficients d'appartenance. L'apprentissage de ces coefficients est réalisé par l'algorithme de Baum-Welch avec une formule de réestimation adéquate.

3.1.1. Calcul de la probabilité d'émission des états

Les probabilités d'émission des états des HMM sont calculées à partir des probabilités conditionnelles des états. L'estimation par K-ppv pose alors le problème de l'affectation des points de référence sur les états du modèle représentant leur classe phonétique. L'introduction des coefficients d'appartenance des points aux états des HMM permet de lever cette difficulté.

3.1.1.1. Calcul de probabilités conditionnelles

Comme nous l'avons introduit pour exposer les principes de l'estimation non-paramétrique (voir §2.2.3), un estimateur du maximum de vraisemblance de la probabilité d'un vecteur x_0 dans un espace de représentation dont nous connaissons n points est donné par :

$$\hat{P}(x_O \in R) = \frac{K}{n} \tag{3-1}$$

où K est le nombre d'échantillons compris dans la région R autour de x_0. Si nous considérons maintenant un espace dans lequel sont représentées C classes, chacune par n_i échantillons, la probabilité de x_0 conditionnée à la classe i peut être estimée par :

$$\hat{P}(x_O \in R / i) = \frac{k_i}{n_i} \text{ avec } \sum_{i=1}^{C} k_i = K \text{ et } \sum_{i=1}^{C} n_i = n \tag{3-2}$$

où k_i est le nombre d'échantillons de la classe i contenus dans la région R et n_i le nombre de points de référence de la classe i. On obtient ainsi un estimateur asymptotiquement sans biais de la vraisemblance de x_0 sur la classe i.

L'utilisation de l'estimateur des K-ppv nous amène à préciser certaines propriétés des probabilités dans un espace continu. La probabilité d'un point d'un espace continu est nulle. En effet, la masse des probabilités étant répartie sur l'ensemble de l'espace, la probabilité associée à une région de cet espace a une probabilité qui tend vers 0 lorsque le volume de la région tend vers 0. Il faut donc estimer la probabilité d'un point par la probabilité d'une région de faible rayon centrée sur ce point. Les fdp mesurent en chaque point la variabilité de la probabilité. Il est alors possible d'obtenir la probabilité d'une région de l'espace en intégrant la fdp sur une région de volume suffisamment faible.

Dans le cas des estimateurs continus, cette intégration est coûteuse et pose le problème du choix de la région d'intégration. Dans le cas des estimateurs non-paramétriques, la définition de la région est intrinsèque au processus d'estimation. Elle est soit fixe dans le cas des estimateurs à noyaux soit variable jusqu'à intégrer K points dans le cas de l'estimateur des K-ppv. L'estimation de la probabilité est assimilée à la probabilité d'un voisinage de ce point.

Il est à noter que la probabilité (3-2) est liée à la région R, variable dans le cas de l'estimation par K-ppv. On peut alors s'interroger sur l'influence sur le processus de reconnaissance des variations de R lors des évaluations successives de P. En reconnaissance locale, la région étant la même pour toutes les classes considérées, R n'influe pas sur l'identification (cf. §2.3.2). De même, lors de la reconnaissance segmentale, le calcul de la probabilité d'un segment sera obtenu par la concaténation des estimations de chaque trame. Ainsi, les variations de la région R utilisée pour l'estimation des probabilités seront partagées par toutes les solutions étudiées et n'auront pas d'effet sur le processus de reconnaissance. Dorénavant, nous ne préciserons donc plus la région d'estimation dans la formulation des probabilités.

De l'estimation de la vraisemblance, il est possible de déduire un estimateur de la probabilité *a posteriori* de la classe i de x_0 en utilisant la règle de Bayes :

$$\hat{P}(i/x_0) = \frac{\hat{P}(x_0/i)\hat{P}(i)}{\hat{P}(x_0)} \qquad (3\text{-}3)$$

Les probabilités *a priori* des classes peuvent être évaluées par simple comptage parmi les n points de l'ensemble de référence :

$$\hat{P}(i) = \frac{n_i}{n} \qquad (3\text{-}4)$$

Et, donc, en utilisant (3-1) et (3-2), on obtient la probabilité *a posteriori* de la classe i de x_0 :

$$\hat{P}(i/x_0) = \frac{\frac{k_i}{n_i} \times \frac{n_i}{n}}{\frac{K}{n}} = \frac{k_i}{K} \qquad (3\text{-}5)$$

Le calcul des probabilités conditionnelles (vraisemblances ou *a posteriori*), à l'aide de l'estimateur des K-ppv, nécessite donc un ensemble de référence dont les points sont associés de manière univoque aux classes considérées (afin de déterminer les k_i). L'introduction de l'estimateur des K-ppv dans les HMM implique d'associer les points de l'ensemble de référence aux états des modèles et non plus aux classes phonétiques.

3.1.1.2. Coefficients d'appartenance

Jusqu'à présent, nos considérations se sont portées sur le calcul des probabilités conditionnelles des vecteurs sur les classes phonétiques. Dans le cadre des HMM, les vecteurs sont émis par les états des modèles. Les vecteurs de l'ensemble de référence doivent être associés aux sous-classes phonétiques représentées par les états markoviens. La probabilité d'émission de l'état i pour le vecteur x_0 donnée par l'estimateur des K-ppv est alors :

$$b_i(x_0) = \frac{k_i}{n_i} \qquad (3\text{-}6)$$

où k_i représente le nombre de plus proches voisins de x_0 associés à l'état i et n_i le nombre total de vecteurs de l'ensemble d'apprentissage associés à l'état i. Le problème auquel nous sommes confrontés est de déterminer l'association entre les vecteurs de référence et les états des HMM.

Par le biais d'un étiquetage de référence, un vecteur est associé au HMM représentant sa classe phonétique. Mais ensuite, l'association du vecteur à un des états du HMM doit être obtenue automatiquement dès lors qu'aucune expertise n'est disponible à ce niveau. Par ailleurs, l'assignation d'un vecteur à un unique état nous paraît une solution peu souhaitable dans la mesure où celle-ci sera obtenue automatiquement.

Les considérations précédentes font apparaître une limitation de l'estimation par K-ppv qui dépasse le cadre de son introduction dans les systèmes markoviens. L'association univoque des vecteurs sur les classes à reconnaître n'offre pas de moyen simple pour gérer l'incertitude ou l'imprécision sur les données de référence. Ce point peut se révéler délicat dans des domaines où même l'avis de l'expert, lorsqu'il est disponible peut-être soumis à caution. Ainsi, certains échantillons de référence peuvent être très caractéristiques de la classe à laquelle ils appartiennent lorsque d'autres peuvent être atypiques.

La théorie des ensembles flous [Zadeh, 1965] permet d'apporter une certaine réponse à cette limitation à travers la notion de degrés d'appartenance. A chaque vecteur x_0 de l'ensemble de référence est attribué un coefficient d'appartenance $u_i(x_0)$ sur chaque classe i tel que :

$$u_i(x_0) \in [0,1] \qquad \sum_{i=1}^{C} u_i(x_0) = 1 \tag{3-7}$$

avec C nombre total de classes (dans le cas d'un système HMM, on aura $C=S$, nombre total d'états dans le système). De façon générale, ces coefficients peuvent être attribués par des experts ou de façon arbitraire ou encore être évalués, par exemple en combinant les degrés d'appartenance des voisins du point. Une alternative, proposée par Keller [Keller, Gray *et al.*, 1985] et prolongée par [Zouhal & Denoeux, 1997], tient compte des distances entre voisins dans cette combinaison. Sur des données synthétiques simples, la règle de décision obtenue est plus performante que la règle des K-ppv standard [Zouhal & Denoeux, 1997].

Cette approche de l'association non univoque des points de référence sur les classes a été formalisée au travers de la théorie de Dempster-Shafer [Denoeux, 1995]. Au-delà des considérations liées aux performances de reconnaissance, ce cadre permet l'introduction d'une information très riche dans l'opérateur de reconnaissance, puisqu'il devient notamment possible d'associer une mesure de confiance à une classification. Les développements issus de cette formalisation visent à améliorer l'opérateur de reconnaissance fondé sur la règle des K-ppv. Dans notre cas, l'opérateur de reconnaissance est fondé sur les HMM. L'intervention des coefficients d'appartenance sera donc limitée à l'expression de la probabilité d'émission d'un état :

$$b_i(x_0) = \frac{\sum_{k=1}^{K} u_i(ppv(k, x_0))}{U_i} \quad \text{avec} \quad U_i = \sum_{v=1}^{V} u_i(v) \tag{3-8}$$

où $ppv(k,x_0)$ est une fonction qui renvoie le $k^{\text{iéme}}$ ppv du vecteur x_0 et U_i est la somme des coefficients d'appartenance à l'état i sur les V vecteurs de référence. Cette formulation de l'estimation de la probabilité d'émission d'un état d'un HMM est la plus pratique à mettre en œuvre. Pourtant, pour des raisons de généralisation, une seconde formulation est introduite :

$$b_i(x_0) = \frac{\sum_{v=1}^{V} u_i(v) \times ppv_K(x_0, v)}{U_i} \tag{3-9}$$

où $ppv_K(x_0,v)$ devient une fonction indicatrice qui vaut 1 si le vecteur v est un des K-ppv de x_0 et 0 sinon :

$$ppv_K(x_0, v) = \begin{cases} 1 \text{ si } v \in \text{K}-\text{ppv de } x_0 \\ 0 \text{ sinon} \end{cases} \tag{3-10}$$

Les deux formulations (3-8) et (3-9) aboutissent au même résultat. Du point de vue pratique, l'équation (3-8) sera préférée puisque qu'elle n'oblige pas à parcourir les V vecteurs de l'ensemble de référence mais uniquement les K-ppv du point considéré. Du point de vue théorique, l'équation (3-9) sera utilisée pour démontrer les formules de réestimation.

L'utilisation de l'estimateur des K-ppv permet une formulation aussi simple de la probabilité *a posteriori* que de la vraisemblance (cf. §3.1.1.1). Dans le cas des GMM, le passage de l'une à l'autre impose un effort de normalisation supplémentaire. Il est donc possible d'utiliser comme probabilité d'émission d'état des probabilités *a posteriori* :

$$b_i(x_0) = \frac{\sum_{v=1}^{V} u_i(v) \times ppv_K(x_0, v)}{K} \tag{3-11}$$

Conformément au §3.1.1.1, la probabilité *a posteriori* est obtenue en divisant le nombre de K-ppv de la classe par K et non plus par une valeur proportionnelle a la probabilité *a priori* de la classe (U_i). Les probabilités *a posteriori* ont l'avantage d'être normalisées. Le recours à une probabilité *a posteriori* dans le système HMM/K-ppv sera étudié dans §3.2.1. Par défaut, nous utiliserons la vraisemblance (3-8).

Il est maintenant nécessaire d'apprendre les degrés d'appartenance des vecteurs de l'ensemble de référence. Ces coefficients constituent les paramètres des fdp basées sur l'estimateur des K-ppv. Ils deviennent avec les matrices de transition les paramètres du système HMM/K-ppv.

3.1.2. Apprentissage des coefficients d'appartenance

L'apprentissage d'un HMM consiste à trouver les meilleures valeurs de ses paramètres au sens d'un critère donné. Comme nous l'avons exposé (cf. §1.1.3.3 et §2.1), un certain nombre de critères existe et leur méthode d'apprentissage associée : MLE, MMIE, MCE… L'estimateur des K-ppv étant explicitement discriminant l'apprentissage par MLE paraît suffisant en première approche.

3.1.2.1. Formules de réestimation

Les paramètres de l'estimateur des K-ppv sont les coefficients d'appartenance (§3.1.1.2) aux états du HMM associés à chaque vecteur de l'ensemble de référence. L'algorithme de Baum-Welch étant itératif, ces paramètres doivent avoir des valeurs initiales. La segmentation uniforme des séquences d'apprentissage fournit une première estimation des coefficients d'appartenance, adaptée au cas des modèles gauche-droits. Les séquences sont divisées en autant de sous-séquences contiguës qu'il existe d'états dans le HMM. Puis, chacun des vecteurs de ces sous-séquences est assigné à un état dans l'ordre d'apparition temporel. Des procédures alternatives seront proposées et testées dans §3.2.3.

L'algorithme de Baum-Welch est une version particulière de l'algorithme d'Espérance-Maximisation (*Expectation-Maximisation*, EM) [Dempster, Laird *et al.*, 1977 ; Redner & Walker, 1984] adaptée au critère MLE. L'algorithme EM peut-être présenté succinctement comme une technique de calcul de maxima locaux de fonctions. Il est généralement introduit pour des problèmes d'optimisation qui seraient simplifiés par la connaissance d'une variable supplémentaire. Dans le cas de l'émission d'un segment acoustique par un HMM, cette variable est la suite d'état qui a généré la séquence acoustique. La donnée manquante est estimée à l'aide des données observées (la séquence acoustique) et des paramètres actuels du modèle générateur. La vraisemblance des données observées sur le modèle est alors maximisée en transformant les paramètres du modèle générateur à l'aide des formules de réestimation. Les paramètres actualisés du modèle permettent une nouvelle estimation de la donnée manquante et ainsi de suite jusqu'à convergence.

L'intérêt de l'algorithme de Baum-Welch est de conduire à une formulation analytique des nouvelles valeurs de paramètres au travers des formules de réestimation. Elles découlent de la maximisation d'une fonction auxiliaire. Deux approches sont présentées pour obtenir les formules de réestimation adaptées aux coefficients d'appartenance.

3.1.2.2. Réestimation Locale

Au-delà des démonstrations rigoureuses permettant de dériver les formules de réestimation, un principe simple se dégage. Lors de la phase E, l'ensemble des chemins possibles (séquences d'états) d'un segment acoustique à travers son HMM est évalué. Les formules de réestimation utilisées lors de la phase de maximisation reviennent alors à évaluer le poids relatif d'un certain nombre de chemins particuliers. Ces chemins sont ceux faisant intervenir le paramètre que l'on cherche à évaluer.

Ainsi, pour associer à chaque vecteur o_t d'une séquence d'apprentissage O son coefficient d'appartenance à un état i, l'estimation peut être obtenue en comptant la part prise par les chemins faisant correspondre o_t et i (*i.e.* $q_t=i$) par rapport à l'ensemble des chemins possibles :

$$\hat{u}_i(o_t) = \frac{P(O, q_t = i / \lambda)}{P(O / \lambda)} \tag{3-12}$$

où λ représente le HMM associé à O selon l'alignement de référence.

Les probabilités du numérateur et du dénominateur de (3-12) sont obtenues lors de l'étape E. Deux passes sont nécessaires pour les obtenir, permettant le calcul des variables directe et rétrograde (voir §1.1.3.1) :

$$\hat{u}_i(o_t) = \frac{\alpha_i(t) \times \beta_i(t)}{\alpha_F(T)} \tag{3-13}$$

avec F état final du modèle et T la longueur de la séquence O.

Ces considérations sont généralisables à un ensemble de Ex segments d'apprentissage. Dans ce cas, le dénominateur et le numérateur sont sommés séparément sur chaque exemple e :

$$\hat{u}_i(o_t) = \frac{\sum_{e=1}^{Ex} \alpha_i^e(t) \times \beta_i^e(t)}{\sum_{e=1}^{Ex} \alpha_F^e(T)} \tag{3-14}$$

Cette approche est simple et a permis de fournir les premiers résultats de nos expériences sur l'introduction de l'estimateur des K-ppv dans le formalisme markovien [Lefèvre, Montacié *et al.*, 1997; Montacié, Caraty *et al.*, 1997]. Elle présente toutefois deux inconvénients majeurs. Tout d'abord, la convergence de l'algorithme de Baum-Welch avec cette formule de réestimation n'est pas démontrée. De plus, elle oblige à apprendre les coefficients d'appartenance en utilisant comme ensemble d'apprentissage l'ensemble de référence lui-même. Or, ceci n'est nullement obligé ni désirable pour des considérations de généralisation. Une deuxième formule de réestimation est donc proposée afin de résoudre les deux limitations de la réestimation locale.

3.1.2.3. Réestimation Propagée

L'équation (1-6) permet d'obtenir la vraisemblance d'une suite de trames acoustiques sur un HMM. Cette équation est fondée sur l'intégration d'une variable cachée ; la séquence d'états parcourue lors de l'émission :

$$Q = q_1, \ldots, q_t, \ldots, q_T \tag{3-15}$$

Une seconde variable cachée est introduite ; la séquence de ppv parcourue lors de l'émission :

$$P = p_1, \ldots, p_t, \ldots, p_T \tag{3-16}$$

Il existe N^T séquences d'états Q et K^T séquences de ppv P. La probabilité d'une séquence d'états et de ppv particulière est donnée par[1] :

$$P(O, Q, P / \lambda) = a_{1q_1} \prod_{t=1}^{T} b_{q_t}^{p_t}(o_t)\, a_{q_t q_{t+1}} \tag{3-17}$$

avec

$$b_{q_t}^{p_t}(o_t) = \frac{u_{q_t}(p_t) \times ppv_K(o_t, p_t)}{U_{q_t}} \tag{3-18}$$

avec $ppv_K(o_t,p_t)$ toujours définie par (3-10) et U_{qt} par (3-8). L'équation (3-18) représente la probabilité d'émission associée à un vecteur particulier de l'ensemble de référence. Une interprétation de l'équation (3-17) est qu'il existe N^T séquences d'états qui peuvent avoir produit l'observation O et pour chacune de ces séquences V^T branchements possibles sur les vecteurs de référence. En pratique, seuls K^T branchements seront effectivement empruntés : la fonction $ppv_K(\ )$ annule les autres. La vraisemblance totale est obtenue en sommant l'équation (3-17) sur toutes les séquences d'états et de ppv possibles :

$$P(O / \lambda) = \sum_{Q,P} \left\{ \prod_{t=1}^{T} b_{q_t}^{p_t}(o_t) \times a_{q_t q_{t+1}} \right\} \tag{3-19}$$

La maximisation de la probabilité d'émission d'un modèle est obtenue par le biais d'une fonction auxiliaire définie sur la mesure de Kullback-Leibler [Kullback & Leibler, 1951] :

$$F(\lambda, \hat{\lambda}) = \sum_{Q,P} P(O, Q, P / \lambda) \times \log P(O, Q, P / \hat{\lambda}) \tag{3-20}$$

[1] Le terme a_{1q1} sera omis dans la suite afin de simplifier les équations. Les résultats démontrés dans le paragraphe s'appliqueront sur cet élément de la matrice transition de la même façon que pour les autres.

Dans [Liporace, 1982] , il est montré qu'accroître F en fonction de $\hat{\lambda}$ accroît la vraisemblance $P(O/\hat{\lambda})$ de l'observation sur le modèle. La maximisation de la fonction auxiliaire peut être divisée en deux maximisations séparées du fait du logarithme :

$$\log P(O,Q,P/\hat{\lambda}) = \sum_{t=1}^{T} \log \hat{a}_{q_t q_{t+1}} + \sum_{t=1}^{T} \log \frac{\hat{u}_{q_t}(p_t)}{U_{q_T}} \times ppv_K(o_t, p_t) \qquad (3\text{-}21)$$

La fonction $ppv_K(\ )$ n'est pas intégrée dans le *log* dans la mesure où elle n'intervient que pour annuler les séquences non rencontrées et pas dans la valuation de la probabilité. La fonction auxiliaire F peut alors s'exprimer :

$$F(\lambda,\hat{\lambda}) = \sum_{i=2}^{N} F_{\hat{a}_i}(\lambda, \{\hat{a}_{ij}\}_{j=2}^{N}) + \sum_{i=1}^{N} F_{\hat{u}_i}(\lambda, \{\hat{u}_i(v)\}_{v=1}^{V}) \qquad (3\text{-}22)$$

avec

$$\begin{aligned} F_{\hat{a}_i}(\lambda, \{\hat{a}_{ij}\}_{j=2}^{N}) &= \sum_{Q,P} P(O,Q,P/\lambda) \sum_{t=1}^{T} \log \hat{a}_{q_t q_{t+1}} \times \delta_i^{q_t} \\ &= \sum_{j=2}^{N} \sum_{t=1}^{T} P(O, q_t = i, q_{t+1} = j/\lambda) \times \log \hat{a}_{ij} \end{aligned} \qquad (3\text{-}23)$$

$$\begin{aligned} F_{\hat{u}}(\lambda, \{\hat{u}_i(v)\}_{v=1}^{V}) &= \sum_{Q,P} P(O,Q,P/\lambda) \sum_{t=1}^{T} ppv(o_t, p_t) \times \log \frac{\hat{u}_{q_t}(p_t)}{U_{q_t}} \times \delta_i^{q_t} \\ &= \sum_{v=1}^{V} \sum_{t=1}^{T} P(O, p_t = v, q_t = i/\lambda) \times ppv_K(o_t, v) \times \log \frac{\hat{u}_i(v)}{U_i} \end{aligned} \qquad (3\text{-}24)$$

avec δ le symbole de Kronecker qui vaut 1 lorsque ses deux indices coïncident, 0 sinon. La maximisation de $F_{\hat{a}i}$ et $F_{\hat{u}i}$ considérant les contraintes :

$$\sum_{j=2}^{N} \hat{a}_{ij} = 1 \text{ et } \hat{a}_{ij} \geq 0 \text{ pour tout } i \qquad (3\text{-}25)$$

$$\sum_{v=1}^{V} \frac{\hat{u}_i(v)}{U_i} = 1 \text{ et } \hat{u}_i(v) \geq 0 \text{ pour tout } v \qquad (3\text{-}26)$$

est donc de la forme $\sum_{i=1}^{I} x_i \log y_i$ avec la contrainte $\sum_{i=1}^{I} y_i = 1$ et $y_i \geq 0$. Cette expression atteind son maximum global pour :

$$y_i = \frac{x_i}{\sum_{i=1}^{I} x_i} \text{ pour tout } i=1 \text{ à } I \qquad (3\text{-}27)$$

La justification de ce résultat se trouve, entre autre, dans [Baum, 1972]. Intuitivement, les x_i étant fixés, on répartit le poids total des y_i (en l'occurrence 1) de manière à faire correspondre les plus fortes valeurs de y_i aux plus fortes valeurs de x_i. On obtient les formules de réestimation suivantes:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} P(O, q_t = i, q_{t+1} = j / \lambda)}{\sum_{t=1}^{T} P(O, q_t = i / \lambda)} \quad (3\text{-}28)$$

$$\frac{\hat{u}_i(v)}{U_i} = \frac{\sum_{t=1}^{T} P(O, p_t = v, q_t = i / \lambda) \times ppv_K(o_t, v)}{\sum_{t=1}^{T} P(O, q_t = i / \lambda) \times ppv_K(o_t, v)} \quad (3\text{-}29)$$

Une solution en $\hat{u}_i$ à l'équation (3-29) sous les conditions (3-26) est donnée par :

$$\hat{u}_i(v) = \frac{\sum_{t=1}^{T} P(O, p_t = v, q_t = i / \lambda) \times ppv_K(o_t, v)}{\sum_{t=1}^{T} P(O, p_t = v / \lambda) \times ppv_K(o_t, v)} \quad (3\text{-}30)$$

sous l'hypothèse que le dénominateur de $\hat{u}_i$ varie peu en fonction de v. Cette hypothèse revient à considérer uniforme la probabilité qu'un point de l'ensemble de référence fasse partie des ppv d'un vecteur d'un segment donné. On vérifie alors que l'équation (3-30) conduit à :

$$U_i = \sum_{v'=1}^{V} \sum_{t=1}^{T} P(O, q_t = i / p_t = v', \lambda) \times ppv_K(o_t, v') \quad (3\text{-}31)$$

Et :

$$\frac{\hat{u}_i(v)}{U_i} = \frac{\sum_{t=1}^{T} P(O, p_t = v, q_t = i / \lambda) \times ppv_K(o_t, v)}{\sum_{t=1}^{T} P(O, p_t = v / \lambda) \times ppv_K(o_t, v) \times \sum_{v'=1}^{V} \sum_{t=1}^{T} P(O, q_t = i / p_t = v', \lambda) \times ppv_K(o_t, v')}$$

(3-32)

L'hypothèse de faible variation du dénominateur de $\hat{u}_i$ permet de réduire le dénominateur de (3-32) par factorisation du membre gauche et de retrouver l'équation (3-29).

Les équations (3-28) et (3-30) peuvent être évaluées à l'aide des variables directe et rétrograde (voir §1.1.3.1) :

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) \times a_{ij} \times b_j(o_{t+1}) \times \beta_{t+1}(j)}{\sum_{t=1}^{T} \alpha_t(i) \times \beta_t(i)} \quad (3\text{-}33)$$

$$\hat{u}_j(v) = \frac{\sum_{t=1}^{T-1}\sum_{i=1}^{N}(\alpha_t(i)\times a_{ij})\times u_j(v)\times ppv_K(o_{t+1},v)\times \beta_{t+1}(j)}{\sum_{t=1}^{T-1}\sum_{i,j=1}^{N}(\alpha_t(i)\times a_{ij})\times u_j(v)\times ppv_K(o_{t+1},v)\times \beta_{t+1}(j)} \tag{3-34}$$

Comme pour la réestimation locale, les formules de réestimation propagée peuvent être généralisées à un ensemble de *Ex* exemples par une sommation séparée des numérateur et dénominateur :

$$\hat{a}_{ij} = \frac{\sum_{e=1}^{Ex}\sum_{t=1}^{T_e-1}\alpha_t^e(i)\times a_{ij}\times b_j(o_{t+1}^e)\times \beta_{t+1}^e(j)}{\sum_{e=1}^{Ex}\sum_{t=1}^{T_e}\alpha_t^e(i)\times \beta_t^e(i)} \tag{3-35}$$

$$\hat{u}_j(v) = \frac{\sum_{e=1}^{Ex}\sum_{t=1}^{T_e-1}\sum_{i=1}^{N}(\alpha_t^e(i)\times a_{ij})\times u_j(v)\times ppv_K(o_{t+1}^e,v)\times \beta_{t+1}^e(j)}{\sum_{e=1}^{Ex}\sum_{t=1}^{T_e-1}\sum_{i,j=1}^{N}(\alpha_t^e(i)\times a_{ij})\times u_j(v)\times ppv_K(o_{t+1}^e,v)\times \beta_{t+1}^e(j)} \tag{3-36}$$

Bien que plus complexe et coûteuse, la formule de réestimation propagée présente l'avantage sur la réestimation locale d'être conforme au critère du MLE. La convergence de l'algorithme de Baum-Welch utilisé avec cette formule est donc montrée, au moins vers des maxima locaux de la fonction auxiliaire.

De plus, contrairement à la réestimation locale, la réestimation propagée permet de séparer les ensembles de référence et d'apprentissage. L'ensemble de référence regroupe les vecteurs parmi lesquels seront recherchés les plus proches voisins. L'ensemble d'apprentissage fournira des exemples de séquences acoustiques permettant l'estimation des paramètres des HMM. En effet, la réestimation ne porte pas sur les vecteurs de la séquence d'observations utilisée mais sur leurs ppv : l'alignement d'une séquence permet de propager la réestimation sur tous les vecteurs de référence qui sont intervenus comme ppv. Ce dernier point est un avantage de la réestimation propagée car il permet d'envisager des traitements spécifiques sur l'ensemble de référence (par exemple l'édition ou la condensation, voir §2.3.5) qui peuvent conduire à le modifier tout en conservant inchangé l'ensemble utilisé pour l'apprentissage des modèles.

3.1.3. Evaluations des réestimations locale et propagée

Deux systèmes HMM/50-ppv sont appris à partir des deux formules de réestimation : locale et propagée. L'évaluation d'un système complet nécessite l'évaluation de ses différents composants afin de tenter de séparer leur influence sur les performances globales.

L'estimateur des 50-ppv a déjà été évalué sur une tâche d'identification locale (voir §2.3.3). Nous désirons maintenant juger de l'interaction entre l'estimateur de probabilité et la modélisation markovienne. Deux niveaux d'appréciation peuvent être distingués : l'alignement dynamique dans les HMM qui sera évalué par une identification segmentale et sa combinaison avec un alignement dynamique entre HMM évalué par un DAP. Par ailleurs, les convergences des systèmes en fonction de la formule de réestimation sont étudiées. On se référera à [Lefèvre, Montacié *et al.*, 1999] pour les mêmes expériences réalisées sur la base TIMIT complète (avec les silences). Auparavant, le protocole expérimental est précisé.

3.1.3.1. Protocole expérimental

Les conditions expérimentales pour les expériences segmentales sont rappelées. Sauf mention contraire, elles resteront identiques tout au long de cette étude.

Le signal acoustique est représenté par 12 MFCC et l'énergie à court-terme calculés sur une fenêtre de 25 ms toutes les centi-secondes. Les modèles de Markov possèdent une topologie gauche-droite à 3 états avec un saut de l'état central et un bouclage sur l'état courant. L'estimateur des 50-ppv sera considéré pour le calcul des probabilités d'émissions des états. Chacune des 39 classes phonétiques (Annexe E) est associée à un modèle (monophone). L'étiquetage de référence de TIMIT est utilisé.

Un certain nombre d'options relatives à la mise au point du système HMM/K-ppv vont être discutés. Après chaque évaluation, la meilleure alternative sera retenue dans le système. Avant cela, l'apprentissage par défaut sera obtenu par la combinaison suivante : initialisation uniforme des coefficients d'appartenance, apprentissage segmental par l'algorithme de Baum-Welch et élimination lors de l'apprentissage des trames de probabilités nulles. L'alternative concernant les formules de réestimation est évaluée immédiatement.

3.1.3.2. *Convergence de l'apprentissage*

Le critère retenu pour mesurer l'influence sur le processus d'apprentissage des formules de réestimation est l'augmentation de la vraisemblance moyenne des modèles. Deux raisons nous empêchent toutefois de l'utiliser directement : la variation de la longueur des segments d'apprentissage et la variation du nombre de segments pris en compte au cours des itérations (ce problème est particulier à l'estimateur des K-ppv et sera discuté dans §3.2.2). La vraisemblance moyenne par trame offre un critère synthétique qui prend explicitement en compte les deux points précédents. La

Figure 3-1 montre l'évolution de ce critère pour les deux formules de réestimation en fonction du nombre d'itérations.

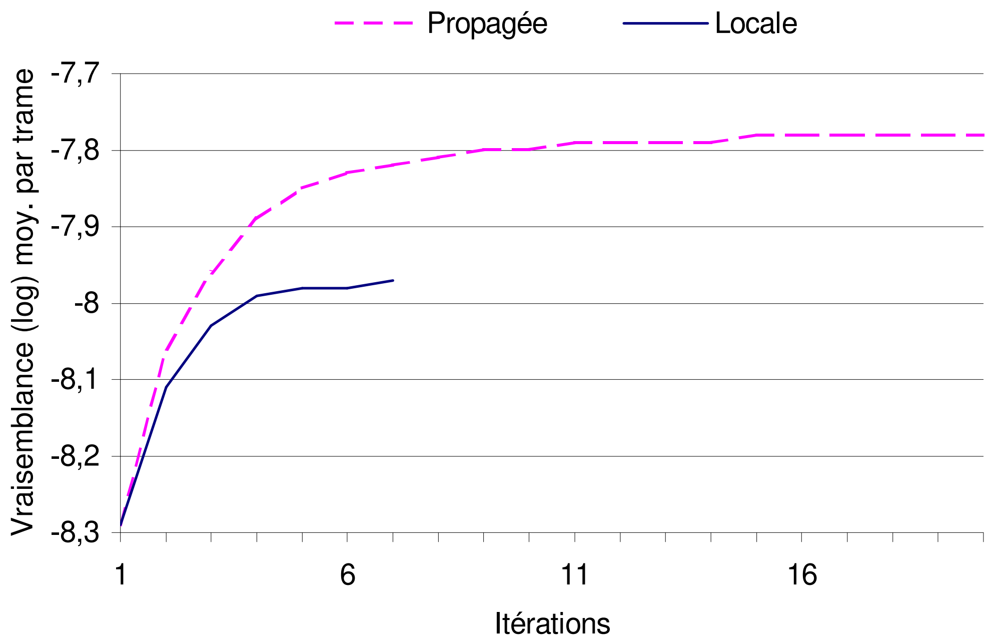


Figure 3-1 Vraisemblances (log) moyennes par trames en fonction du nombre d'itérations pour les réestimations propagée et locale.

La preuve de la convergence de la réestimation locale n'a pas été apportée. Effectivement, lors de l'apprentissage du système correspondant, de nombreux modèles n'ont pas convergé. Le comportement typique de ces modèles est de s'approcher rapidement d'un maximum local proche des paramètres initiaux puis de diverger ensuite sans rencontrer de nouveau maximum (malgré quelques tentatives poursuivies au-delà du millier d'itérations). Afin de pallier ce problème, la procédure d'apprentissage est arrêtée lorsque que la variation de la vraisemblance moyenne par trame devient négative. Malgré cette correction « manuelle » à l'absence de convergence de la réestimation locale, sa vraisemblance moyenne par trame reste inférieure à celle de la réestimation propagée. La stagnation de la vraisemblance moyenne de la réestimation locale au bout de 7 itérations est due à l'arrêt de l'apprentissage de l'ensemble des 39 modèles. Dans le cas de la réestimation propagée, il apparaît qu'une quinzaine d'itérations suffit pour converger.

3.1.3.3. Reconnaissances segmentales phonétiques

Nos expériences ont été limitées à la reconnaissance phonétique car elle est simple à mettre en œuvre. Notre participation à une campagne d'évaluation de dictée vocale (voir Annexe D) nous a permis de mesurer la difficulté de la mise au point et du test d'un système de reconnaissance lexical. Hormis le temps de calcul très important, la difficulté à dégager clairement l'influence des différents composants du système justifie notre recours aux évaluations par reconnaissance phonétique pour mesurer l'apport de l'estimateur des K-ppv à la modélisation phonétique par HMM.

Afin de mesurer l'influence des formules de réestimation sur les résultats de reconnaissances du système HMM/K-ppv, deux expériences de reconnaissances phonétiques sont pratiquées : l'identification segmentale et le DAP. La comparaison des résultats avec le système de l'état de l'art HMM/GMM sera faite ultérieurement dans le §3.3.

Identification segmentale

L'identification segmentale se distingue du décodage acoustico-phonétique par la connaissance des frontières entre les phonèmes. Cette expérience permet une évaluation des estimateurs de fdp introduits dans les HMM sans influence de la segmentation entre modèles lors de la reconnaissance. Les segments de test sont alignés sur chacun des modèles à l'aide de l'algorithme de Viterbi. La classe retenue correspond au modèle de plus grande vraisemblance.

Trois systèmes ont été considérés en fonction de la formule de réestimation utilisée lors de leur apprentissage : aucune (seule l'initialisation uniforme est réalisée), locale et propagée. Les résultats de l'identification segmentale, présentés dans le Tableau 3-1, confirme l'intérêt de la réestimation propagée avec un gain en performance de l'ordre de 2% sur l'apprentissage et le test alors que la réestimation locale conduit à une dégradation de performances.

Réestimation	Taux d'identification	
	Appr	Test
Aucune	60,9	50,9
Locale	55,5	47,5
Propagée	63,2	**52,9**

Tableau 3-1 Taux d'identification segmentale d'un système HMM/50-PPV en fonction des formules de réestimation.

Après l'introduction de l'estimateur des K-ppv dans le contexte des HMM, une expérience est réalisée afin de nous convaincre de l'intérêt de cette démarche. Son but est de vérifier que des méthodes plus simples ne permettent pas d'utiliser aussi bien les propriétés de l'estimateur pour la reconnaissance segmentale. Pour cette expérience, la classe phonétique possédant le plus grand nombre de ppv est retenue. Le nombre de ppv est calculé sur l'ensemble du segment (soit à considérer l'estimateur des Tx50-ppv, T étant la longueur du segment considéré).

Modélisation	Taux d'identification	
	Appr	Test
Aucune	58,2	51,5
HMM	63,2	52,9

Tableau 3-2 Taux d'identification segmentale avec et sans modélisation markovienne.

Les résultats de l'expérience montrent que la modélisation markovienne permet d'obtenir de meilleurs taux d'identification qu'une utilisation intégralement non-paramétrique de l'estimateur des K-ppv (Tableau 3-2). Ce résultat est satisfaisant dans la mesure où le problème de la segmentation n'était pas pris en compte dans la méthode simple, alors qu'il est dans le cadre de la modélisation markovienne.

DAP

L'intérêt du DAP a beaucoup diminué depuis que l'avènement des systèmes markoviens ne fait plus de celui-ci une étape préalable à la reconnaissance lexicale. Toutefois, cette dernière reste étroitement liée à la qualité de la modélisation acoustique, et une amélioration de la reconnaissance phonétique se traduit généralement par une augmentation des performances de reconnaissance lexicale [Gauvain, Lamel *et al.*, 1994].

Les taux de DAP (voir l'Annexe B pour la signification des différents taux) sont donnés dans le Tableau 3-3 pour l'évaluation des systèmes HMM/50-ppv sur les ensembles d'apprentissage et de test de TIMIT en fonction du type de réestimation.

Réestimation	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Aucune	61,8	19,3	18,9	1,3	60,5	1,5
Locale	59,3	23,2	17,5	3,3	56,0	2,0
Propagée	64,4	19,3	16,3	1,6	62,8	1,5
Test						
Aucune	52,1	27,1	20,8	2,6	49,6	
Locale	51,4	29,3	19,3	4,5	46,9	
Propagée	54,6	27,1	18,2	3,2	**51,4**	

Tableau 3-3 Taux de DAP du système HMM/50-ppv en fonction du type de réestimation.

La réestimation propagée donne de meilleurs résultats aussi bien sur l'apprentissage que sur le test (cf. Tableau 3-3). Sur les deux ensembles d'évaluations, la réestimation locale dégrade les taux de reconnaissance. Ces évaluations confirment nos hypothèses concernant les deux types de réestimation. La réestimation locale, plus simple à mettre en œuvre et plus rapide, ne permet pas de bons résultats faute d'une réelle convergence de l'apprentissage. La réestimation propagée conduit à une augmentation des performances du système.

Dans la suite de notre étude, nous considérerons la réestimation propagée par défaut. Le choix de la formule de réestimation étant fait, une étude de l'apprentissage va nous permettre de fixer les autres options.

3.2. Etude de l'apprentissage du système HMM/K-ppv

Dans la section précédente, un protocole minimal pour le développement d'un système HMM/K-ppv a été obtenu. Il est fondé sur le protocole décrit dans §3.1.3.1 auquel s'ajoute le choix de la réestimation propagée pour l'apprentissage des coefficients d'appartenance.

L'existence d'une procédure d'apprentissage convergente permet l'étude des variantes de mise en œuvre du système. Deux variantes liées aux propriétés de l'estimateur des K-ppv sont d'abord abordées : le choix de la probabilité conditionnelle et la gestion des probabilités nulles. Puis, le choix d'une procédure d'initialisation des coefficients d'appartenance et l'opportunité de la réestimation connectée sont étudiés.

3.2.1. Choix de la probabilité conditionnelle

Comme nous l'avons signalé dans §3.1.1.1, l'estimateur des K-ppv permet une formulation aussi simple de la vraisemblance et de la probabilité *a posteriori* d'un point sur une classe. L'emploi des probabilités *a posteriori* conduit à la formulation de nouveaux modèles mais peut aussi être envisagé dans les HMM conventionnels.

3.2.1.1. Probabilités a posteriori dans les HMM

Les probabilités d'émissions des états des HMM sont généralement calculées comme des vraisemblances du vecteur sur l'état concerné. Toutefois, le formalisme markovien n'impose pas l'utilisation d'une vraisemblance : les probabilités d'émission d'états peuvent être estimées par des probabilités *a posteriori* [Bourlard & Wellekens, 1990].

Dans le cas de l'utilisation de modèles de langage, les probabilités linguistiques *a priori* sont factorisées dans la structure de décodage et donc l'emploi de probabilités *a posteriori* est incohérent [Renals, Morgan *et al.*, 1992]. En fait, le recours aux probabilités *a posteriori* entraîne l'utilisation de modèles particuliers que sont les modèles de transitions [Konig, 1996]. Ces modèles sont différents des HMM dans la mesure ou ils ne sont plus des modèles de production mais des modèles de reconnaissance. La mise en œuvre de tels modèles n'a, pour l'instant, été étudiée que dans le cadre des réseaux de neurones. Il n'est pas encore possible de dégager clairement leurs avantages réels.

Par ailleurs, le calcul des probabilités *a posteriori* a été introduit dans les HMM afin d'accélérer le décodage [Renals & Hochberg, 1995]. Dans ce cas, les probabilités *a posteriori* fournissent une estimation locale de l'adéquation entre un modèle et les observations qui permet de désactiver certains phonèmes lors du décodage. Un réseau de neurones de type MLP permet d'estimer la probabilité *a posteriori* des phonèmes pour chaque trame à décoder. Lors du décodage, les phonèmes les plus improbables peuvent donc être identifiés préalablement et la prolongation des chemins par ces phonèmes est évitée. L'intégration de cette technique dans le décodeur NOWAY a permis une amélioration d'un facteur 10 du temps de décodage d'une phrase sur une tâche de reconnaissance avec un vocabulaire de 20 k-mots [Renals & Hochberg, 1995].

3.2.1.2. Probabilités a posteriori dans le système HMM/K-ppv

Les deux types de probabilités conditionnelles (vraisemblance et *a posteriori*) sont expérimentés sur un DAP dans le système HMM/50-ppv.

Probabilité	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Vraisemblance	64,4	19,3	16,3	1,6	62,8	1,5
A posteriori	60,4	21,3	18,3	2,3	58,1	1,0
Test						
Vraisemblance	54,6	27,1	18,2	3,2	51,4	
A posteriori	53,5	26,7	19,8	3,2	50,3	

Tableau 3-4 Taux de DAP du système HMM/50-ppv en fonction du type de probabilité conditionnelle.

Les résultats montrent que l'influence du type de probabilité n'est pas déterminante sur le test (Tableau 3-4). L'écart de 1,1% entre les taux de DAP avec vraisemblance et probabilité *a posteriori* représente, en effet, le rayon de l'intervalle de confiance pour cette expérience (cf. Annexe B). Il semble toutefois que la vraisemblance soit plus conforme avec la méthode d'apprentissage du système puisqu'on relève un écart de 6% en auto-cohérence.

La rapidité du décodage n'étant pas une préoccupation de notre étude, l'utilisation des probabilités *a posteriori* pour désactiver les phonèmes les moins probables n'a pas été réalisée. Pour la suite de l'étude, les vraisemblances seront utilisées par défaut.

3.2.2. Gestion des probabilités nulles

Dans le cas des estimateurs continus, les probabilités d'émissions étant estimées directement par les valeurs des densités de probabilités peuvent devenir très faibles mais ne sont jamais strictement nulles. Un problème particulier est posé par l'utilisation de l'estimateur des K-ppv dans les HMM : les vecteurs ayant des probabilités d'émission nulles sur tous les états de leur modèle.

Ce point est problématique lors de l'apprentissage par l'algorithme de Baum-Welch, car la vraisemblance du segment considéré sur son HMM est alors nulle. Les trames concernées correspondent manifestement à un problème d'étiquetage, d'alignement ou d'analyse. Un certain nombre de propositions vise à remédier à ce problème :

- affecter une *probabilité minimum* très faible : cette est inspirée des fdp continues qui parviennent toujours à affecter une probabilité d'émission même très faible à une trame. Le problème posé est la détermination de la valeur.
- *éliminer* la trame : on opère alors une compression du segment. L'inconvénient réside dans une modification de la structure temporelle du segment.
- affecter la probabilité de la *trame précédente* : cette solution permet de conserver la même durée au segment. La trame est remplacée par la trame précédente. L'hypothèse est que la mauvaise trame peut-être étiquetée comme la précédente. Or, les trames problématiques correspondent généralement à des zones de fortes instabilités spectrales qui dénotent des transitions phonétiques.
- affecter une *probabilité moyenne* entre les deux trames adjacentes : cette proposition constitue un raffinement de la solution précédente.

L'évaluation systématique de chacune des propositions précédentes n'a pas été menée. Les expériences préliminaires nous avaient conduits à préférer l'élimination des trames et à constater de faibles variations de performance entre les différentes techniques. Cette technique sera retenue par défaut lors de l'apprentissage des systèmes HMM/K-ppv.

Il peut arriver, malgré l'application des techniques précédentes, que la vraisemblance d'un segment soit nulle. Ce problème naît de l'incohérence entre chemins : si une trame a au moins un état avec une probabilité non nulle, elle est valide et n'est donc pas éliminée. Néanmoins, cela ne nous assure pas que la trame suivante n'aura pas des probabilités nulles sur les seuls états qui pourrait prolonger le chemin. Dans ce cas, le segment ne sert pas à l'apprentissage. Toutefois, il n'est pas écarté de l'ensemble d'apprentissage. Il est donc possible que, lors d'une itération ultérieure, la modification des paramètres du système lui permette d'obtenir une vraisemblance non nulle. Ce phénomène peut occasionner une fluctuation du nombre d'exemples pris en compte lors des itérations successives de l'algorithme d'apprentissage. Cet effet a été observé uniquement dans le cas de la réestimation locale.

3.2.3. Initialisation des paramètres

Dans le cadre des méthodes d'optimisation de type EM, l'initialisation des paramètres revêt une grande importance. Ces méthodes sont sensibles à la présence de maxima locaux de la fonction auxiliaire. Le risque est donc de situer le point d'initialisation proche d'un maximum local.

Deux niveaux d'initialisation sont à prévoir : la topologie du modèle (nombre d'états et matrice de transition initiale) et la répartition des données sur cette topologie.

3.2.3.1. Inférence de topologie

L'apprentissage auto-organisateur constitue l'un des attraits principaux des HMM. Mais la qualité de cet apprentissage est étroitement liée à la représentation des données à travers la topologie des modèles. Le recours à une même topologie pour des phonèmes très différents semble peu adapté ; ainsi, généralement, voyelles et consonnes sont apprises à partir de la même topologie malgré des durées et des caractéristiques spectrales et temporelles très différentes. La topologie, fixée *a priori*, des HMM est un des éléments qui compromet le plus leur qualité.

La topologie la plus simple est un modèle gauche-droit à trois états émetteurs. Ce modèle proposé par Bakis [Bakis, 1976] est le plus utilisé dans les systèmes de reconnaissance de l'état de l'art avec des variations sur le nombre d'états (de 1 à 5 pour la modélisation de phonèmes). L'absence de méthodes automatiques d'inférence de topologie ne permet pas de remettre en cause efficacement le choix d'une telle topologie, de même que l'emploi d'une topologie unique pour tous les modèles. Des expériences consistant à choisir pour chacun des phonèmes la topologie maximisant la probabilité d'émission sur l'ensemble d'apprentissage ont été menées dans le cas des systèmes HMM/GMM [Barras, 1996]. Les topologies gauche-droites ne diffèrent que par leur nombre d'états (de 2 à 4). Aucune amélioration significative des résultats n'a pu être obtenue.

L'approche que nous proposons est l'extension des techniques d'apprentissage des HMM à l'inférence automatique des topologies des HMM à partir des observations. Des études sur l'inférence de topologie ont déjà été menées [Casacuberta, Vidal *et al.*, 1990; Jouvet, Mauuary *et al.*, 1991 ; Takami & Sagayama, 1992; de Mori, Galler *et al.*, 1995 ; Takara, Iha *et al.*, 1998]. Des expériences préliminaires d'inférence de topologie à partir de clusters de vecteurs spectraux nous ont conduit à écarter momentanément les approches intégralement non-supervisée au profit d'une approche *a posteriori* [Belluardo, 1997].

Le paradigme de l'inférence de topologie *a posteriori* est l'apprentissage d'un modèle ergodique, complètement connecté. Les probabilités de transitions d'états inutiles sont presque annulées. Après élagage des parties inutilisées, un tel modèle présente une structure optimale par rapport à sa complexité. En pratique, les données disponibles limitent la complexité des structures des modèles que l'on peut effectivement apprendre.

En tenant compte de cette limitation, nous proposons une structure initiale (Figure 3-2), compromis entre complexité et possibilité d'apprentissage. Ce modèle peut être vu comme une version parallèle d'un modèle de Bakis trois-états.

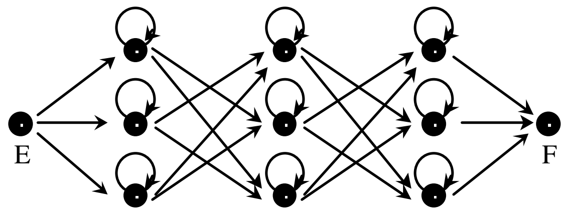


Figure 3-2 Un modèle de Bakis trois-états parallèle.

L'initialisation du modèle est obtenue à partir de trois clusters spatiaux des vecteurs d'observation des séquences d'apprentissage du modèle. Une rangée d'états est attribuée à chaque cluster. Chaque vecteur est assigné aux états de la rangée de son cluster avec des probabilités d'occupation d'état équiprobables (1/3 sur nos exemples).

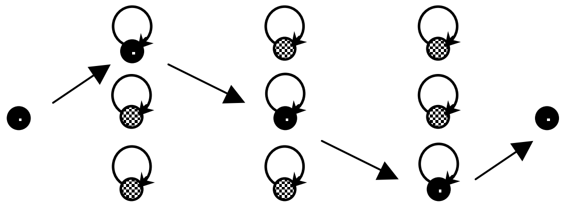


Figure 3-3 Topologie du modèle de la voyelle /ay/ après réestimation.

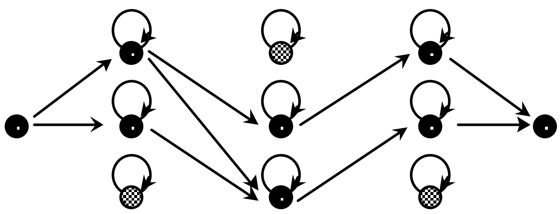


Figure 3-4 Modèle de la consonne /dh/ après réestimation.

Les Figure 3-3 et 3-4 montrent des exemples de topologies obtenues sur la base TIMIT avec l'estimateur des 50-PPV par la méthode d'inférence de topologie *a posteriori*. Pour la voyelle /ay/, la topologie obtenue est celle d'un modèle gauche-droit, tandis que l'inférence pour la consonne /dh/ donne un modèle plus complexe qui peut être vu comme deux modèles gauche-droits parallèles.

Topologie	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
G-D 3 états	64,4	19,3	16,3	1,6	62,8	1,5
9 états	61,2	20,0	18,8	2,4	58,8	1,5
Test						
G-D 3 états	54,6	27,1	18,2	3,2	51,4	
9 états	49,2	29,6	21,2	4,6	44,6	

Tableau 3-5 Taux de DAP du système HMM/50-ppv en fonction de la topologie des HMM : gauche-droite à 3 états ou inférée à 9 états.

Une expérience de DAP a été réalisée avec un système HMM/50-ppv utilisant les topologies inferées. Les résultats du DAP montrent que ces modèles dégradent considérablement les performances du système (Tableau 3-5). Il est vraisemblable que la contrainte imposée par la topologie initiale soit encore trop forte. Le recours à des méthodes d'inférence grammaticale plus élaborées (voir, par exemple, [Miclet, 1990; Dupont & Miclet, 1998]) doit être envisagé. Dans la suite de nos travaux, la topologie utilisée restera un modèle gauche-droit à 3 états avec un saut de l'état central (voir Figure 1-2).

3.2.3.2. Initialisation des coefficients d'appartenance

Trois procédures ont été envisagées pour initialisation des coefficients d'appartenance :

- *la répartition équiprobable*

Cette méthode repartit uniformément chaque trame sur l'ensemble des états du HMM :

$$u_i(o_t) = \frac{1}{N} \quad \forall(i,t) \tag{3-37}$$

Aucune information temporelle n'est introduite dans le système, mais l'equiprobabilité entre état est strictement respectée.

- *la segmentation uniforme*

Cette méthode d'initialisation est la plus couramment utilisée dans les systèmes HMM/GMM. Elle consiste à répartir uniformément les trames d'un segment sur le nombre d'états du HMM. Les séquences sont segmentées en fonction du nombre d'états du modèle, les vecteurs des sous-séquences ainsi obtenus sont associés successivement aux états du modèle :

$$u_i(o_t) = 1 \; si \; t > i \times \frac{T}{N} \tag{3-38}$$

Cette méthode a de bonnes propriétés : elle respecte la dimension temporelle et est proche de l'equiprobabilité entre états.

- *la projection temporelle*

Cette méthode originale vise à introduire, avant la réestimation, une information temporelle plus grande que celle obtenue par la segmentation uniforme. Pour cela, l'information apportée par l'estimateur des K-ppv va être utilisée préalablement aux itérations d'apprentissage.

Les paramètres sont initialisés par une segmentation uniforme des séquences d'observations. Puis, une projection du temps vers les états est réalisée : chaque vecteur est associé à chaque état proportionnellement au nombre de ses K-ppv dans cet état (d'après la segmentation uniforme). Enfin, les séquences de vecteurs sont divisées en clusters temporels proportionnels. Les projections temps/états sont moyennées sur l'ensemble des vecteurs de chaque cluster temporel. Ces projections temporelles moyennes sont utilisées comme valeurs initiales des coefficients d'appartenance.

La Figure 3-5 illustre la projection temporelle dans un modèle gauche-droit 3-états pour la voyelle /uw/ avec l'estimateur des 50-ppv. Par exemple, les coefficients d'appartenance des observations appartenant au cinquième intervalle (*i.e.*, compris entre 40 et 50% de la durée de leur séquence) sont 0,26 pour le premier état, 0,46 pour le second et 0,27 pour le dernier. Les projections temporelles pour l'ensemble des classes phonétiques confirment la relation existant entre les proximités spatiales et temporelles et ainsi l'utilisation, en première approche, d'un modèle de type gauche-droit.

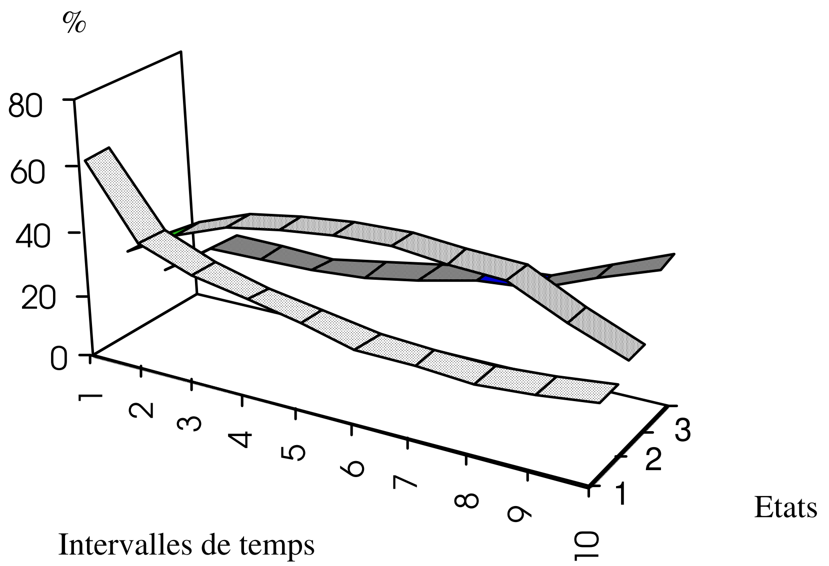


Figure 3-5 Projection temporelle pour la voyelle /uw/.

Les trois procédures d'initialisation sont comparées sur une expérience de DAP dont les résultats sont présentés dans le Tableau 3-6. Les initialisations équiprobable et par projection temporelle conduisent à des résultats identiques. L'initialisation équiprobable sera donc retenue pour le système de base dès lors que, pour des résultats comparables, elle est beaucoup plus simple à mettre en œuvre que l'initialisation par projection temporelle.

Initialisation	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Equiprobable	70,3	17,0	12,7	2,6	67,7	1,5
Uniforme	64,4	19,3	16,3	1,6	62,8	1,5
Projection Temps/Etats	70,3	17,0	12,7	2,6	67,7	1,5
Test						
Equiprobable	57,6	27,8	14,6	5,0	52,5	
Uniforme	54,6	27,1	18,2	3,2	51,4	
Projection Temps/Etats	57,5	27,5	15,1	5,0	52,5	

Tableau 3-6 Taux de DAP du système HMM/50-ppv en fonction de la procédure d'initialisation.

Jusqu'à présent, nous nous sommes uniquement intéressés à la réestimation séparée des modèles dans le système HMM/K-ppv. Il est aussi envisageable de réaliser une réestimation connectée.

3.2.4. Réestimation connectée

La réestimation simple suppose que l'on segmente la base d'apprentissage afin d'apprendre chaque modèle avec les segments acoustiques de sa classe phonétique. Dans le cas de la réestimation connectée, la base d'apprentissage n'est pas segmentée et les HMM sont appris simultanément sur toutes les phrases d'apprentissage (voir §1.3.1). Ainsi, la réestimation connectée peut autoriser certaines trames à avoir des coefficients d'appartenance non nuls sur d'autres modèles que celui de sa classe.

Bien que plus complexe à mettre en œuvre, la réestimation connectée utilise les mêmes formules que la réestimation simple. L'unique modification consiste en la prise en compte de l'état final du modèle précédent lors de la propagation des chemins dans un modèle (voir, par exemple, [Young, Jansen *et al.*, 1995] pour le détail des formules). Cette procédure étant coûteuse, on ne procède qu'à quelques itérations (classiquement de 2 à 5). Dans le cas du système HMM/50-ppv, les performances en DAP baissent dés la première itération (Tableau 3-7) et semble ne plus évoluer ensuite.

Réestimation	Phonème Identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Propagée	70,3	17,0	12,7	2,6	67,7	1,5
+ 1 itération	66,6	19,1	14,4	2,7	63,9	1,5
+ 2 itérations	66,2	18,3	15,3	2,5	63,8	1,5
Test						
Propagée	57,6	27,8	14,6	5,0	52,5	
+ 1 itération	56,5	27,7	15,8	4,6	51,9	
+ 2 itérations	56,4	27,6	16,0	4,5	51,9	

Tableau 3-7 Taux de DAP du système HMM/50-ppv en fonction du nombre d'itérations de la réestimation connectée.

La procédure de réestimation connectée ne sera donc pas appliqué à l'apprentissage des systèmes HMM/50-ppv. Un des intérêts de cette méthode d'apprentissage est de permettre une remise en cause des frontières phonétiques durant l'apprentissage. Les méthodes d'adaptation des références devraient nous permettre d'obtenir le même effet avant l'apprentissage.

3.2.5. Adaptation des références

Dans le paragraphe §2.3.4, plusieurs tentatives ont été présentées ayant pour but l'adaptation de l'ensemble de référence à l'estimateur des K-ppv : la correction des frontières phonétiques et l'édition/condensation de l'ensemble de référence. Elles sont reprises dans le cadre du système HMM/K-ppv.

3.2.5.1. Corrections des frontières phonétiques

L'ajustement des frontières entre segments phonétiques a été montré profitable dans le cadre de l'identification locale par la règle des K-ppv (cf. §2.3.4.2). L'étiquetage de l'ensemble de référence de TIMIT obtenu après re-segmentation a été utilisé pour l'apprentissage du système HMM/50-ppv. Aucun gain n'est constaté par rapport au système appris sur l'ensemble de référence non corrigé (Tableau 3-8).

Ensemble de référence	Phonème identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Initial	70,3	17,0	12,7	2,6	67,7	1,5
Re-segmenté	68,0	19,1	12,9	3,7	64,2	1,5
Test						
Initial	57,6	27,8	14,6	5,0	52,5	
Re-segmenté	57,7	28,1	14,2	6,4	51,2	

Tableau 3-8 Taux de DAP du système HMM/50-ppv en fonction du traitement de l'ensemble de référence.

3.2.5.2. Edition et condensation

Les méthodes d'édition et de condensation, introduites en §2.3.5, n'ont pas permis d'améliorer significativement la qualité de l'estimateur des 50-ppv sur une tâche d'identification locale par la règle des 50-ppv. Les ensembles de référence obtenus sont repris et évalués sur un DAP (Tableau 3-10). Les paramètres des modèles du système HMM/50-ppv sont alors appris à partir des ensembles de référence édités/condensés.

L'utilisation des techniques d'édition/condensation dans le cadre du système HMM/50-ppv conduit dans tous les cas à une dégradation des résultats de DAP. Dans le cas de l'édition, on note une augmentation très forte du taux d'insertion (malgré une adaptation du facteur d'échelle) pour les deux méthodes testées. Les résultats sont plus faiblement dégradés dans le cas de la condensation (-3% sur le test et l'apprentissage).

Ensemble de référence	Phonème identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Initial	70,3	17,0	12,7	2,6	67,7	1,5
Edition 1	61,0	6,4	32,6	0,6	60,5	0,5
Edition 2	63,6	19,6	16,8	2,2	61,4	1,0
Condensation	67,1	18,8	14,0	3,1	64,0	1,0
Test						
Initial	57,6	27,8	14,6	5,0	52,5	
Edition 1	57,9	34,1	8,0	**19,1**	38,8	
Edition 2	60,9	31,5	7,6	**16,4**	44,5	
Condensation	56,1	30,1	13,8	6,7	49,4	

Tableau 3-9 Taux de DAP du système HMM/50-ppv en fonction du traitement de l'ensemble de référence.

Dans la mesure où les résultats d'identification locale n'ont pas montré d'amélioration avec l'application des techniques d'édition/condensation, il était prévisible que les résultats de reconnaissance segmentale ne soit pas non plus améliorés. Toutefois, la préservation des résultats initiaux présente un intérêt en terme de coût de calcul puisque alors ils sont obtenus avec des ensembles de référence de tailles plus réduites.

Une hypothèse permettant de justifier les résultats du Tableau 3-10 est que l'utilisation des ensembles de référence comme ensembles d'apprentissage ne soit pas adaptée. Les techniques d'édition/condensation s'appliquent au niveau local et ne font intervenir aucune considération segmentale. Ainsi, après l'application de ces techniques, il est possible que les segments phonétiques amputées d'un certain nombre de leurs vecteurs ne soient plus représentatifs des structures temporelles que l'on tente de modéliser. Afin de vérifier cette hypothèse, nous avons renouvelé nos expériences en utilisant pour l'apprentissage l'ensemble complet dont les ppv sont recherchés parmi les ensembles de référence édités/condensés.

Les résultats de DAP confirme notre hypothèse dans la mesure où l'ensemble des taux est relevé (Tableau 3-10). On constate alors qu'il est possible de réduire de moitié la taille de l'ensemble de référence sans diminuer le niveau de performance du système. Cette information ne prendra son intérêt que lorsque se posera le problème de l'optimisation du coût de calcul de l'estimateur des K-ppv ce qui n'interviendra qu'après que les performances du système HMM/K-ppv ait été amenées au niveau de l'état de l'art des systèmes HMM/GMM.

Ensemble de référence	Phonème identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Initial	70,3	17,0	12,7	2,6	67,7	1,5
Edition 1	59,4	27,0	13,5	8,2	51,3	3,0
Edition 2	65,9	20,1	14,1	2,8	63,1	1,5
Condensation	69,6	18,5	12,0	3,0	66,6	1,0
Test						
Initial	57,6	27,8	14,6	5,0	52,5	
Edition 1	54,1	31,7	14,2	8,0	46,2	
Edition 2	56,9	27,5	15,6	4,1	**52,8**	
Condensation	56,0	29,9	14,1	5,8	50,2	

Tableau 3-10 Taux de DAP du système HMM/50-ppv en fonction du traitement de l'ensemble de référence (l'ensemble d'apprentissage n'étant pas modifié).

3.3. Performances comparées des systèmes HMM/K-ppv et HMM/GMM

Une fois définie la procédure d'apprentissage complète pour les systèmes HMM/K-ppv, ils peuvent être évalués et leurs performances comparées à celles d'un système HMM/GMM. Le paradigme d'évaluation décrit dans le §3.1.3 est repris. Afin de mesurer l'influence de la segmentation inhérente au processus de décodage deux expériences de reconnaissances segmentales sont considérées : l'identification segmentale et le DAP. De plus, dans le but de s'assurer de l'optimalité du protocole d'apprentissage retenu, les combinaisons entre les options étudiées dans §3.2 sont évaluées en ce qui concerne la formule de réestimation, le type d'initialisation et le type de probabilité.

3.3.1. Identification segmentale

Il est à noter que, lors de l'identification segmentale, certains segments peuvent ne pas être identifiés. Ceci est dû aux erreurs de segmentation qui font que certains segments appartenant à deux classes phonétiques ne peuvent être alignés sur un seul modèle (présence de trames de probabilités nulles annulant les vraisemblances de l'ensemble des chemins à travers le modèle). La proportion de ces segments est sensiblement la même pour les deux systèmes et se situe entre 2 et 3% du nombre total de segments aussi bien pour l'auto-cohérence que pour le test. Elle ne varie pas en fonction des options des systèmes. Lorsqu'un segment ne peut être décodé, il est comptabilisé comme une erreur.

Les résultats de l'identification segmentale sont rapportés dans le Tableau 3-11 pour le système HMM/50-ppv en fonction des combinaisons d'options d'apprentissage et dans le Tableau 3-12 pour le système HMM/GMM en fonction du nombre de gaussiennes par état. Les résultats du système HMM/50-ppv remettent en cause nos précédentes conclusions en ce qui concerne le type d'initialisation et le type de probabilité. Lors de leur évaluation séparée, les meilleures options étaient une initialisation équiprobable et l'utilisation de la vraisemblance. Le meilleur résultat est finalement donné par la combinaison de l'initialisation uniforme avec des probabilités *a posteriori* (64,1% en auto-cohérence et 56% sur le test). Il est aussi notable que l'écart de performance entre les deux types de réestimation (locale et propagée) est, en général, faible ; la configuration testée précédemment (initialisation uniforme et vraisemblance) correspondant à l'écart le plus élevé.

Réestimation	Initialisation	Probabilité	Taux d'identification	
			Appr	Test
Aucune	Equiprobable	Vraisemblance	59,2	49,1
		A posteriori	59,6	52,2
	Uniforme	Vraisemblance	60,9	50,9
		A posteriori	61,9	53,8
	Projection Temps/Etats	Vraisemblance	59,9	49,6
		A posteriori	60,4	52,5
Locale	Equiprobable	Vraisemblance	63,1	52,0
		A posteriori	60,1	52,7
	Uniforme	Vraisemblance	55,5	47,5
		A posteriori	58,9	52,4
	Projection Temps/Etats	Vraisemblance	63,3	52,2
		A posteriori	61,8	53,6
Propagée	Equiprobable	Vraisemblance	65,6	51,1
		A posteriori	65,0	52,1
	Uniforme	Vraisemblance	63,2	52,9
		A posteriori	64,1	56,0
	Projection Temps/Etats	Vraisemblance	68,9	51,1
		A posteriori	67,1	52,4

Tableau 3-11 Taux d'identification segmentale du système HMM/50-PPV en fonction des paramètres de l'apprentissage : formules de réestimation, type d'initialisation et type de probabilité.

Dans le cas du système HMM/GMM, les résultats de l'identification segmentale sont améliorés avec l'augmentation du nombre de gaussiennes ; contrairement à ce que nous avions observé en identification locale. Les meilleures performances sont obtenues pour 32 gaussiennes par état. Il est vraisemblable que les taux continuent à progresser au-delà de 32 gaussiennes. Le nombre de gaussiennes par état a été arrêté à cette valeur pour des considérations de temps de calcul lors des expériences.

Nb de gaussiennes par états	Taux d'identification	
	Appr	Test
1	40,2	37,9
8	46,9	42,6
16	50,0	43,3
32	53,9	44,5

Tableau 3-12 Taux d'identification segmentale du système HMM/GMM en fonction du nombre de gaussiennes par état.

Les systèmes HMM/GMM utilisés pour les expériences d'identification segmentale ont été appris spécifiquement : la réestimation par phrases qui suit habituellement l'augmentation du nombre de gaussiennes a été remplacée par une réestimation segmentale. On peut se convaincre de l'intérêt de cette modification en comparant les résultats obtenus par les modèles appris avec et sans réestimation connectée (Tableau 3-13).

Type de réestimation	Taux d'identification	
	Appr	Test
Segmentale	53,9	44,5
Connectée	51,4	41,9

Tableau 3-13 Taux d'identification segmentale du système HMM/GMM en fonction du type de réestimation.

Ces résultats sont cohérents avec le fait que la réestimation connectée remet en cause les frontières de l'étiquetage de référence. Or l'identification segmentale est évaluée à partir de l'étiquetage de référence. L'utilisation de la réestimation connectée doit donc être limitée aux systèmes évalués par un DAP.

3.3.2. DAP

Les expériences de DAP sont réalisées sur les mêmes configurations des systèmes que pour l'identification segmentale. Les résultats de DAP (Tableau 3-15) sont en accord avec les conclusions de l'étude de l'apprentissage. La meilleure combinaison est composée de la réestimation propagée, de l'initialisation équiprobable et de l'utilisation de la vraisemblance (67,7% de reconnaissance en auto-cohérence et 52,5% sur le test). Il est possible d'utiliser l'initialisation par projection temps/états avec des performances comparables. Toutefois, cette procédure d'initialisation est plus coûteuse. Il lui sera donc préféré l'initialisation équiprobable. Ces constations sont à nuancer sur le test pour lequel, la combinaison réestimation locale, initialisation équiprobable et vraisemblance permet d'obtenir un taux de DAP équivalent.

Lors de ces expériences, nous n'avons pas fait varier le nombre K de ppv. Cette opération (nous obligeant à renouveler l'ensemble des expériences plusieurs fois) aurait été très coûteuse. De plus, les résultats d'identification locale ont montré clairement l'amélioration apportée par l'augmentation du nombre de ppv. En ce qui concerne le système HMM/GMM, l'augmentation du nombre de gaussiennes par GMM est intégrée au processus d'apprentissage. Dans ces conditions, il est peu coûteux d'évaluer le système lors de chaque itération. Ces expériences intermédiaires présentent l'intérêt de montrer que, contrairement à l'identification locale, l'augmentation des gaussiennes conduit à une amélioration des résultats de DAP aussi bien sur l'apprentissage que sur le test (Tableau 3-14). Le système HMM/GMM voit ainsi son taux de reconnaissance augmenté de presque 7% sur le test (10% en auto-cohérence) entre 1 et 32 gaussiennes par états.

Les systèmes HMM/GMM de l'état de l'art offrent des taux de reconnaissance de l'ordre de 70% [Lamel & Gauvain, 1993; Ljolje, 1994; Robinson, 1994; Young & Woodland, 1994]. Les principales différences avec le système de base présenté dans ce paragraphe sont : une modélisation contextuelle et l'adjonction de coefficients delta dans les vecteurs de paramètres. Ces éléments seront étudiés dans le chapitre suivant.

GMM	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
1	49,4	25,2	25,4	1,6	47,9	3,0
8	55,5	24,7	19,8	2,1	53,5	3,0
16	56,6	23,9	19,5	1,9	54,7	3,0
32	59,7	21,7	18,6	1,8	57,9	3,0
Test						
1	46,8	27,1	26,1	1,4	45,5	
8	52,1	26,6	21,3	2,6	49,6	
16	53,0	25,9	21,1	2,3	50,7	
32	54,6	25,3	20,1	2,3	52,3	

Tableau 3-14 Taux de DAP du système HMM/GMM en fonction du nombre de gaussiennes par état.

Réesti-mation	Initialisation	Probabilité	Phn ident	Phn subst	Phn omis	Phn ins	Phn rec	Fact éch
Apprentissage								
Aucune	Equiprobable	Vraisemblance	61,3	19,6	19,1	1,6	59,7	1,5
		A posteriori	56,1	22,0	21,9	2,5	53,5	1,0
	Uniforme	Vraisemblance	61,8	19,3	18,9	1,3	60,5	1,5
		A posteriori	57,1	20,5	22,4	1,7	55,4	1,0
	Projection Temps/Etats	Vraisemblance	61,9	19,5	18,7	1,5	60,3	1,5
		A posteriori	56,1	21,3	22,6	2,2	53,8	1,0
Locale	Equiprobable	Vraisemblance	66,5	19,3	14,2	2,8	63,7	1,5
		A posteriori	58,1	22,5	19,4	3,1	55,1	1,0
	Uniforme	Vraisemblance	59,3	23,2	17,5	3,3	56,0	2,0
		A posteriori	57,3	23,2	19,5	3,1	54,2	1,0
	Projection Temps/Etats	Vraisemblance	66,4	19,3	14,2	2,7	63,7	1,5
		A posteriori	58,9	21,7	19,5	2,7	56,1	1,5
Propagée	Equiprobable	Vraisemblance	70,3	17,0	12,7	2,6	67,7	1,5
		A posteriori	60,1	21,0	18,9	2,6	57,5	1,0
	Uniforme	Vraisemblance	64,4	19,3	16,3	1,6	62,8	1,5
		A posteriori	60,4	21,3	18,3	2,3	58,1	1,0
	Projection Temps/Etats	Vraisemblance	70,3	17,0	12,7	2,6	67,7	1,5
		A posteriori	60,7	20,1	19,1	2,6	58,1	1,0
Test								
Aucune	Equiprobable	Vraisemblance	52,3	26,8	20,9	2,8	49,5	
		A posteriori	50,4	26,6	23,0	3,1	47,3	
	Uniforme	Vraisemblance	52,1	27,1	20,8	2,6	49,6	
		A posteriori	51,3	24,9	23,8	2,5	48,8	
	Projection Temps/Etats	Vraisemblance	52,6	27,0	20,4	3,0	49,6	
		A posteriori	50,2	25,8	24,0	2,9	47,3	
Locale	Equiprobable	Vraisemblance	57,1	27,2	15,7	4,6	52,6	
		A posteriori	52,1	27,1	20,8	3,9	48,2	
	Uniforme	Vraisemblance	51,4	29,3	19,3	4,5	46,9	
		A posteriori	51,5	27,9	20,6	3,7	47,9	
	Projection Temps/Etats	Vraisemblance	57,0	27,3	15,8	4,6	52,3	
		A posteriori	52,3	26,6	21,1	3,5	48,7	
Propagée	Equiprobable	Vraisemblance	57,6	27,8	14,6	5,0	52,5	
		A posteriori	52,7	27,0	20,3	3,8	48,9	
	Uniforme	Vraisemblance	54,6	27,1	18,2	3,2	51,4	
		A posteriori	53,5	26,7	19,8	3,2	50,3	
	Projection Temps/Etats	Vraisemblance	57,5	27,5	15,1	5,0	52,5	
		A posteriori	51,9	26,8	21,3	3,8	48,1	

Tableau 3-15 Taux de DAP du système HMM/50-ppv en fonction des paramètres de l'apprentissage : les formules de réestimation, le type d'initialisation et le type de probabilité.

3.3.3. Mise en relation des résultats

Les résultats de l'identification segmentale et du DAP des systèmes HMM/50-ppv et HMM/32-GMM sont mis en relation dans la Figure 3-6. Les résultats du système HMM/50-ppv sont donnés pour la combinaison réestimation propagée, initialisation propagée et vraisemblance bien que cette configuration ne soit pas la meilleure pour l'identification segmentale. Il est ainsi possible de comparer les résultats entre les expériences d'identification segmentale et de DAP. Notre objectif étant la RAP continue, le DAP constitue l'expérience de référence.

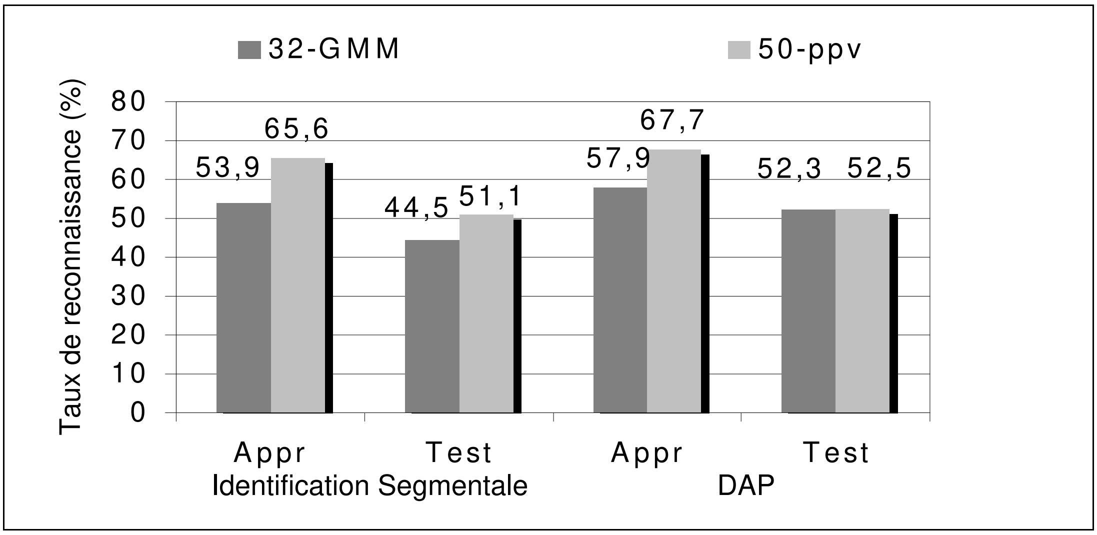


Figure 3-6 Résultats de reconnaissance phonétique des systèmes HMM/50-ppv et HMM/32-GMM.

La mise en relation des résultats permet de confirmer en reconnaissance segmentale l'écart de performance relevé en identification locale entre les estimateurs 32-GMM et 50-ppv (cf. Figure 3-6). Une augmentation de l'ordre de 10% est constatée en auto-cohérence sur les deux types d'expériences. Cet écart se réduit à +6,6% sur le test de l'identification segmentale. L'exception notable concerne le test du DAP pour lequel les deux systèmes présentent des taux comparables (52,3% pour le système HMM/32-GMM et 52,5 pour le système HMM/50-ppv).

Afin de préciser les différences de comportement entre les deux systèmes, masquées par des taux semblables, le taux de DAP du test est détaillé pour chaque étiquette phonétique (Tableau 3-16). Les étiquettes ont été ensuite regroupées par macro-classes phonétiques : voyelles, diphtongues, liquides, nasales, occlusives, fricatives et occlusions (Tableau 3-17).

Phn	Nb Phn	Taux de reconnaissance	
		50-ppv	32-GMM
Voyelles			
iy	243	60,5	51,0
ih	579	49,3	47,5
eh	189	23,3	29,1
ae	105	37,1	37,1
aa	228	51,3	42,1
ah	321	35,2	36,1
uw	74	33,8	44,6
uh	29	3,4	20,7
er	233	62,8	51,5
ey	114	52,6	49,1
ay	89	49,4	30,3
oy	16	-12,5	6,2
aw	30	13,3	13,3
ow	89	30,3	30,3
Liquides			
l	291	47,8	52,6
r	270	44,1	52,2
y	50	14,0	22,0
w	144	50,7	41,7
Occlusions			
cl	982	78,5	79,4

Phn	Nb Phn	Taux de reconnaissance	
		50-ppv	32-GMM
Nasales			
m	204	47,5	45,6
n	385	52,2	45,5
ng	52	26,9	38,5
Fricatives			
z	181	63,5	68,0
v	93	29,0	37,6
f	131	78,6	80,2
th	38	18,4	28,9
s	320	83,8	69,4
sh	77	68,8	62,3
hh	77	44,2	44,2
ch	40	50,0	50,5
jh	42	33,3	61,9
dh	128	21,1	25,0
Occlusives			
b	132	34,1	48,5
d	113	30,1	29,2
dx	90	21,1	15,6
g	66	22,7	42,4
p	136	54,4	56,6
t	187	58,3	64,2
k	165	60,6	72,1

Tableau 3-16 Taux de DAP des systèmes HMM/50-ppv et HMM/32-GMM détaillés par classes phonétiques (le meilleur est en grisé).

Macro-classe	Taux de reconnaissance	
	50-ppv	32-GMM
Voyelles	44,9	41,9
Liquides	44,8	48,3
Occlusions	78,5	79,4
Nasales	48,7	44,9
Fricatives	59,3	58,2
Occlusives	44,5	51,2

Tableau 3-17 Taux de DAP des systèmes HMM/50-ppv et HMM/32-GMM détaillés par macro-classes phonétiques (le meilleur est en grisé).

Si l'on excepte les occlusives, les taux de reconnaissance en macro-classes entre les deux systèmes montrent de faibles divergences. Cette constatation est aussi vraie lorsque l'on détaille les classes phonétiques : il n'est pas possible de corréler les taux de reconnaissance avec les macro-classes phonétiques. Les différences de taux de reconnaissance se compensent entre classes appartenant à la même macro-classe. La qualification des différences de comportement entre les deux estimateurs nécessite donc une analyse plus précise. Avant cela, pour tenter de quantifier ces différences, nous proposons de mesurer la complémentarité des estimateurs K-ppv et GMM au niveau local.

3.3.4. Complémentarité des estimateurs K-ppv et GMM

La complémentarité entre différents estimateurs peut être obtenue par construction (par exemple en les apprenant sur des données différentes) ou être constatées empiriquement. Les expériences présentées au cours de cette étude nous ont permis de constater des différences dans le comportement des estimateurs K-ppv et GMM.

L'alignement de leur décision en identification locale peut permettre d'évaluer la complémentarité des deux estimateurs. Pour chaque trame, trois types de configurations peuvent apparaître : les deux reconnaissances sont correctes, l'une seulement des reconnaissances est correcte ou les deux reconnaissances sont erronées. Le premier cas décrit l'ensemble des trames sans problème. Le second cas correspond à l'ensemble des trames qui peuvent être correctement reconnues par la combinaison idéale des deux estimateurs. Finalement, le troisième cas regroupe les trames incorrectes pour les deux estimateurs.

Cet alignement nous permet d'évaluer le gain maximal en identification qui pourrait être obtenu par la combinaison optimale des deux estimateurs. Ce faisant, nous ne sommes assurés que de l'amélioration des performances d'identification locale. La reconnaissance segmentale ne constitue pas une simple concaténation de décisions locales. Il est donc envisageable que l'amélioration de l'identification locale n'ait que peu d'influence sur le processus de reconnaissance segmental comme le montrent les résultats de DAP (cf. Figure 3-6). Les résultats de mesure de complémentarité entre les estimateurs 50-PPV et 32-GMM sont donnés dans le Tableau 3-18. Nous constatons que plus d'un tiers des trames du noyau de test ne sont reconnues par aucun des estimateurs (38,1%).

50-ppv\32-GMM	Correct	Incorrect
Correct	29,7	16,2
Incorrect	16,0	38,1

Tableau 3-18 Complémentarité des estimateurs 50-ppv et 32-GMM.

Le taux d'identification d'une combinaison idéale est donc de l'ordre de 62%, représentant un gain absolu d'environ 16% pour l'un comme pour l'autre estimateur. Il existe donc des différences de comportement entre les deux estimateurs qui peuvent être exploitées au sein d'un estimateur hybride K-ppv/GMM. Mais surtout, l'analyse plus avant de ces différences devrait fournir, par comparaison, des informations plus précises sur le fonctionnement de l'un et l'autre estimateur.

3.4. Conclusion

Dans ce chapitre, les étapes nécessaires à la construction d'un système de reconnaissance HMM/K-ppv de base ont été décrites. Un cadre formel a été développé pour obtenir un algorithme d'apprentissage de type EM dont nous avons montré la convergence selon le critère du maximum de vraisemblance. L'étude expérimentale de l'apprentissage a permis de dégager un protocole optimal. Dans ce dernier, la formule de réestimation propagée est utilisée en combinaison avec une initialisation partagée. Les autres options retenues sont l'utilisation d'une vraisemblance et l'élimination des vecteurs de probabilité nulle.

L'étude des performances du système de base HMM/50-ppv montrent sa supériorité sur un système HMM/32-GMM à une exception notable près puisqu'elle concerne le test du DAP qui constitue notre expérience de référence. Alors que ses performances sont supérieures en identification segmentale et pour l'auto-cohérence du DAP (de 6 à 10% d'écart), le système HMM/50-ppv présente un taux de reconnaissance comparable à celui du système HMM/32-GMM sur le test du DAP. Ce résultat semble montrer que l'amélioration des qualités locales de l'estimation probabiliste des trames de parole a une influence positive sur les performances d'identification du système mais pas sur ses capacités de segmentation. Cette hypothèse est difficilement vérifiable dans la mesure où le DAP par les modèles markoviens rend indissociables les étapes de segmentation et d'identification. Une analyse approfondie des résultats de reconnaissance doit être menée afin de mettre en avant d'une part les différences internes au système HMM/K-ppv (notamment entre les tâches d'identification et de reconnaissance) et d'autre part les différences de comportement entre les deux systèmes HMM/K-ppv et HMM/GMM.

Toutefois, les performances des systèmes étudiés n'étant pas celles de l'état de l'art, le risque est grand de mettre en avant par cette analyse des phénomènes dont l'influence pourrait être remise en cause par l'intégration de techniques performantes. Notre préoccupation dans l'immédiat doit donc être l'augmentation des performances du système HMM/K-ppv. Les techniques de l'état de l'art des systèmes HMM/GMM doivent nous permettre d'évoluer vers un système de référence HMM/K-ppv. L'existence du système de base constitue, en ce sens, un atout très important car il permettra d'évaluer de manière objective la pertinence des traitements proposés.

4. ADAPTATION DE L'ESPACE DE REPRESENTATION ET DES REFERENCES

L'étude menée dans le chapitre précédent a permis le développement d'un système HMM/K-ppv de base. Afin d'améliorer ses performances, des traitements particuliers sont proposés portant sur deux aspects de l'opérateur de reconnaissance : l'espace de représentation et les références. Le choix de ces traitements a été largement guidé par la volonté de faire bénéficier le système HMM/K-ppv des techniques de l'état de l'art des systèmes HMM/GMM.

Les phonèmes, définis comme l'ensemble minimal des unités sonores permettant de distinguer deux mots d'une langue, constituent des unités élémentaires de reconnaissance appropriées (bien que discutées, cf. §1.2.1) pour la parole. Ils permettent une interprétation linguistique simple des séquences acoustiques décodées. Il est toutefois possible de redéfinir les classes utilisées lors de la reconnaissance pour autant qu'un retour aux classes phonétiques puisse être opéré *in fine*. Deux approches opposées sont envisageables : la subdivision en sous-classes ou la constitution de macro-classes utilisées lors d'un décodage préalable. La première approche permet de limiter les confusions lors de la reconnaissance par une redéfinition globale des classes en sous-classes plus homogènes. La seconde propose, au contraire, par une séparation préalable en macro-classes, de spécialiser l'opérateur sur des ensembles de classes particulièrement confus. L'objectif dans ce cas est d'améliorer les performances par la prise en compte d'un moins grand nombre de classes à chaque étape.

En préalable à l'adaptation des références dans le système HMM/K-ppv, l'adaptation de l'espace de représentation est envisagée. Cette démarche est justifiée par une prise en compte de la dimension temporelle finalement peu réaliste dans les HMM du fait de l'hypothèse d'indépendance des trames. Parmi les méthodes permettant d'y remédier, l'introduction de paramètres temporels dans l'espace de représentation est la plus répandue. D'utilisation extrêmement simple, elle a permis des gains de reconnaissance importants dans les systèmes HMM/GMM. Afin de qualifier précisément l'influence des paramètres temporels dans les systèmes HMM/GMM, plusieurs hypothèses sont étudiées. Les diverses modalités (vecteurs composites ou flux de données) pour l'application de la méthode au système HMM/K-ppv sont évaluées.

4.1. Introduction de paramètres temporels dérivés dans l'espace de représentation

Une meilleure prise en compte de la dynamique temporelle du signal lors de la reconnaissance est envisagée pour le système HMM/K-ppv. Parmi les différentes approches possibles, l'introduction de paramètres temporels dérivés (coefficients delta) dans l'espace de représentation semble la plus prometteuse. Les différentes modalités possibles de cette introduction sont présentées. Une étude spécifique de l'influence des coefficients delta dans les systèmes HMM/GMM est finalement menée.

4.1.1. Représentation temporelle dans les HMM

Les coefficients delta représentent le paradigme de l'introduction de l'information temporelle au sein d'un système markovien. Depuis son introduction [Furui, 1986], cette technique a été appliquée avec succès à la majorité des tâches de reconnaissance de la parole [Lee, 1988; Rabiner, Wilpon *et al.*, 1989; Lee, Giachin *et al.*, 1992; Wilpon, Lee *et al.*, 1993]. D'autres méthodes ont été envisagées pour incorporer l'information temporelle dans les paramètres des systèmes HMM : fonctions de variations spectrales (*Spectral Variation Function*) [Brugnara, de Mori *et al.*, 1992], cepstre dynamique (*dynamic cepstrum*) [Aikawa, Singer *et al.*, 1993]… Mais aucune n'a jamais allié autant efficacité et simplicité que les coefficients delta.

4.1.1.1. Alternatives aux coefficients delta

Les premiers efforts pour introduire de l'information temporelle dans les paramètres des systèmes de RAP consistaient à concaténer des trames adjacentes [Nadàs, Mercer *et al.*, 1981]. Cette méthode entraînait deux problèmes majeurs : l'augmentation considérable du nombre de paramètres en entrée des systèmes et l'introduction de problèmes d'estimation.

Les méthodes usuelles de réductions du nombre de paramètres (ACP et ALD, §2.2.1) ne permirent pas d'augmenter les performances sur des tâches de reconnaissance de grands vocabulaires. Cette méthode fut donc abandonnée en 1988 devant le succès croissant des coefficients delta. Elle fut toutefois introduite de nouveau dans le cadre des systèmes hybrides HMM/réseaux de neurones (voir §2.2.2) qui utilisent conjointement les coefficients delta et les trames adjacentes dans des vecteurs de paramètres pouvant aller jusqu'à plusieurs centaines de coefficients (234 coefficients, par exemple, pour [Konig, 1996]), au détriment du temps de calcul.

Afin de mesurer l'apport d'une méthode permettant de ne pas introduire de nouveaux paramètres dans le système, une variante à la concaténation de trames est testée dans le système HMM/K-ppv. Pour chaque trame, les *K/2* ppv des deux trames adjacentes sont pris en compte en plus de ses propres K ppv. Une trame est donc classée par *2K* ppv dont la moitié appartient aux trames voisines.

Estimateur	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
50-ppv	70,3	17,0	12,7	2,6	67,7	1,5
100-ppv adj.	67,6	16,1	16,3	0,9	66,7	1,0
50-ppv	57,6	27,8	14,6	5,0	52,5	
100-ppv adj.	54,1	27,1	18,9	2,6	51,5	

Tableau 4-1 Taux de DAP du système HMM/50-ppv avec et sans concaténation des ppv des trames adjacentes.

Les résultats du DAP avec concaténation des ppv (Tableau 4-1) confirment les limites de cette approche. On note une réduction de moitié du taux d'insertion par rapport au système standard. Mais l'augmentation du taux d'omissions conduit finalement à une baisse de 1% des taux de reconnaissance sur l'apprentissage et sur le test. Ainsi, même s'il existe une corrélation très forte entre les classes des trames adjacentes, l'utilisation conjointe de leur ppv n'introduit pas d'amélioration dans le système. Il semble donc nécessaire d'introduire explicitement l'information temporelle dans le système, par exemple, sous la forme de nouveaux paramètres.

4.1.1.2. Motivations pour les coefficients delta

Plusieurs éléments motivent le recours aux coefficients delta dans les systèmes de RAP. Ainsi, Furui a réaffirmé (après [Ruske, 1982]) le rôle important que jouent les variations temporelles du spectre dans la perception humaine. Des expériences perceptives sur l'identification de syllabes consonne-voyelle en japonais dont le début ou la fin sont tronqués ont montré que les zones critiques du point de vue de la perception sont les zones de plus fortes transitions spectrales [Furui, 1986]. De plus, dans le cadre de la reconnaissance indépendante du locuteur, si la localisation des formants peut être très variable entre locuteurs, les experts en lecture de spectrogrammes s'accordent à dire que la forme de leur évolution temporelle est relativement cohérente pour un même locuteur. Il paraît donc intéressant d'adjoindre aux coefficients initiaux des coefficients offrant une approximation de la pente de leur évolution temporelle.

Les coefficients delta dans les systèmes HMM/GMM et HMM/K-ppv de cette étude sont calculés à l'aide de la formule de régression :

$$D_i(t) = \frac{\sum_{n=1}^{n_0} n \times \left(C_i(t+n) - C_i(t-n)\right)}{2\sum_{n=1}^{n_0} n^2} \tag{4-1}$$

où $D_i(t)$ est un coefficient delta de la trame *t*, correspondant au coefficient statique $C_i(t)$. La longueur de l'intervalle est généralement de l'ordre de 5 trames (n_0=2), représentant 50 ms de parole. Des coefficients différentiels du second ordre peuvent être aussi obtenus par une régression linéaire sur les coefficients du premier ordre. Ces coefficients ont montré un bon comportement pour de la parole bruitée ou soumise à l'effet Lombard. Dans le cas de parole de bonne qualité, la faible amélioration des résultats ne justifie pas le coût de calcul induit.

Dans nos notations, l'adjonction de coefficients différentiels aux coefficients statiques sera notée +D. Ainsi, le jeu de paramètres que nous utilisons dans nos évaluations MFCC+E devient MFCC+E+D lorsqu'il intègre des coefficients différentiels. Le sous-vecteur constitué uniquement des coefficients delta est noté MFCC+E_D.

4.1.1.3. Apport des coefficients delta

L'introduction des coefficients delta dans un système de reconnaissance donne généralement lieu a une augmentation notable des performances. L'adjonction de coefficients delta aux coefficients initiaux d'un système HMM/GMM améliore les résultats de reconnaissance sur la base TIMIT de l'ordre de 6 à 10% (cf. l'étude menée dans [Barras, 1996]). L'augmentation des performances due à l'utilisation des coefficients delta met en avant une des limitations majeures des HMM. Bien que l'information contenue dans les paramètres différentiels soit directement disponible dans le jeu de paramètres initial, il faut qu'elle soit fournie explicitement pour que les HMM en tiennent compte. Ceci est une conséquence de l'hypothèse d'indépendance qui permet de modéliser l'émission de chaque observation indépendamment des précédentes et des suivantes.

Une étude de Bocchieri et Wilpon a permis de mieux préciser l'influence réciproque entre les coefficients statiques et dynamiques [Bocchieri & Wilpon, 1993]. Leur objectif était de réduire le nombre de dimensions de l'espace de représentation des trames de parole sans diminuer la qualité de la reconnaissance. Ils ont pour cela proposé une procédure automatique de sélection de coefficients basée sur une analyse discriminante de type ALD (voir §2.2.1.2). Leurs expériences sur la base TIMIT ont montré qu'il est possible de réduire de moitié la taille de l'espace de représentation sans dégrader la reconnaissance. Un phénomène intéressant est que la procédure automatique a tendance à conserver associés les coefficients statiques et delta de même ordre. Ainsi, à partir de 38 coefficients (12 statiques, 12 delta, 12 delta-delta, delta-énergie et delta-delta-énergie), les 12 coefficients conservés comprennent les 2 énergies (ce qui n'est pas surprenant connaissant l'importance de l'énergie pour la reconnaissance [Lee & Hon, 1989]), les statiques de 1 à 6 et les delta de 3 à 6. Les coefficients delta semble donc une approche prometteuse pour l'introduction d'une information temporelle dans le système HMM/K-ppv.

4.1.2. Modalités de l'introduction des coefficients delta dans le système HMM/K-ppv

Les coefficients delta sont introduits dans le système HMM/K-ppv. Dans une première approche, les coefficients statiques et delta sont utilisés conjointement au sein d'un vecteur composite, ce qui représente la configuration habituelle dans les systèmes HMM/GMM. Dans une seconde approche, les coefficients statiques et delta sont séparés en deux flux. La probabilité d'émission d'une observation est obtenue par multiplication des probabilités de chacun des flux.

4.1.2.1. Vecteur et distance composites

Pour combiner des coefficients cepstraux et des coefficients delta-cepstraux, l'utilisation d'une métrique composite [Furui, 1986; Shikano, Lee *et al.*, 1986] peut être envisagée avec une distance définie par :

$$d_{composite} = d_E(MFCC + E) + w \times d_E(MFCC + E\_D) \qquad (4\text{-}2)$$

avec d_E la distance euclidienne et w un poids qui reflète l'importance relative des sous-vecteurs MFCC+E et MFCC+E_D. Le poids optimal proposé dans [Lee, 1988] est de 0,8. Il a été obtenu après une série d'expériences de reconnaissance phonétique.

Nos tentatives pour utiliser les coefficients delta dans un vecteur composite dans le système HMM/K-ppv n'ont pas pu être menées à terme. Pour d'obtenir un temps de calcul acceptable pour l'algorithme des K-ppv rapide, le poids des coefficients delta doit être au maximum de 0,05. Leur influence est alors quasiment annulée.

L'espace obtenu en combinant les dimensions statiques et dynamiques n'a pas de bonnes propriétés topologiques. Cet aspect n'intervient pas dans le cas des estimateurs continus, du fait notamment de l'hypothèse d'indépendance entre les dimensions. Cette analyse est confirmée par K-F Lee qui, soumis à des problèmes identiques avec l'estimateur discret, a du avoir recours à des dictionnaires de prototypes multiples [Lee, 1988]. L'intégration des coefficients delta devrait donc être facilitée par la considération de deux espaces de représentation distincts.

4.1.2.2. Flux de données

Lorsque les vecteurs d'observations contiennent des données hétérogènes (coefficients cepstraux, énergie, coefficients delta…), il est possible de découper les vecteurs d'observation en plusieurs groupes de coefficients appelés flux de données et de modéliser ces flux indépendamment Cela revient à dissocier les espaces de représentation associés aux dimensions des différents flux.

La probabilité d'émission d'un état i d'un HMM pour l'observation o_t est alors reconstruite par le produit des probabilités d'émission de chacun des flux :

$$b_i(o_t) = \prod_F \left(b_i^f(o_t)\right)^{\gamma_f} \tag{4-3}$$

où b_i^f est la probabilité d'émission du flux f parmi F sur l'état i appliquée sur le sous-vecteur correspondant de o_t et γ_f le facteur de pondération du flux f. Ce principe est applicable des estimateurs continus comme discrets. K.F. Lee, qui utilise des modèles discrets, construit trois dictionnaires de prototypes, un par type de donnée (MFCC statiques, delta et énergie) [Lee, 1988; Lee & Hon, 1989].

Cas des GMM

L'approche par flux de données n'est généralement pas retenue dans le cas d'une modélisation par GMM. Les coefficients étant jugés suffisamment indépendants, il est habituel d'utiliser des matrices de covariance diagonales. Chaque coefficient peut alors être considéré comme un flux. Toutefois, cette considération n'est vraie que dans le cas des 1-GMM.

Formellement, une gaussienne multidimensionnelle à matrice de covariance diagonale est équivalente au produit des gaussiennes monodimensionnelles de chacun de ses coefficients, ce qui peut être comparé à l'association d'un flux par dimension :

$$N(x;\mu,\Sigma_{diag}) = \prod_{i=1}^{d} N(x_i;\mu_i,\sigma_i) = \prod_{f=1}^{F} N(x_f;\mu_f,\sigma_f) \tag{4-4}$$

avec $F = d$. Si nous considérons maintenant un GMM, l'utilisation d'une matrice diagonale ne conduit plus à la même formulation que l'association exacte d'un flux à chaque dimension. Dans le cas d'un GMM à G gaussienne, l'expression de la fdp est :

$$\begin{aligned}\sum_{g=1}^{G} N(x;\mu_g,\Sigma_{diag_g}) &= \sum_{g=1}^{G}\prod_{i=1}^{d} N(x_i;\mu_{g_i},\sigma_{g_i}) \\ &= \sum_{g=1}^{G}\prod_{f=1}^{F} N(x_f;\mu_{g_f},\sigma_{g_f})\end{aligned} \tag{4-5}$$

Or, formellement une modélisation par flux des dimensions nous donne :

$$\sum_{g=1}^{G} N(x;\mu_g,\Sigma_{diag_g}) = \prod_{f=1}^{F}\sum_{g=1}^{G} N(x_f;\mu_{g_f},\Sigma_{g_f}) \tag{4-6}$$

Il y a donc inversion de la somme et du produit. Ceci à une conséquence concrète dans le système. Des expériences menées sur 2 flux (MFCC statiques et énergie, et MFCC delta) ont abouti à des scores de reconnaissance inférieurs de quelques pour-cent à ceux d'une modélisation sans flux. Ces résultats tendent à montrer que les deux espaces (statique et dynamique) doivent être modélisés simultanément avec l'estimateur GMM.

Application au système HMM/K-ppv

La modélisation par flux de données est appliquée au système HMM/K-ppv. Pour cela, les formules de réestimation, introduites dans le chapitre 3, doivent être adaptées.

Chaque vecteur de l'ensemble de référence possède deux ensembles de coefficients d'appartenance aux états : un correspondant au sous-vecteur des coefficients statiques $\{u^S_i(v)\}$ et l'autre au sous-vecteur des coefficients delta $\{u^D_i(v)\}$. Les formules de réestimation pour chaque groupe de coefficients nécessitent la prise en compte de la participation de l'autre dans le calcul de la probabilité d'émission. Pour les coefficients d'appartenance correspondant au flux des paramètres statiques, nous obtenons :

$$\hat{u}_i^S(v) = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N}(\alpha_{t-1}(j)\times a_{ji})\times\left(u_i^S(v)\right)^{\gamma_S}\times ppv_K^S(o_t,v)\times\left(b_i^D(o_t)\right)^{\gamma_D}\times\beta_t(i)}{\sum_{t=1}^{T}\sum_{j,i=1}^{N}(\alpha_{t-1}(j)\times a_{ji})\times\left(u_i^S(v)\right)^{\gamma_S}\times ppv_K^S(o_t,v)\times\left(b_i^D(o_t)\right)^{\gamma_D}\times\beta_t(i)} \quad (4\text{-}7)$$

avec

$$b_i^D(o_t) = \frac{\sum_{v=1}^{V} u_i^D(v)\times ppv_K^D(o_t,v)}{U_i} \quad (4\text{-}8)$$

L'exposant *S* ou *D* des fonctions implique qu'elles s'appliquent au sous-vecteur concerné (MFCC+E ou MFCC+E_D). Deux coefficients de pondérations γ_S et γ_D sont utilisés afin de régler l'influence de chacun des groupes de paramètres. La formule de réestimation pour les coefficients d'appartenance correspondant aux paramètres delta s'obtient par échange des indices *S* et *D* dans (4-7) et (4-8).

4.1.3. Evaluation de l'apport des coefficients deltas

Le paradigme d'évaluation introduit dans le chapitre 3 est repris. Afin de mesurer comparativement l'apport des coefficients delta sur les systèmes HMM/GMM et HMM/K-ppv, deux groupes d'expériences sont considérés : l'identification locale et les reconnaissances segmentales (identification segmentale et DAP). Chaque expérience est réalisée pour chacun des systèmes HMM/GMM et HMM/K-ppv avec et sans coefficients delta dans le vecteur de paramètres.

4.1.3.1. Identification locale

Les résultats de l'identification locale par l'estimateur GMM (Tableau 4-2) ont été obtenus sur la base de donnée TIMIT après séparation silence/parole (voir §2.3.4.1) ; ce qui explique les différences de résultats de la première colonne avec le Tableau 2-3 (voir §2.3.3.2).

Nombre de Gaussiennes/ Etats	12 MFCC+E		12 MFCC+E+D	
	Appr.	Test	Appr.	Test
1	37,4	36,6	38,7	37,9
25	40,0	36,8	46,9	43,8
50	41,7	36,9	49,3	44,8
75	42,9	36,7	51,1	45,2
100	43,9	36,7	52,6	45,3

Tableau 4-2 Taux d'identification locale de l'estimateur GMM en fonction du nombre de gaussienne par somme pondérée et du jeux de paramètres.

Les résultats du Tableau 4-2 sont repris dans la Figure 4-1. On constate que l'intégration des coefficients delta dans les paramètres permet des gains en identification locale aussi bien sur le test qu'en auto-cohérence. Ces gains, relativement faible dans le cas du 1-GMM (environ 1% sur l'apprentissage et le test) s'accentuent avec l'augmentation du nombre de gaussiennes. Ainsi, l'écart de performance entre les paramètres MFCC+E et MFCC+E+D est de l'ordre de 9% pour le 100-GMM.

Avant de comparer leurs résultats, nous rappelons que les modalités de prise en compte des paramètres delta ne sont pas les mêmes pour les deux estimateurs. Ainsi, dans le cas du GMM, la technique standard (*i.e.* celle utilisée dans les systèmes HMM/GMM de l'état de l'art) correspond à une distance composite avec un poids de 1 (distance incluse dans le calcul des gaussiennes). Pour l'estimateur des K-ppv, la technique des flux de données est appliquée. Quelques valeurs sont testées en identification locale afin de déterminer un facteur de pondération optimal entre les flux associés aux coefficients statiques et delta (Tableau 4-3). Une pondération de 1 correspond aux meilleurs résultats sur les deux ensembles d'évaluation avec une augmentation des taux de l'ordre de 3% par rapport aux paramètres statiques seuls.

Pondérations	Taux d'identification	
	Appr	Test
0,0	52,1	45,8
0,5	53,5	47,1
1,0	55,2	49,0
1,5	46,3	41,6

Tableau 4-3 Taux d'identification locale de l'estimateur 50-ppv en fonction de la pondération des ppv des coefficients delta.

Lorsque l'on rapproche les résultats des estimateurs 100-GMM et 50-ppv, on constate que si le 50-ppv conserve les meilleurs résultats l'apport des coefficients delta est plus important avec l'estimateur GMM. Les écarts de performances se réduisent de 5% après l'introduction des coefficients delta.

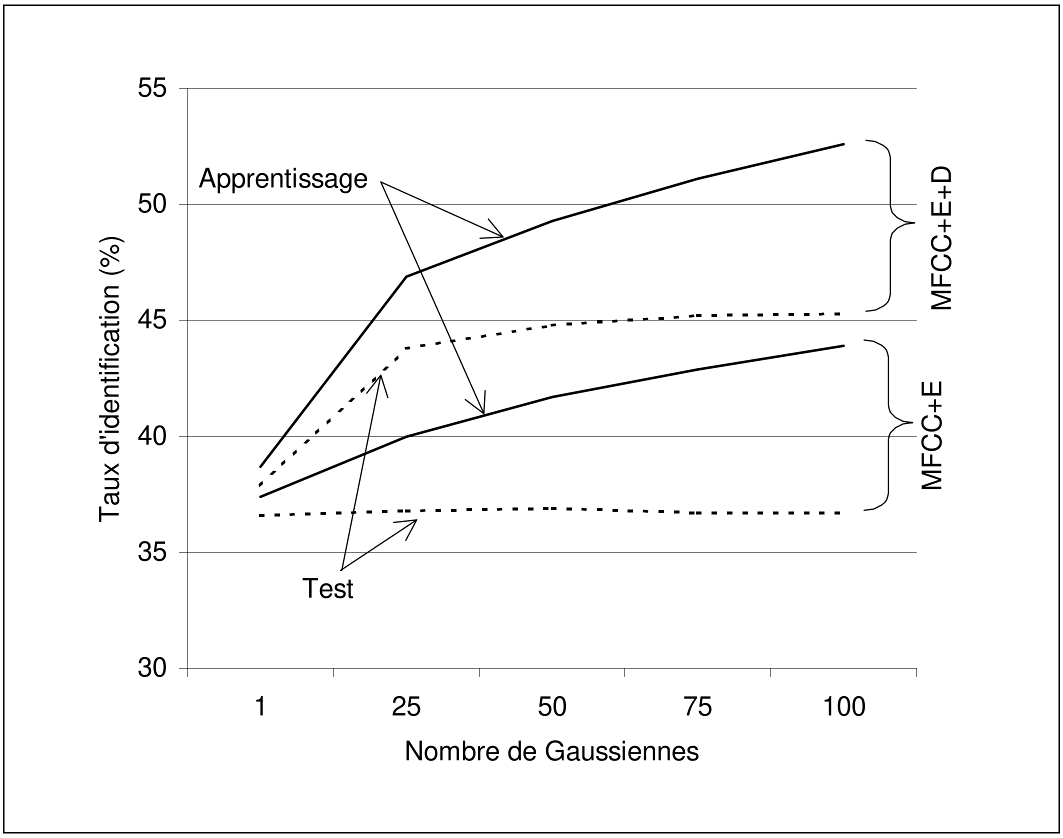


Figure 4-1 Evolution du taux d'identification locale des GMM avec et sans coefficients delta en fonction du nombre gaussiennes par mixture.

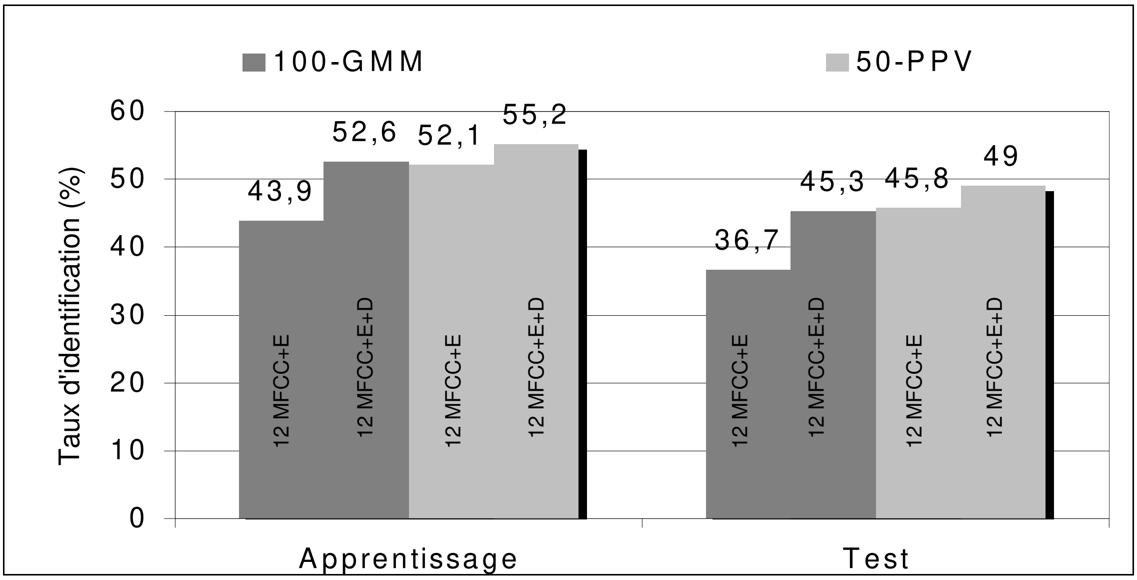


Figure 4-2 Taux d'identification locale des estimateurs 50-ppv et 100-GMM en fonction du jeu de paramètres utilisé.

4.1.3.2. Reconnaissances segmentales

Le protocole expérimental du §3.3 est repris. Afin de mesurer l'influence de l'introduction des coefficients delta, les deux types de reconnaissances segmentales sont réalisés. Les deux systèmes testés correspondent aux deux meilleurs systèmes en l'absence de coefficients delta : 50-ppv et 32-GMM.

Systèmes	Paramètres	Taux d'identification	
		Appr	Test
50-ppv	12 MFCC+E	65,6	51,1
	12 MFCC+E+D	65,3	53,9
32-Gauss	12 MFCC+E	53,9	43,6
	12 MFCC+E+D	54,0	43,6

Tableau 4-4 Taux d'identification segmentale des systèmes HMM/50-ppv et HMM/32-GMM en fonction du jeu de paramètres utilisé.

Les résultats d'identification segmentale montrent que les coefficients delta n'ont pas d'influence sur le système HMM/32-GMM (Tableau 4-4). Par contre, une amélioration du taux d'identification de 2,8% est observée pour le test du système HMM/50-ppv.

Systèmes	Paramètres	Phonèmes identifiés	Phonèmes Substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur 'échelle
Apprentissage							
50-ppv	12 MFCC+E	70,3	17,0	12,7	2,6	67,7	1,5
	12 MFCC+E+D	70,6	19,2	10,2	8,8	61,8	2,0
32-Gauss	12 MFCC+E	59,7	21,7	18,6	1,8	57,9	3,0
	12 MFCC+E+D	71,5	17,2	11,3	1,9	69,6	3,0
Test							
50-ppv	12 MFCC+E	57,6	27,8	14,6	5,0	52,5	
	12 MFCC+E+D	61,8	27,5	10,7	**11,6**	50,2	
32-Gauss	12 MFCC+E	54,6	25,3	20,1	2,3	52,3	
	12 MFCC+E+D	66,4	21,0	12,6	2,5	63,9	

Tableau 4-5 Taux de DAP des systèmes HMM/50-ppv et HMM/32-GMM en fonction du jeu de paramètres utilisé.

Les résultats de DAP (Tableau 4-5) montrent un apport très important des coefficients delta dans le système HMM/32-GMM (environ +11,6% sur l'apprentissage et le test). Le gain en performance doit être imputable à une amélioration de la qualité de la segmentation ; ce qui explique l'absence de progrès en identification segmentale. L'introduction des coefficients delta au sein du système HMM/50-ppv conduit à une dégradation des performances (-6% sur l'apprentissage et -2,3% sur le test). Pourtant, une amélioration des performances avait été relevée pour les identifications locale et segmentale. On note que les résultats des deux estimateurs ne permettent pas d'aboutir aux même conclusions.

La prise en compte des coefficients delta dans le système HMM/GMM implique donc un phénomène qui n'est pas reproduit dans le système HMM/K-ppv. Une étude plus précise de l'influence des paramètres delta dans le système HMM/GMM est menée afin de le mettre à jour.

4.1.4. Etude des paramètres delta dans le système HMM/GMM

Les expériences précédentes permettent de conclure que les coefficients delta ont un effet bénéfique sur la segmentation durant le processus de reconnaissance. Deux hypothèses sont étudiées afin de préciser le rôle des coefficients delta dans le système HMM/GMM : la création d'îlots de confiance durant le processus de reconnaissance et une meilleure adéquation des données à la topologie des modèles. Au préalable, leur influence sur le processus d'apprentissage des systèmes HMM/GMM est évaluée.

4.1.4.1. Influence des coefficients delta lors de l'apprentissage

L'influence des coefficients delta sur l'apprentissage des modèles HMM/GMM est évaluée au travers d'une expérience de découplage. Le principe est d'entraîner deux systèmes : l'un avec des coefficients statiques (12 MFCC+E), l'autre avec des coefficients statiques et delta (12 MFCC+E+D). Lors du test, les sous-vecteurs moyenne et variance des gaussiennes correspondant aux coefficients delta (12 MFCC+E_D) sont supprimés dans le second système et ajoutés aux coefficients statiques dans le premier système. Pour cela, il a été nécessaire de réaliser l'expérience avec des GMM à une gaussienne par état, faute de ne pas savoir comment mettre en correspondance les sous-vecteurs moyenne et variance entre les modèles des deux systèmes.

Paramètres	Systèmes	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage							
12 MFCC+E	Initiaux	49,4	25,2	25,4	1,6	47,9	3,0
	Composites	45,4	30,5	24,1	2,6	42,9	3,0
12 MFCC+E+D	Initiaux	57,0	25,2	17,9	1,9	55,1	3,0
	Composites	49,7	24,3	26,0	1,4	48,3	3,0
Test							
12 MFCC+E	Initiaux	46,8	27,1	26,1	1,4	45,5	
	Composites	47,6	25,2	27,2	1,7	45,9	
12 MFCC+E+D	Initiaux	55,8	25,6	18,6	2,1	53,6	
	Composites	43,9	31,4	24,7	2,8	41,0	

Tableau 4-6 Taux de DAP des systèmes HMM/1-GMM composites.

Les résultats de DAP sur TIMIT (Tableau 4-6) montrent que l'introduction des coefficients delta après l'apprentissage (41%) ne permettent pas d'obtenir d'aussi bon résultat que lorsque les coefficients statiques et delta sont entraînés conjointement (53,6%). Par contre l'interaction entre les deux types de coefficients lors de l'apprentissage n'a pas d'influence directe sur les paramètres du système correspondant coefficients statiques dès lors que les performances du système en l'absence des coefficients delta (45,9%) sont équivalentes à celles d'un système n'ayant été appris qu'avec des coefficients statiques (45,5%). L'influence des paramètres delta doit donc être recherchée dans le processus de reconnaissance.

4.1.4.2. Ilots de confiance durant le processus de reconnaissance

Dans la formulation théorique des HMM, les états émettent des symboles avec une certaine probabilité. Avec des estimateurs de fdp non-paramétriques, ces probabilités sont directement estimées à partir des points réels. Avec des estimateurs de fdp paramétriques continues, les probabilités sont extrapolées à partir des densités de probabilités. Généralement, les valeurs de fdp représentent une estimation convenable des probabilités. Toutefois, une différence notable est l'absence de normalisation des probabilités d'émission. Lorsque les variances prennent des valeurs faibles et très variables entre les différentes gaussiennes, des effets de résonance peuvent apparaître qui peuvent être perçues comme des îlots de confiance du processus de reconnaissance.

Dynamique des vraisemblances

Les effets de résonance sont liés à la dynamique des vraisemblances. Les valeurs de ces dernières sont comprises entre 0 (jamais atteint) et une valeur maximale (atteinte pour le vecteur moyenne de la gaussienne). L'augmentation de la valeur maximale va donc conduire à une plus grande dynamique des vraisemblances et permettre des écarts très important entre les probabilités d'émissions associées à une trame.

Dans le cas de l'estimateur GMM, la probabilité d'émission d'un état i pour l'observation o_t est calculée comme :

$$b_i(o_t) = \sum_{g=1}^{G} c_{ig} \times N(o_t; \mu_{ig}, \Sigma_{ig}) \tag{4-9}$$

avec G le nombre de gaussiennes dans la mixture, c_{ig} le poids de la g$^{\text{ième}}$ gaussienne, μ_{ig} son vecteur moyenne et Σ_{ig} sa matrice de covariance de la gaussienne. En développant l'expression de la loi normale, l'équation (4-9) devient :

$$b_i(o_t) = \sum_{g=1}^{G} c_{ig} \times F \times \exp\left(-\frac{1}{2}(o_t - \mu_{ig})^T \Sigma_{ig}^{-1}(o_t - \mu_{ig})\right)$$

$$\text{avec } F = \frac{1}{\sqrt{(2\pi)^d \left|\Sigma_{ig}\right|}} \tag{4-10}$$

avec $|\Sigma_{ig}|$ le déterminant de la matrice de covariance (dans le cas des matrices diagonales, il s'agit donc de la trace de la matrice). L'exposant de l'exponentielle étant toujours négatif ou nul, le facteur F représente la valeur maximale que peut atteindre la probabilité d'émission. Ainsi, lorsque *F* devient très grand, les vecteurs d'observations proches du vecteur moyenne de la gaussienne vont conduire à un effet de résonance. La Figure 4-3 représente l'histogramme des logarithmes du facteur *F* calculé à partir des 4.416 fdp gaussiennes du système HMM/32-GMM dans les cas de paramètres MFCC+E et MFCC+E+D.

Dans la mesure où les fdp estiment des probabilités, il semble raisonnable de considérer *F* inférieur ou égal à 1 comme contrainte. Elle permet, de plus, de s'assurer de l'influence des paramètres de transition qui sont eux strictement normalisés et bornés par 1. Lorsque, seuls les coefficients MFCC+E sont utilisés, on relève que 90% des fdp gaussiennes respectent cette contrainte. Cette proportion chute à 4% lorsque les coefficients delta sont introduits. La dispersion des valeurs maximales est aussi augmentée. Notre hypothèse est que la forte augmentation de la valeur maximale des GMM (la valeur 1 ne constituant qu'une référence, cohérente avec le fait que l'on estime des probabilités) ainsi qu'une plus grande dispersion des ces valeurs entre gaussiennes conduisent à l'existence, lors du décodage, de trames dont l'influence est démultipliée jusqu'à constituer des îlots de confiance.

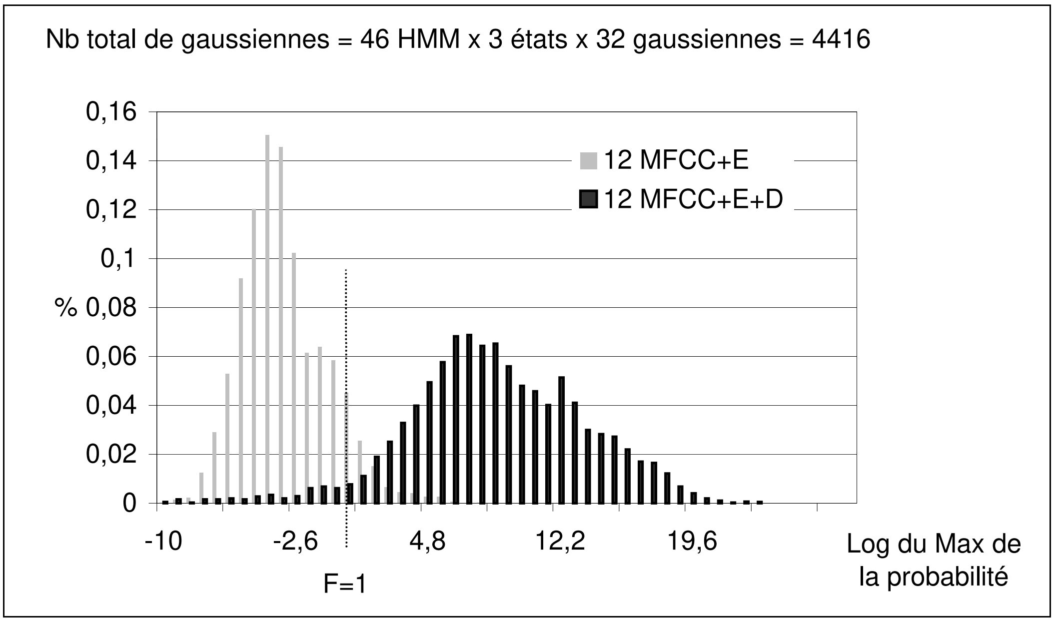


Figure 4-3 Histogramme des logarithmes des maxima de probabilités (F) dans un système HMM/32-GMM en fonction du jeu de paramètres.

L'impact de l'effet de résonance est mesuré sur un DAP sur la base TIMIT. Il n'est pas marginal dès lors qu'il intervient pour 21% des trames parmi les 48.622 du noyau de test. Les classes phonétiques les plus concernées sont /cl/ (occlusions), /ey/, /f/ et surtout /s/. Nous proposons de saturer les valeurs des vraisemblances lors du décodage afin d'évaluer l'influence réelle de l'effet de résonance.

Saturation des vraisemblances lors du décodage

La saturation s'opère en ramenant à 1 les probabilités d'émission plus grande que 1. Cette méthode, quelque peu rude, a peu d'influence sur la reconnaissance dont le taux n'a diminué que de 1%. Une normalisation permettant de ramener la dynamique des vraisemblances à celle des paramètres de transition a été appliquée : lorsqu'une résonance se produit, les probabilités d'émission de tous les états pour cette trame sont divisées par la valeur du pic de résonance de façon à conserver la plus grande des probabilités d'émission inférieure ou égale à 1. Cette expérience n'a montré aucune influence sur le taux de reconnaissance du DAP.

Ces expériences ne permettent pas de conclure que le processus de reconnaissance est sensible à l'effet de résonance. Ainsi, l'influence des coefficients delta sur le processus de reconnaissance n'apparaît pas être le fait du calcul de la fdp gaussienne. Cette hypothèse étant écartée, la modification de la relation entre les données et la topologie des HMM impliquée par l'utilisation des coefficients delta est étudiée.

4.1.4.3. Meilleure adéquation à la topologie

Du point de vue de la théorie de l'information, la topologie d'un HMM est d'autant plus adaptée aux données que ses états jouent un rôle égal. Ceci signifie, au moins théoriquement, que la qualité d'une topologie peut être évaluée aussi au travers de la quantité de données modélisées par chacun de ses états après convergence de la procédure d'apprentissage. L'entropie du modèle nous semble un bon critère pour mesurer le niveau d'adéquation entre un modèle et ses données.

Le décodage d'une séquence acoustique sur un modèle permet de déterminer la suite d'états qui a le plus probablement engendrée la séquence. Cette suite d'états peut-être vue comme l'émission des symboles du HMM. Ainsi, l'utilisation de chaque état dans la séquence est considérée comme indépendante des précédents. Il est possible d'estimer la probabilité d'apparition liée à chacun des états et donc d'estimer l'entropie du modèle du point de vue de l'utilisation de ses états.

Un HMM est considéré comme une source d'information émettant N symboles différents. Dans ce cas, les symboles sont les états activés lors du décodage d'une séquence acoustique. Le calcul de l'entropie du modèle nécessite donc l'estimation des fréquences d'activation des états. Ces fréquences sont liées aux séquences d'états cachées (et donc inconnues) des segments à travers leur modèle. Pour les obtenir, les segments sont alignés à l'aide de l'algorithme de Viterbi qui fournit la séquence d'état la plus probable. Il est alors possible d'estimer l'entropie de la toplogie sur l'ensemble des séquences de test. L'entropie des HMM obtenue est comparée à l'entropie théorique d'un modèle équiprobable.

Une source émettant N symboles de manière équiprobable a une entropie valant $\log_2 N$. Dans le cas d'un modèle gauche-droit à 3 états, N vaut 3. Les valeurs des entropies moyennes pour le système HMM/32-GMM sont évalués sur le noyau de test de TIMIT avec et sans coefficients delta (Tableau 4-7).

Systèmes	MFCC+E	MFCC+E+D	Maximum Théorique
Entropie	1,56	1,57	1,58

Tableau 4-7 Entropie moyenne des topologies pour le système HMM/32-GMM avec (MFCC+E+D) et sans (MFCC+E) coefficients delta.

Il apparaît que le système utilisant les coefficients delta possède une entropie moyenne plus proche de l'entropie théorique d'un modèle à 3 états. Toutefois, le faible écart entre les entropies des systèmes avec et sans coefficients delta ne permet pas d'aboutir à une conclusion franche. Lors d'une précédente expérience menée sur la base TIMIT complète (*i.e.* avec les silences), un écart plus grand (+0,05) avait été observé [Lefèvre, Montacié *et al.*, 1998]. La relation entre l'introduction des coefficients delta et une meilleure adéquation des données à la topologie des HMM paraît donc surtout liée à la modélisation de la classe des silences.

4.1.5. Conclusion

L'introduction des coefficients delta dans les paramètres du système HMM/K-ppv n'a pas permis une amélioration de ses performances en DAP. Ce résultat est surprenant dans la mesure où des gains sont observés sur les expériences d'identification locale et segmentale. Nous avions noté dans le chapitre précédent que l'augmentation des performances sur les tâches d'identification n'était pas nécessairement répercutée sur le DAP. Nous observons maintenant qu'il peut arriver que les résultats de DAP soient dégradés. Inversement, alors que les performances en identification du système HMM/GMM sont inchangées, l'introduction des coefficients delta améliore considérablement ses résultats en DAP.

Une étude complémentaire visant à mettre en évidence l'influence réelle des coefficients delta dans le système HMM/GMM a été menée. Ses conclusions ne nous ont pas permis de définir un artefact permettant de recréer un phénomène comparable dans le système HMM/K-ppv. L'adaptation de l'espace de représentation n'ayant pas permis d'améliorer les performances du système HMM/K-ppv, l'évolution du système de base HMM/K-ppv est poursuivie par l'étude de la modélisation contextuelle présentée dans le cadre plus large de l'adaptation des références.

4.2. Adaptation des références

Un opérateur de RdF peut être amélioré par la remise en cause de l'association *a priori* entre les données et les classes. Afin d'augmenter la cohésion intra-classe, il peut être profitable de subdiviser les classes en sous-classes. Dans le cadre de la RAP, l'application de ce principe aux classes phonétiques conduit à différents types de modélisation en fonction du critère de subdivision utilisé.

Inversement, afin d'augmenter la précision de l'opérateur de reconnaissance, il peut être profitable de diminuer le nombre de classes pris en compte simultanément. Pour cela, le processus de reconnaissance débute par une phase une reconnaissance en macro-classes. Dans les deux cas, le recours à l'adaptation des références est motivé par le problème du recouvrement des classes.

4.2.1. Recouvrement des classes

Le "recouvrement" dans l'espace de représentation des différentes classes à identifier est une difficulté classique en RdF. Il s'agit d'une des remises en cause les plus graves de l'hypothèse d'agrégation [Simon, 1984] dans la mesure où des vecteurs de classes différentes occupent des régions communes de l'espace de représentation. Ce recouvrement peut n'être que partiel et n'affecter que très faiblement la reconnaissance. Il peut-être total et entraîner la non reconnaissance d'une classe complète (c'est le cas de la diphtongue /oy/ pour le DAP du système HMM/50-ppv, cf. Tableau 3-16).

La RAP est un domaine où le recouvrement affecte très fortement la reconnaissance et ce quel que soit l'espace de représentation proposé, du moins actuellement. Le choix des classes, lorsqu'il est permis, est donc d'autant plus important que l'objet à identifier peut être déformé (variantes allophoniques). Une première approche consiste à déterminer un critère qui permette la partition automatique des classes d'apprentissage en sous-classes homogènes respectant au mieux les "déformations-types". Il s'agit de la modélisation sub-phonétique. Compte tenu de déformations ainsi minimisées au sein d'une même classe phonétique, on peut espérer mieux modéliser et reconnaître les éléments de cette classe. Ce critère peut être issu de connaissances expertes de la tâche de reconnaissance ou déduit des données de référence.

Plusieurs approches de modélisations sub-phonétiques peuvent être envisagées. Les plus classiques issues de l'état de l'art des systèmes HMM/GMM sont fondées sur un apport de connaissance dans le système. Un critère issu de la connaissance experte de la tâche permet de séparer les représentants d'une classe en plusieurs sous-classes plus homogènes. Ce critère peut être :

- le *genre* du locuteur qui permet de considérer deux sous-classes pour chaque classe phonétique ;

- le *contexte phonétique* qui permet de dégager un nombre variable de sous-classes en se fondant sur un critère de production de la parole.

Une troisième approche originale, non paramétrique, consiste à dégager des sous-classes cohérentes à partir des données en se fondant uniquement sur un critère d'*agrégation* dans l'espace de représentation.

La subdivision des classes n'est pas la seule manière de gérer le problème du recouvrement. Une seconde approche est basée sur la constitution de groupes de classes (agrégation en macro-classes) particulièrement confus. L'objectif est alors d'augmenter la précision de l'opérateur de reconnaissance en diminuant le nombre de classes prises en compte simultanément.

4.2.2. Modélisation par genre

Dans les systèmes HMM/GMM, la modélisation par genre est généralement assimilée à une méthode d'adaptation. Ceci est justifié car, dans leur cas, les modèles par genre sont obtenus par un ré-apprentissage des modèles globaux avec, éventuellement, un critère de type MAP ou MLLR (voir §1.3.3.3). Dans le cas d'un système HMM/K-ppv, les modèles par genre sont obtenus par la partition de l'ensemble de référence et l'apprentissage direct de deux jeux séparés de modèles. Il s'agit donc de la subdivision de chaque classe phonétique en deux sous-classes basées sur un critère lié au sexe du locuteur. Des différences notables dans le dimensionnement de l'appareil phonatoire entre hommes et femmes ayant une incidence acoustique justifie cette taxinomie [Calliope, 1989].

La normalisation par la longueur du canal vocal (VTLN, voir §1.3.3.2), dont le principe est d'introduire une transformation linéaire de l'échelle des fréquences lors de l'estimation des paramètres acoustiques, représente une alternative récente qui tend à se développer [Lee & Rose, 1996]. La modélisation par genre a été retenue dans nos expériences car elle est fondée sur des connaissances précises facilement vérifiables et simples à mettre en œuvre. Une augmentation de quelques pour-cent du taux de reconnaissance est communément relevée sur le DAP de TIMIT après l'application de cette technique. Deux modes d'application de la modélisation par genre peuvent être considérés : la détermination préalable du genre ou l'approche multi-modèle.

4.2.2.1. Détection préalable du genre du locuteur

La modélisation par genre est un cas particulier de la modélisation sub-phonétique dans la mesure où son intérêt ne réside pas uniquement en une définition plus fine des classes dans l'espace de représentation. Le développement des techniques de reconnaissance du locuteur permet de connaître avec une assez grande réussite (proche de 100% dans de bonnes conditions) le sexe d'un locuteur. Il est alors envisageable d'effectuer un pré-traitement permettant de déterminer le sexe du locuteur afin d'utiliser le jeu adéquat de modèle lors de la reconnaissance.

Notre but étant de mesurer l'apport maximum d'une modélisation par genre dans un système HMM/K-ppv, nous nous sommes placés dans le cas d'une identification préalable idéale du sexe du locuteur. Cette expérience est réalisable car le sexe du locuteur nous est donné pour toutes les phrases de TIMIT.

Modélisation	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Complète	70,3	17,0	12,7	2,6	67,7	1,5
Homme	70,4	16,7	12,9	2,4	68,0	1,5
Femme	71,8	16,5	11,8	2,6	69,1	1,5
Homme et Femme	70,8	16,6	12,6	2,5	68,3	
Test						
Complète	57,6	27,8	14,6	5,0	52,5	
Homme	57,6	27,3	15,1	5,0	52,7	
Femme	58,3	26,4	15,3	4,1	54,2	
Homme et Femme	57,9	27,0	15,2	4,7	53,2	

Tableau 4-8 Taux de DAP d'un système HMM/50-PPV avec décodage séparé des genres.

La séparation par genre conduit à une amélioration non significative du taux de reconnaissance global (homme et femme) du système (Tableau 4-8). Seuls les modèles féminins profitent réellement de la subdivision avec une augmentation de 1,7% du taux de reconnaissance sur le test. Toutefois, les locuteurs masculins étant plus nombreux dans le test, cette amélioration n'a que peu d'effet sur le taux global.

4.2.2.2. Approche multi-modèle

Afin de mesurer l'apport de la modélisation par genre dans le cadre strict de la modélisation sub-phonétique, la détermination préalable du sexe du locuteur n'est plus considérée. On se place dans le cadre d'une approche multi-modèle dans laquelle l'ensemble des modèles sont considérés simultanément lors de la reconnaissance [Jouvet, Dautremont *et al.*, 1994]. Le choix entre les différents modèles de la même unité phonétique découle alors directement de la qualité de la modélisation obtenue par la considération de sous-classes plus homogènes.

Modélisation	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Complète	70,3	17,0	12,7	2,6	67,7	1,5
Homme et Femme	71,2	17,6	11,2	4,1	67,1	1,5
Test						
Complète	57,6	27,8	14,6	5,0	52,5	
Homme et Femme	58,4	28,7	12,8	7,5	51,0	

Tableau 4-9 Taux de DAP d'un système HMM/50-PPV avec modélisation par genre sans détection préalable.

L'approche multi-modèle ne conduit pas à une augmentation des performances (Tableau 4-9). On en déduit que l'information sur le sexe du locuteur n'est pertinente pour le système HMM/K-ppv que lorsqu'elle est fournie explicitement lors du décodage. L'objectif d'une modélisation en sous-classes plus homogènes n'est pas atteint, un critère de subdivision plus performant doit être utilisé.

4.2.3. Modélisations sub-phonétiques

Deux types de modélisations sub-phonétiques sont considérées : la modélisation contextuelle et la modélisation non-supervisée.

4.2.3.1. Modélisation contextuelle

L'approche la plus classique de la modélisation sub-phonétique supervisée nous est donnée par une considération de production de la parole : la co-articulation. La conservation de l'énergie lors du mouvement articulatoire conduit à minimiser ou renforcer certains mouvements en fonction des suivants. Ainsi, certains phonèmes auront des prononciations différentes en fonction des phonèmes les accompagnant : les variantes allophoniques contextuelles.

<u>Principes</u>

Différentes formes de modélisation contextuelle existent. Le triphone tient compte des phonèmes précédents et suivants, les biphones droits du phonème suivant et les biphones gauches du phonème précédent. Des extensions à un voisinage de 4 ou 5 phonèmes ont été développées [Woodland, Leggetter *et al.*, 1995]. Mais leur emploi reste marginal du fait d'un gain en performance assez faible comparée à la difficulté de leur apprentissage.

Pour un ensemble initial de 39 phonèmes, il existe 59.319 triphones possibles. La constitution d'un corpus d'apprentissage suffisant pour chaque classe est rendue difficile du fait de leur démultiplication. De façon générale, ce problème peut être résolu par la technique du partage de paramètres (§1.3.2) qui a pour effet d'augmenter artificiellement le nombre d'exemples pris en compte lors de l'apprentissage d'un modèle. L'application de cette technique au problème de la modélisation d'allophones contextuelles a donné lieu a diverses solutions. Elles peuvent être séparées selon deux critères :

- le *niveau de partage* de paramètres : quatre niveaux sont généralement considérés (le modèle, l'état, la fdp et les paramètres de la fdp). L'ensemble du modèle est partagé dans le cas des allophones généralisés [Lee, Hon *et al.*, 1990]. Dans les approches plus récentes, le partage est effectué au niveau de l'état [Ljolje, 1994 ; Young & Woodland, 1994; Hwang, Huang *et al.*, 1996 ; Junqua & Vassallo, 1996].

- le *type du critère* de regroupement : la décision peut être basée sur des connaissances expertes (critère supervisé) ou uniquement guidé par les données (critère non-supervisé). Les quasi-triphones reposant sur des connaissances de nature phonologique sont un exemple de méthode supervisée [Ljolje, 1994]. Ils sont obtenus à partir de la concaténation de modèles dépendant du contexte gauche ou droit pour en faire des modèles gauche et droit.

  Un regroupement automatique des paramètres implique la définition d'une mesure de similarité. Une fois celle-ci définie, la procédure est classique et s'applique quel que soit le niveau de partage de paramètres. Initialement, chaque triphone rencontré dans le corpus d'apprentissage est modélisé par un HMM. Puis, chaque HMM est associé à un cluster (il y a au départ autant de clusters que de HMM dans le système). La mesure de similarité permet d'identifier les 2 clusters les plus proches. Ils sont regroupés. Ces étapes sont répétées jusqu'à réalisation du critère de convergence choisi (nombre final de contextes indépendants, données suffisantes pour chacun des contextes…). Les mesures de similarité entre HMM sont nombreuses : entropie croisée, divergence, discrimination [Juang & Rabiner, 1985], mesure du χ_2 [Paul & Martin, 1988], information mutuelle [Lee, 1988] ou encore mesure de l'adéquation à une loi de Poisson [Bahl, 1990].

  L'application directe de ces mesures ne permet pas de modéliser les contextes non rencontrés dans l'ensemble d'apprentissage. Afin d'y remédier, il est possible de réaliser le partage des paramètres par le biais d'un arbre de décision construit en minimisant un critère d'entropie [Bahl, de Souza *et al.*, 1991]. L'arbre est construit par séparation progressive des contextes en 2 sous-ensembles qui sont le résultat d'une question binaire maximisant leur homogénéité. Cette opération est répétée jusqu'à réalisation d'un critère d'arrêt. Il est alors possible de déterminer les paramètres à utiliser pour n'importe quel contexte, y compris ceux qui n'ont pas été rencontrés lors de sa construction, en parcourant l'arbre jusqu'à l'une de ses feuilles.

Sur la base TIMIT, une augmentation de prés de 10% en valeur absolue du taux de reconnaissance phonétique a pu être obtenue uniquement à l'aide de modèles en contexte droit [Lee & Hon, 1989].

<u>Modélisation contextuelle dans le système HMM/K-ppv</u>

La modélisation contextuelle est appliquée au système HMM/K-ppv. En première approche, une modélisation par biphones droits sera utilisée. Une étude précédente a montré que dans le cadre de la base de données TIMIT l'écart de performance est faible entre les modélisations par triphone et par biphone [Barras, 1996]. De plus, les taux de reconnaissance des modèles contextuels droits sont légèrement meilleurs que ceux des modèles gauches. Une identification accrue des occlusives pré-vocaliques peut expliquer la différence de comportement entre les modèles gauches et droits [Lee, 1988]. Une autre hypothèse, plus globale, est que l'anticipation prime sur la rémanence [Derouault, 1987].

Seuls les biphones droits possédant au moins 50 occurrences dans l'ensemble d'apprentissage sont retenus. A partir des 39 classes phonétiques initiales, on obtient 519 modèles sur les 1521 possibles (480 biphones droits et 39 monophones afin de modéliser les biphones non représentés). Ces modèles sont entraînés selon le protocole d'apprentissage déduit du chapitre 3.

Afin d'augmenter la qualité de la modélisation, le partage de paramètres est appliqué aux biphones droits au niveau des états. Nous proposons d'utiliser la divergence de Kullback [Kullback & Leibler, 1951], dans sa version symétrique, comme mesure de similarité entre les fdp des états. La mesure s'applique sur les fdp b_i et b_j des états i et j comme:

$$KL(b_i \| b_j) = \int_x b_i(x) \log \frac{b_i(x)}{b_j(x)} dx + \int_x b_j(x) \log \frac{b_j(x)}{b_i(x)} dx \qquad (4\text{-}11)$$

L'intégrale sur l'espace de représentation est rapportée à une sommation sur l'ensemble O des vecteurs d'observation o de l'ensemble d'apprentissage :

$$\begin{aligned} KL(b_i \| b_j) &= \sum_O \left\{ b_i(o) \log \frac{b_i(o)}{b_j(o)} - b_j(o) \log \frac{b_j(o)}{b_i(o)} \right\} \\ &= \sum_O \left(b_i(o) - b_j(o) \right) \times \left(\log b_i(o) - \log b_j(o) \right) \end{aligned} \qquad (4\text{-}12)$$

Le calcul de la divergence entre 2 fdp nécessite l'évaluation des fdp sur tous les vecteurs de l'ensemble d'apprentissage. Cette procédure a un coût de calcul global en $o(VS^2)$ avec V le cardinal de l'ensemble d'apprentissage et S le nombre total d'états dans le système. Avec V de l'ordre de 1 million et S autour de quelques milliers, cette procédure est excessivement coûteuse. La solution que nous avons retenue pour amoindrir son coût consiste à diminuer le nombre de vecteurs utilisés pour l'estimation de la divergence. Ainsi, $V/100$ vecteurs ont été considérés au lieu de V ; soit 9.050 vecteurs dans le cas de l'ensemble d'apprentissage de TIMIT.

Plusieurs approches sont envisageables pour extraire ce sous-ensemble de vecteurs. Il semble raisonnable que les proportions des classes phonétiques dans ce dernier soient conformes à celles de l'ensemble complet. Pour ce faire, les probabilités *a priori* des classes ont été calculées puis une ponction aléatoire a été opérée sur l'ensemble d'apprentissage jusqu'à obtenir le nombre de vecteurs désiré pour chaque classe.

Les résultats du DAP sont donnés dans le Tableau 4-10 en fonction du nombre d'états indépendants du système. Il apparaît que la modélisation par biphones droits (correspondant à 1557 états pour 519 modèles) n'est pas adaptée, en l'état, au système

HMM/K-ppv dans la mesure où elle conduit à une forte dégradation des performances (-9,3% en auto-cohérence et –13,9% sur le test). On note, de plus, qu'afin d'obtenir ces résultats l'influence du modèle de langage (bigramme phonétique) a du être renforcée puisque l'on passe d'un facteur d'échelle de 1,5 avec les monophones à un facteur 3,5 avec les biphones droits. Une tentative d'explication de ces résultats sera faite dans §4.2.3.3. Les expériences de partages de paramètres d'états ont été réalisées malgré les résultats observés avec les biphones droits. Leur intérêt est de vérifier si une amélioration des performances peut être espérée de la technique.

Le choix du nombre d'état après regroupement est arbitraire. Il est possible d'introduire un critère d'optimalité ; la variation du taux de reconnaissance en auto-cohérence par exemple. Notre but, dans ces expériences préliminaires, est simplement de vérifier que l'augmentation du nombre de données servant à l'estimation des fdp a une influence sur le taux de reconnaissance global. Le nombre d'états regroupés a donc été fixé de manière arbitraire à 200 ; ce qui correspond à un système final de 1357 états indépendants. Faute de temps, le facteur d'échelle n'a pas été optimisé pour le système à 1357 états. Il est vraisemblance que la dynamique des vraisemblances n'est que faiblement modifiée par la diminution du nombre d'états. Le facteur d'échelle obtenu avec les biphones droits sans états liés a donc été réutilisé.

Nb états total	Phonèmes identifiés	Phonèmes Substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
117 (39x3)	70,3	17,0	12,7	2,6	67,7	1,5
1557 (519x3)	75,2	19,5	5,4	**16,7**	58,5	3,5
1357	74,8	19,7	5,5	**16,6**	58,2	3,5
Test						
117 (39x3)	57,6	27,8	14,6	5,0	52,5	
1557 (519x3)	60,5	33,7	5,8	**28,9**	31,6	
1357	60,7	33,6	5,7	**29,3**	31,5	

Tableau 4-10 Taux de DAP du système HMM/50-ppv en contexte droit en fonction du nombre total d'états dans le système.

On constate que le partage d'états entre les biphones droits du système HMM/50-ppv n'a pas d'influence significative sur le taux de reconnaissance du DAP (Tableau 4-10). Deux hypothèses peuvent expliquer ce résultat : le critère de regroupement n'est pas adapté (éventuellement mal estimé à cause d'un nombre trop faible de vecteurs) ou le nombre d'états liés est encore insuffisant pour permettre une estimation fiable des fdp de chaque état. Toutefois, le niveau de performance de la modélisation contextuelle étant très inférieur à celui du système de base, il nous semble raisonnable d'envisager une méthode de modélisation plus adaptée au système HMM/K-ppv avant d'étudier plus avant la technique de partage de paramètres.

4.2.3.2. Modélisation sub-phonétique non-supervisée

Les modélisations introduites précédemment (par genre et contextuelle) reposent toutes deux sur l'introduction d'une connaissance externe. Dans un cas, il s'agit de l'influence du sexe du locuteur, dans l'autre de l'influence du contexte phonétique. Ces connaissances ne s'appliquent pas toujours aussi bien quelle que soit la tâche concernée. Ainsi, il a pu être montré que, sous certaines conditions, une subdivision des classes selon l'accent régional du locuteur était préférable à une séparation par son genre [Van Compernolle, Smolders *et al.*, 1991]. De même, comme nous l'avons signalé dans la section précédente, l'emploi des biphones à la place des triphones peut se révéler mieux adapté à une tâche en fonction de la taille de l'ensemble d'apprentissage.

Une approche non supervisé permettra d'obtenir une division sub-phonétique des classes toujours adaptée à la tâche. Le critère de la division est déduit de la proximité spatiale des séquences d'apprentissages. Un algorithme de type k-moyennes est utilisé (*LGB*, [Linde, Buzo *et al.*, 1980]) pour séparer les segments d'une classe en sous-ensembles qui permettront l'apprentissage de sous-classes non-supervisées. Le nombre de sous-classes peut être obtenu à partir d'un seuil de distorsion. Nous avons choisi de le fixer *a priori*.

L'ensemble des trames d'une classe phonétique est séparé en *N* clusters. Chaque séquence est associée au cluster majoritaire parmi les trames la composant. Chaque classe est ainsi subdivisée en *N* sous-classes (correspondant aux *N* clusters précédents). Pour chaque sous-classe un modèle est appris sur les données de la sous-classe exclusivement. Lors du décodage, l'ensemble des modèles sont mis en parallèle et associés à leur classe phonétique initiale (pour le modèle de langage). Des systèmes à 3 et 6 sous-classes phonétiques ont été appris. Les taux de reconnaissance en DAP, bien que supérieurs à ceux obtenus avec la modélisation par biphones droits, sont encore inférieurs à ceux du système de base, toujours en raison d'un fort tau d'insertions (Tableau 4-11). De plus, l'augmentation du nombre de sous-classes ne permet pas d'augmenter les résultats sur le test alors qu'un gain de quelques pour-cent est constaté en auto-cohérence entre 3 et 6 sous-classes.

Cette technique est à rapprocher de l'inférence de topologie *a posteriori* (§3.2.3.1) qui présente des performances comparables. Dans le cas de l'inférence, la clusterisation des trames phonétiques ne servait qu'à fournir une initialisation aux modèles. Il s'agissait d'une contrainte moins stricte dans la mesure où, lors du décodage, les segments pouvaient être alignés sur des états appartenant à des sous-classes différentes.

Nb de sous-classes	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
0	70,3	17,0	12,7	2,6	67,7	1,5
3	71,3	19,6	9,1	9,6	61,7	2,5
6	73,4	18,8	7,8	8,9	64,5	2,5
Test						
0	57,6	27,8	14,6	5,0	52,5	
3	58,2	31,2	10,6	**13,2**	45,0	
6	59,0	32,1	8,9	**15,3**	43,7	

Tableau 4-11 Taux de DAP du système HMM/50-ppv avec modélisation sub-phonétique non-supervisée en fonction du nombre de sous-classes engendrées.

Cette méthode, comme les précédentes, ne permet pas d'améliorer les performances du système HMM/K-ppv. Il semble que la subdivision des classes phonétiques ne soit pas adaptée dans le cas de l'estimateur des K-ppv. Les raisons de cette inadéquation peuvent être recherchées dans la nature des paramètres du système HMM/K-ppv.

4.2.3.3. Limites de la modélisation sub-phonétique pour le système HMM/K-ppv

Au-delà des résultats des expériences précédentes, quelques considérations permettent de mesurer les limites de l'application de la modélisation sub-phonétiques au système HMM/K-ppv.

Dans le cas du système HMM/GMM, la modélisation sub-phonétique intervient généralement comme une méthode d'adaptation. Des modèles phonétiques initiaux sont appris ; puis ces modèles sont dupliqués en autant de modèles sub-phonétiques qui sont adaptés sur le sous-ensemble de données correspondant. Cette méthode présente l'avantage d'éviter le sous-apprentissage des modèles. Or, dans le cas du système HMM/K-ppv, les modèles sub-phonétiques sont appris directement sur leurs données propres. L'impossibilité d'initialiser les paramètres des modèles sur l'ensemble des exemples de la classe phonétique est due à la nature des paramètres.

Dans le cas du système HMM/GMM, les vecteurs d'une sous-classe phonétique ont participé à l'estimation initiale des paramètres pour l'ensemble des autres modèles de leur classe phonétique. Pour le système HMM/K-ppv, les coefficients d'appartenance définissent l'appartenance des vecteurs de manière définitive et ne permettent donc pas en l'état l'application de techniques d'adaptation. Une solution permettant de remédier à ce problème est de lever la contrainte de normalisation des coefficients d'appartenance. Ainsi, le modèle d'une classe phonétique pourra être appris à l'aide de l'ensemble des données disponibles ; puis les modèles sub-phonétique, déduits par duplication du modèle initial, seront réappris sur la portion qui leur est associé de l'ensemble complet. A chaque vecteur de référence sera associé deux jeux de coefficients d'appartenance : l'un pour son association à la classe phonétique initiale et l'autre pour son association à la sous-classe phonétique. Cette technique pose alors le problème de la prise en compte simultanée de plusieurs ensembles de coefficients d'appartenance (à l'instar de la modélisation par flux, dont nous avons vu qu'elle ne conduisait pas à une amélioration des performances avec les coefficients delta §4.1.3). Nous avons choisi d'étudier au préalable l'opportunité d'une méthode d'adaptation des références basée sur une modélisation en macro-classes phonétiques.

4.2.4. Agrégation en macro-classes phonétiques

L'intérêt de la subdivision des classes phonétiques est de produire des sous-classes plus homogènes et donc mieux modélisées dans le système. Le point de vue, dans ce cas, est qu'une meilleure définition des références peut améliorer la reconnaissance. Une autre approche consiste à scinder le processus de reconnaissance en deux étapes : la reconnaissance en macro-classes puis l'identification des classes phonétiques. L'intérêt de cette méthode est de diminuer le nombre de classes pris en compte à chaque étape.

La notion de macro-classe est liée à des considérations acoustiques. Les phonèmes regroupés en macro-classes partagent des caractéristiques de production communes. Ces dernières se manifestent au travers de traits acoustiques : le voisement, la nasalité, la présence d'occlusion…

4.2.4.1. Présentation des macro-classes

Six macro-classes phonétiques sont considérées : les voyelles (avec les diphtongues), les liquides (avec les semi-voyelles), les nasales, les fricatives (avec les affriquées), les occlusives et les occlusions. Les classes phonétiques composant chaque macro-classe peuvent être trouvées dans le Tableau E-2 de l'Annexe E. L'importance relative des différentes macro-classes phonétiques est mesurée par leur nombre de vecteurs de référence (cf. Tableau 4-12, qui reprend aussi le nombre de classes phonétiques incluses dans chaque macro-classe ainsi que le nombre de segments dans l'ensemble de référence). Les voyelles sont les plus représentées (presque 45% de l'ensemble total), suivies des occlusions et des fricatives. Les liquides, les nasales et les occlusives représentent chacune moins de 10% de l'ensemble.

Macro-classe	Nb de classes phonétiques	Nb de vecteurs de référence
Occlusions	1	111129 (11,7%)
Fricatives	10	191871 (20,2%)
Liquides	4	80865 (8,5%)
Nasales	3	71822 (7,5%)
Occlusives	7	68484 (7,2%)
Voyelles	14	426629 (44,9%)

Tableau 4-12 Importance relative des macro-classes dans l'ensemble de référence.

4.2.4.2. Reconnaissance en macro-classes

L'opérateur de reconnaissance en macro-classes est composé de 7 sous-opérateurs : un opérateur pour les macro-classes et 6 opérateurs spécialisés par macro-classe. Dans une première étape, les phrases de l'ensemble d'évaluation sont décodées en macro-classe. Ensuite, chaque segment obtenu est identifié par le sous-système correspondant à sa macro-classe. Les différentes étapes du DAP avec reconnaissance en macro-classes sont reprises dans la Figure 4-4.

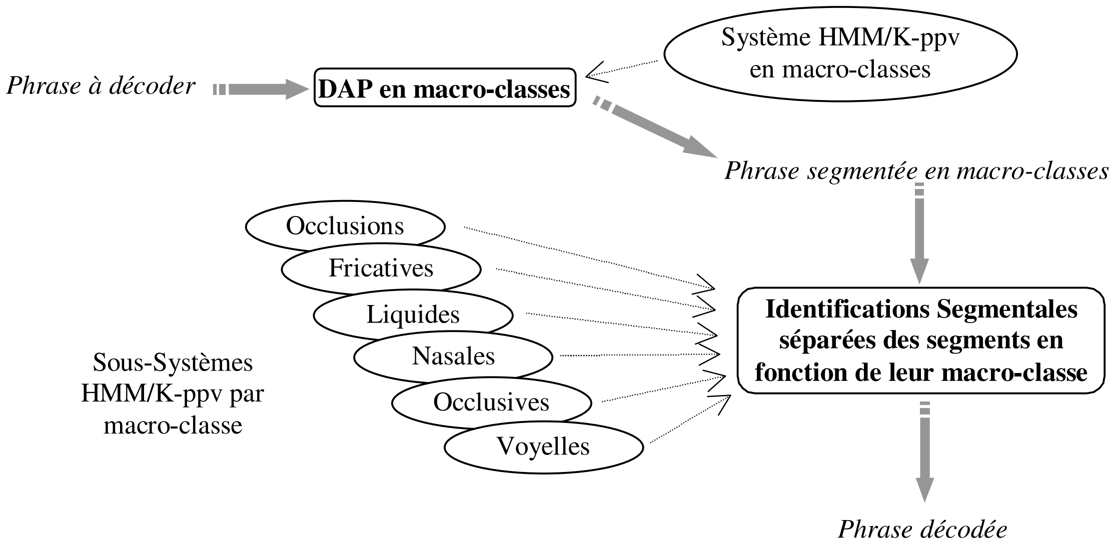


Figure 4-4 Synoptique du DAP avec reconnaissance en macro-classes.

Le taux de reconnaissance du décodage en macro-classes par le système HMM/50-ppv est 70,4% (Tableau 4-13). Or, le regroupement en macro-classes d'un décodage réalisé à partir d'un système standard (avec 39 classes) permet d'atteindre un taux de 74,6%, grâce notamment à une meilleure reconnaissance des voyelles. Cet écart de performance peut s'expliquer par une tendance du système en macro-classes à sous-segmenter le signal lors de la reconnaissance ; ce qui conduit à une forte augmentation du taux d'omission (+6% entre les deux systèmes). Cet effet est compensé, mais insuffisamment, par une réduction du taux d'insertion de l'ordre de 2%.

Modélisation	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Standard (39cl)	83,3	4,1	12,6	2,5	80,8	1,5
Macro-classes	75,7	5,3	**19,0**	1,6	74,1	1.
Test						
Standard (39cl)	79,4	6,3	14,3	4,8	74,6	
Macro-classes	72,9	6,8	**20,3**	2,5	70,4	

Tableau 4-13 Taux de DAP du système HMM/50-ppv avec décodage en macro-classes.

Les expériences de décodage à partir d'un système en macro-classes n'ont pas été réalisées dans le cadre des HMM/GMM. Toutefois, une expérience réduite sur TIMIT, livrée avec HTK, donne un taux de reconnaissance en macro-classe d'environ 70% pour un système HMM/GMM contextuel.

4.2.4.3. Identification segmentale séparée des macro-classes

Une fois la reconnaissance en macro-classes réalisée, les segments étiquetés sont transmis aux sous-systèmes pour leur identification en 39 classes. Cette étape est comparable à l'identification segmentale à partir du système standard HMM/K-ppv à 39 classes (voir Tableau 3-13) à deux exceptions près : la localisation des segments n'est pas issue de l'étiquetage de référence et la macro-classe du segment est déterminée préalablement (elle est d'ailleurs fausse dans 30% des cas, cf. Tableau 4-13).

Afin de mesurer uniquement l'apport de la répartition du processus d'identification en 6 sous-opérateurs indépendants, une expérience d'identification segmentale a été réalisée en utilisant l'étiquetage de référence de TIMIT afin de simuler une reconnaissance préalable des macro-classes idéale. Les taux d'identification se révèlent nettement supérieurs (+11,3% sur le test) à ceux obtenus à partir d'une modélisation en 39 classes (Tableau 4-14). L'intégration d'une connaissance supplémentaire lors du décodage (la macro-classe des segments) explique en grande partie ces résultats.

Modélisation	Taux d'identification	
	Appr	Test
Standard (39cl.)	64,1	56,0
Macro-classes	78,8	67,3

Tableau 4-14 Taux d'identification segmentale dans le cas d'une modélisation séparée des macro-classes.

Après une expérience complète de DAP avec reconnaissance préalable en macro-classes, une baisse des performances est observée par rapport à une modélisation directe en 39 classes (Tableau 4-15). L'écart de performance est plus grand que celui observé au niveau de la reconnaissance en macro-classes (6,2% sur le test contre 4,2% précédemment, cf. Tableau 4-13). Il est essentiellement imputable à une augmentation supplémentaire de 3% du taux d'omission qui s'explique par l'impossibilité d'identifier un certain nombre de segments (cf. §3.3.1). Pour remédier à ce fort taux d'omissions, l'identification segmentale a été remplacée par un DAP. Dans ce cas, une augmentation très importante du taux d'insertions a été constatée.

Modélisation	Phonèmes identifiés	Phonèmes substitués	Phonèmes omis	Phonèmes insérés	Phonèmes reconnus	Facteur d'échelle
Apprentissage						
Standard (39cl.)	70,3	17,0	12,7	2,6	67,7	1,5
Macro-classes	57,4	21,7	**20,9**	1,4	56,0	X
Test						
Standard (39cl.)	57,6	27,8	14,6	5,0	52,5	
Macro-classes	49,1	27,6	**23,3**	3,0	46,2	

Tableau 4-15 Taux de DAP du système HMM/50-ppv directement en 39 classes et après reconnaissance en macro-classes.

Cette étude sur l'opportunité de la division du processus de décision avec une étape préalable de reconnaissance en macro-classes nous conduit à constater que le décodage direct en classes phonétiques offre de meilleurs résultats. En effet, les erreurs introduites lors de la reconnaissance en macro-classes par un système en macro-classes par rapport au système standard en 39 classes ne sont pas compensées par l'accroissement de performance de l'identification segmentale des sous-systèmes spécialisés sur les macro-classes. Cette affirmation a pu être vérifiée au travers d'une expérience supplémentaire.

La reconnaissance en macro-classes obtenue avec le système standard en 39 classes après transformation des étiquettes a été utilisée pour l'identification segmentale séparée. Cette expérience revient à remettre en cause la classe phonétique obtenue par un opérateur global en 39 classes à l'aide d'un opérateur plus local (dans la mesure où il est spécialisé sur la macro-classe à laquelle appartient la classe reconnue). On constate alors que les 4% d'écart sur le taux de la reconnaissance en macro-classes (cf. Tableau 4-13) se retrouvent dans le résultat final en 39 classes (le taux de DAP en 39 classes est alors de 50,8%). L'amélioration des performances par cette méthode suppose donc d'augmenter la qualité de l'étape de reconnaissance en macro-classes. Or, il est vraisemblable que pour y parvenir on s'oriente vers une remise en cause de la modélisation de chacune des macro-classes dans le sens d'une meilleure prise en compte de leurs composantes internes ; par exemple en modélisant directement les 39 classes phonétiques. Cette approche nous conduirait alors à réaffirmer le choix du modèle HMM/K-ppv standard.

4.3. Conclusion

Aux travers des expériences décrites dans ce chapitre, on peut observer que les techniques traditionnellement appliquées au système HMM/GMM ne sont pas immédiatement transférables au système HMM/K-ppv. Ainsi, l'accroissement de l'espace de représentation par l'introduction de paramètres temporels (coefficients delta) ou la modélisation contextuelle, abordée sous son aspect plus général de l'adaptation des classes phonétiques, n'ont pas permis d'augmenter les performances du système

HMM/K-ppv. Elles ont conduit, au contraire, à une diminution de ses performances par rapport au système de base.

Dans le cas de la modélisation contextuelle, l'intérêt de son application au système HMM/K-ppv peut finalement être remis en cause. Même s'il est vrai que la généralisation de son emploi dans les systèmes HMM/GMM rendait incontournable son évaluation pour l'augmentation des performances du système HMM/K-ppv. Le fondement des techniques de modélisation contextuelle est de dégager efficacement les modes probabilistes des données. Dans le cadre des estimateurs paramétriques continus, chaque mode est représenté par une fdp élémentaire et les différents modes sont combinés linéairement. En l'absence d'un critère particulier, la séparation des modes pour chaque état est opérée de manière non-supervisée à l'aide d'un algorithme de type k-moyennes. L'introduction d'un critère expert basée sur une considération acoustique (en l'occurrence la co-articulation) permet une estimation plus robuste des fdp composant la somme pondérée d'un état par une séparation préalable des données en ensembles plus homogènes. Dans le cas des estimateurs non-paramétriques, l'estimation est locale et la prise en compte des modes des données est intrinsèque au processus d'estimation. L'apport de la modélisation contextuelle ne pourra donc qu'être marginal au sein du système HMM/K-ppv. En ce sens, l'introduction d'une information temporelle nous paraît d'un enjeu plus fondamental.

La défection des paramètres temporels dans le système HMM/K-ppv est surprenante. Malgré une amélioration des résultats en identification locale et segmentale, les résultats du DAP sont dégradés après l'introduction des coefficients delta dans les paramètres du système. Il semble pourtant que, moyennant une méthode d'application adaptée, l'information qu'ils contiennent devrait s'appliquer quel que soit le type d'estimateur de probabilité utilisé. L'étude de leur influence sur le système HMM/GMM n'a pas permis de dégager les éléments permettant une transposition favorable au système HMM/K-ppv. Une analyse plus complète des résultats de DAP du système de base doit donc être réalisée avant que les performances du système puissent être améliorées.

CONCLUSION

La reconnaissance de la parole continue est actuellement traitée par une modélisation probabiliste markovienne. Le recours aux modèles de Markov cachés a permis de résoudre un grand nombre des difficultés inhérentes à la reconnaissance des formes segmentale. Toutefois, sa formulation rigoureuse ne doit pas faire oublier que cette modélisation est fondée sur des hypothèses simplificatrices. Une grande part des travaux de recherche du domaine depuis une dizaine d'années consiste à lever ces hypothèses ou, au moins, à diminuer leur influence. Nous avons choisi d'orienter nos travaux vers l'amélioration de la discrimination des modèles markoviens.

Des modèles de Markov discriminants sont habituellement obtenus par de nouvelles itérations d'apprentissage des modèles initiaux à l'aide d'un critère discriminant. Contrairement à ces approches qui se sont révélées coûteuses en temps de calcul et d'utilisation complexe pour un apport finalement restreint, nous avons choisi d'introduire la discrimination dans les modèles de Markov par le biais d'un estimateur de probabilité discriminant : l'estimateur des K-ppv. Les qualités théoriques de cet estimateur ont été confirmées par des expériences d'identification locale sur la base de données TIMIT comportant plus d'un million de vecteurs acoustiques. Son application à la reconnaissance de la parole continue, rendue possible par le développement d'algorithmes de calcul rapides, a permis la mise en œuvre d'un système de base HMM/K-ppv.

Un apprentissage convergent de type EM basé sur le critère du maximum de vraisemblance a été obtenu. Les paramètres de cet apprentissage sont les coefficients d'appartenance associant chaque vecteur de l'ensemble de référence aux états markoviens des modèles. L'étude des options de l'apprentissage a conduit à la validation d'un protocole optimal. Ainsi, deux formules de réestimation ont été proposées et testées. La réestimation propagée, dont la convergence a été montrée, est retenue en combinaison avec une initialisation partagée des coefficients d'appartenance. Le système de base HMM/K-ppv a finalement pu être comparé à un système de base HMM/GMM. Les performances des deux systèmes sur une tâche de reconnaissance acoustico-phonétique sont comparables. Plusieurs orientations ont alors été envisagées pour améliorer les performances du système HMM/K-ppv.

Les traitements proposés ont porté sur deux aspects de l'opérateur de reconnaissance : la définition des classes et l'espace de représentation. Afin de diminuer la confusion entre classes phonétiques, deux approches ont été testées : la division en classes sub-phonétiques et le regroupement en macro-classes. Ces méthodes n'ont pas permis d'obtenir de meilleurs résultats de reconnaissance.

Afin d'améliorer la modélisation temporelle, l'ajout de coefficients deltas aux coefficients statiques a été testé. Les expériences montrent que l'information introduite n'est pas exploitée par l'estimateur des K-ppv. Les coefficients delta conduisent même à une dégradation des performances de reconnaissance du système HMM/K-ppv. L'étude de l'influence des coefficients delta dans le cas du système HMM/GMM n'a pas permis en l'état d'aboutir à une utilisation appropriée des coefficients delta dans le système HMM/K-ppv.

De nombreuses voies restent à explorer pour une meilleure utilisation de l'estimateur des K-ppv au sein d'un système markovien. Nous avons montré que des techniques de l'état de l'art de la modélisation markovienne comme la modélisation contextuelle ou les coefficients delta se révèlent finalement beaucoup plus liées au type d'estimation de probabilité utilisée que ne le laisse supposer leur formulation. Parmi les voies déjà abordées dans ce travail, deux doivent être, à notre avis, prolongées en priorité : l'inférence de topologie et l'utilisation des coefficients delta.

La première car, si la qualité de la représentation locale des données a augmenté avec l'utilisation de l'estimateur des K-ppv, elle reste contrainte au niveau segmental par un cadre très rigide qui ne lui est pas adapté ; en l'occurrence un modèle gauche-droit à 3 états. Le recours aux méthodes d'inférence grammaticale devrait permettre une approche non-paramétrique, et donc mieux adaptée aux données, de l'inférence de topologie. D'autre part, l'étude des coefficients delta doit être poursuivie pour deux raisons principales : tout d'abord car l'inadéquation entre les coefficients deltas et l'estimateur des K-ppv reste un résultat inattendu qu'il conviendra d'expliquer pour une meilleure connaissance de l'estimateur ; ensuite car l'introduction d'une information temporelle explicite dans le système HMM/K-ppv doit toujours être réalisée.

Enfin, l'existence de deux estimateurs de probabilité présentant des comportements différents doit être mise à profit afin de déterminer plus précisément les mécanismes réels qui régissent l'intervention des estimateurs de probabilité dans la modélisation markovienne de la parole. Pour cela, une étude plus approfondie des résultats de reconnaissance doit être réalisée ; notamment à l'aide de l'alignement des reconnaissances phonétiques de chacun des systèmes de base HMM/GMM et HMM/K-ppv.

L'étude des systèmes HMM/GMM a débuté au milieu des années 80. Son usage ne s'est répandu qu'une dizaine d'année plus tard. Des apports déterminants ont été nécessaires entre son introduction et sa généralisation (passage des mono-gaussiennes aux sommes de gaussiennes, partage de paramètres...). Cette étude constitue un premier pas vers l'utilisation de l'estimateur des K-ppv pour la RAP continue. Ses qualités théoriques, confirmées expérimentalement, permettent de le considérer comme la meilleure solution actuelle au problème de l'estimation locale pour la RAP. Notre travail doit maintenant s'orienter vers la recherche de traitements adaptés permettant de répercuter au mieux ses qualités au niveau segmental. Le fait que le système HMM/K-ppv se place d'emblée au même niveau de performance en reconnaissance phonétique qu'un système HMM/GMM dans les mêmes conditions nous encourage dans cette voie.

Annexe A REPRESENTATION DU SIGNAL DE PAROLE ET METRIQUE ASSOCIEE

La première étape de l'étude du signal de parole consiste à le représenter dans un espace réunissant deux qualités : avoir un sens en terme de reconnaissance et ne conserver que l'information pertinente. Il faut pour cela qu'il soit le moins possible sensible aux différences de locuteurs, aux canaux de transmission et aux phénomènes para-linguistiques comme l'état émotionnel du locuteur.

A.1 Calcul des paramètres

L'acquisition du signal est la première étape permettant son analyse automatique. Le signal physique continu est discretisé par échantillonnage : on ne retient ses valeurs qu'en un certain nombre de points répartis à intervalle constant. Cet intervalle est fixé par la fréquence d'échantillonnage. Un des inconvénients de l'échantillonnage est de perdre l'information contenue dans les fréquences supérieures à la moitié de la fréquence d'échantillonnage (théorème de Shannon). Ainsi pour le signal de parole, l'information utile étant contenue entre 0 et 8 kHz, il est nécessaire d'opérer un échantillonnage dont la fréquence est au moins égale à 16 kHz.

Le processus de parole est supposé quasi-stationnaire. Ceci implique que, sur un intervalle de quelques milli-secondes, ses caractéristiques spectrales sont constantes (ou varient suffisamment peu). Ainsi, la fonction principale de l'étape de représentation du signal de parole est de diviser le signal en trames qui représentent un segment stationnaire et pour chaque trame de réaliser une analyse à court-terme. Les trames sont typiquement espacées de quelques dizaines de ms (de 10 à 30 ms) et elles se chevauchent pour permettre une plus grande longueur d'analyse (généralement 25 ms). Une procédure de pré-traitement classique pourra comporter les 5 étapes suivantes :

- Filtrage passe-bas anti-repliement à 8 kHz
- Echantillonnage à 16 kHz et quantification sur 16 bits
- Pré-accentuation par un filtre d'équation $1\text{-}0.95z^{-1}$ pour compenser le rayonnement aux lèvres
- Elimination de la composante continue
- Fenêtrage de Hamming (ou Blackman) pour atténuer les discontinuités aux bords

Ainsi, si nous retenons 16 kHz comme fréquence d'échantillonnage et 10 ms pour la durée d'une trame, chaque trame sera représentée par 160 points. A partir de ces points, deux types d'analyses à court terme peuvent être menés afin de réduire le nombre de points :

- *analyse spectrale non paramétrique* : énergies et analyse spectrale discrète (*Fast Fourier Transform*, FFT [Cooley & Tukey, 1965]).
- *analyse paramétrique* : calcul des coefficients d'un modèle de prédiction linéaire (*Linear Prediction Coefficients*, LPC [Markel & Gray, 1976] ; *Perceptual Linear Predicitive*, PLP [Hermansky, 1990]), rapports des fonctions d'aire (*Log Area Ratio*, LAR).

Il existe ensuite un certain nombre de transformations supplémentaires qui peuvent être appliquées avant d'obtenir les coefficients finaux. Nous ne détaillerons que la procédure permettant l'obtention des coefficients MFCC [Davis & Mermelstein, 1980] (Figure A-1).

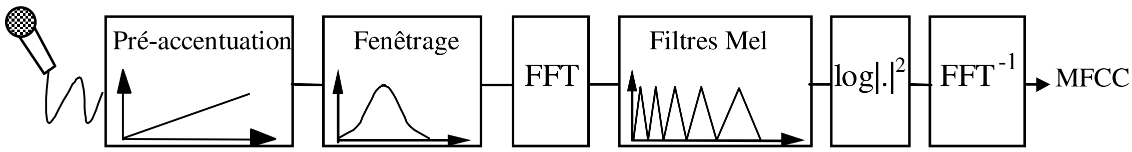


Figure A-1 Chaîne d'analyse du signal permettant l'obtention des MFCC.

Le spectre à court-terme obtenu après l'application de la FFT sur la fenêtre de signal est filtré par un banc de filtre en échelle Mel. L'intérêt de l'échelle Mel est d'être conforme à l'échelle de perception sonore et aux bandes passantes critiques de l'oreille. Ensuite, les coefficients cepstraux sont obtenus par une transformée en cosinus discrète à partir des logarithmes des énergies issues du banc de filtres. L'avantage de la transformation cepstrale est de fournir des coefficients peu corrélés et de mieux séparer l'influence du locuteur.

Il existe un grand nombre de distances pouvant être associées à l'espace des MFCC. Les distances classiques, parmi lesquelles la distance euclidienne et la distance de Mahalanobis qui normalise les coefficients par leur matrice de covariance, sont les plus utilisées. Des distances spécifiques aux espaces de représentation de la parole ont été développées : la mesure d'Itakura [Itakura, 1975] qui permet de comparer les modèles auto-régressifs, les distances cepstrales pondérées [Juang, Rabiner *et al.*, 1987] ou encore la distance par Appariement de Pics Spectraux [Caraty & Rodet, 1985].

A.2 Choix des MFCC

La représentation par MFCC est aujourd'hui la plus répandue dans les systèmes de reconnaissance. [Picone, 1993] relève que, parmi une trentaine de systèmes à travers le monde, les 2/3 des systèmes ont recours à une analyse cepstrale et parmi ceux-ci, 2/3 utilisent des MFCC, contre seulement 1/3 qui utilisent des LPCC. Ceci s'explique certainement par la plus grande simplicité d'utilisation de l'échelle Mel après une FFT qu'avec les coefficients d'un filtre auto-régressif. Néanmoins, en présence de bruit, le recours à d'autres coefficients s'est révélé nécessaire (RASTA-PLP [Hermansky & Morgan, 1994] par exemple). L'ensemble de nos travaux est basé sur l'utilisation des MFCC. Une trame de 25 ms, calculée chaque centi-seconde, est représentée par 12 MFCC plus l'énergie (MFCC+E). Les valeurs des 13 coefficients sont normalisées par leur écart-type mesuré sur l'ensemble des trames d'apprentissage.

Annexe B EVALUATION D'UN SYSTEME DE RAP

L'évaluation d'un système de RAP est obtenue en comparant le décodage d'un certain nombre de phrases de test avec un étiquetage de référence de ces phrases. On peut alors estimer un taux de réussite moyen. La précision de cet estimateur dépend du nombre d'unités testées qui permet de définir l'intervalle de confiance de l'estimation.

B.1 Calcul du taux de reconnaissance

Le décodage d'une phrase fournit une suite d'unités acoustiques. Pour évaluer la qualité du décodage et quantifier les différences avec le résultat idéalement attendu, il faut disposer d'un étiquetage de référence des phrases de test. Les deux chaînes de mots ou de symboles sont alignées par un algorithme de programmation dynamique, tel que l'algorithme de Wagner et Fisher [Wagner & 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 :

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

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 :

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

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

B.2 Intervalle de confiance

La fiabilité de l'estimation du taux de reconnaissance sur un corpus donné dépend du nombre de tests réalisés, 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. [Montacié & Chollet, 1987] propose de modéliser les réussites par une distribution binomiale. Soit *Ex* le nombre d'exemples disponibles pour le test, et P le taux de reconnaissance mesuré, l'intervalle de confiance à $x\%$ du taux de reconnaissance est donné par :

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

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 Ex 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 :

$$\left| P - \tilde{P} \right| \leq_{95\%} 1{,}96 \sqrt{\frac{P(1-P)}{Ex}} \tag{B-2}$$

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.

Dans le cas des ensembles d'apprentissage et de test de TIMIT (après extraction des segments de silence, voir §2.3.4.1), la variation de l'intervalle de confiance en fonction du taux mesuré est donnée dans la Figure B-1 dans le cas de l'identification locale et de la reconnaissance segmentale. Le paramètre des courbes de variation est le nombre d'exemples pris en compte (cf. Tableau B-1).

Tâche	Appr	Core Test
Identification locale (trames)	950.800	48.622
Rec segmentale (phonèmes)	130.908	6.733

Tableau B-1 Nombre d'exemples pour chaque ensemble d'évaluation de TIMIT en fonction de la tâche.

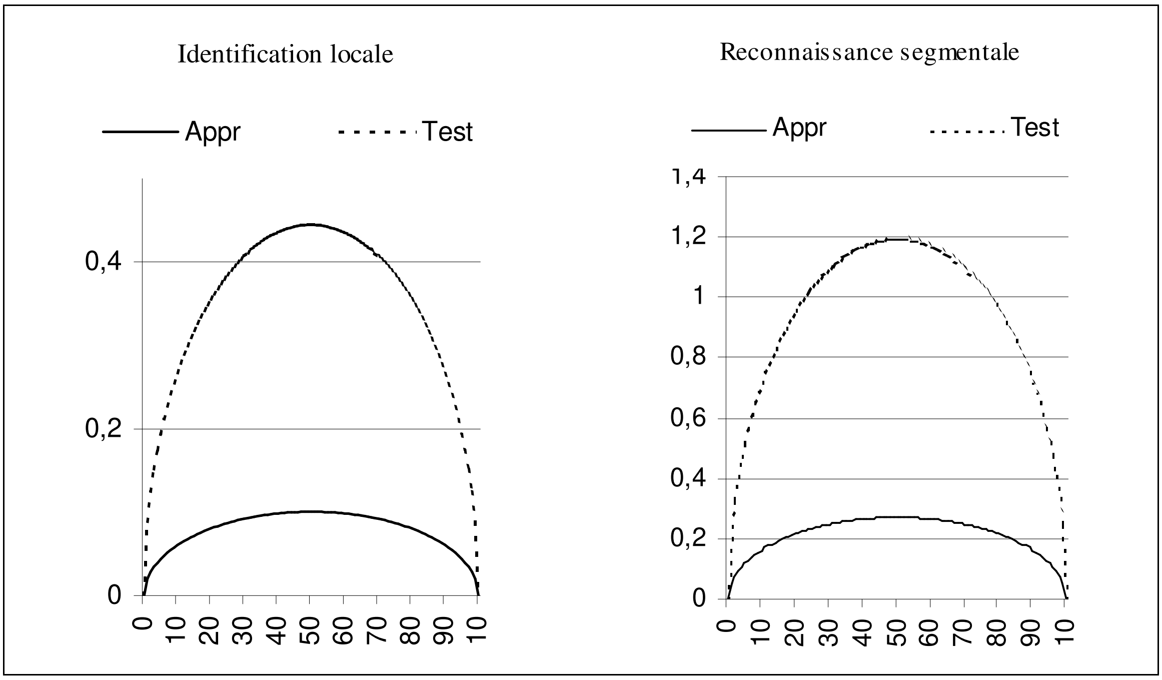


Figure B-1 Rayons de l'intervalle de confiance à 95% pour les différentes tâches et ensembles d'évaluation sur TIMIT en fonction du taux de reconnaissance constaté.

Annexe C 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. Elle est constituée d'un ensemble d'outils logiciels qui permettent de construire des systèmes de reconnaissance de la parole continue à base de HMM [Young, Jansen *et al.*, 1995].

Les modèles peuvent représenter des mots ou tout type d'unité sub-lexicale, et leur topologie peut être librement configurable. Les densités de probabilité d'émission, qui sont associées aux états, peuvent être discrètes ou continues (GMM). Les modèles sont initialisés avec l'algorithme de Viterbi, puis réestimés 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.

C.1 Présentation d'HTK

HTK, dans sa version 2.0, est structuré en 17 librairies utilisées par 18 outils de base (cf. Tableau C-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 *Perl* ou en *C-shell* si l'on travaille sous *Unix*).

Librairies		Outils de base	
HShell	Interface système d'exploitation	HLEd	Edition des fichiers d'étiquettes
HMath	Procédures mathématiques	HHEd	Edition des modèles
HSigP	Procédures de traitement du signal	HCopy	Calcul des paramètres du signal
HDBase	Stockage en mémoire des paramètres	HBuild	Formatage des modèles de langage
HSpIO	Transformations du signal	HCompV	Calcul des moyennes et variances des données d'apprentissage
HAudio	Acquisition du signal		
HWave	Gestion du signal	HDMan	Manipulation des dictionnaires
HParm	Calcul des paramètres	HParse	Génération du graphe de décodage
HVQ	Gestion de la QV	HQuant	QV pour modèles discrets
HLabel	Gestion des fichiers d'étiquettes	HSGen	Génération aléatoire de phrases test
HModel	Gestion des modèles	HSmooth	Lissage des paramètres des modèles
HTrain	Gestion de l'apprentissage	HInit	Initialisation d'un modèle
HLM	Gestion des modèles de langages	HRest	Réestimation d'un modèle
HNet	Gestion des réseaux	HERest	Réestimation des modèles enchaînés
HDict	Gestion des dictionnaires	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

Tableau C-1 Librairies et outils de base de HTK.

C.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 (Figure C-1).

C.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 MFCC ou LPCC ainsi que l'énergie. Les coefficients différentiels du premier et du second ordre sont calculés dynamiquement lors de la lecture des fichiers de paramètres, ce qui économise leur stockage permanent en mémoire. 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 offre un certain nombre de modification standard des étiquettes, par exemple le regroupement de plusieurs phonèmes dans une seule classe ou la fusion de deux segments adjacents.

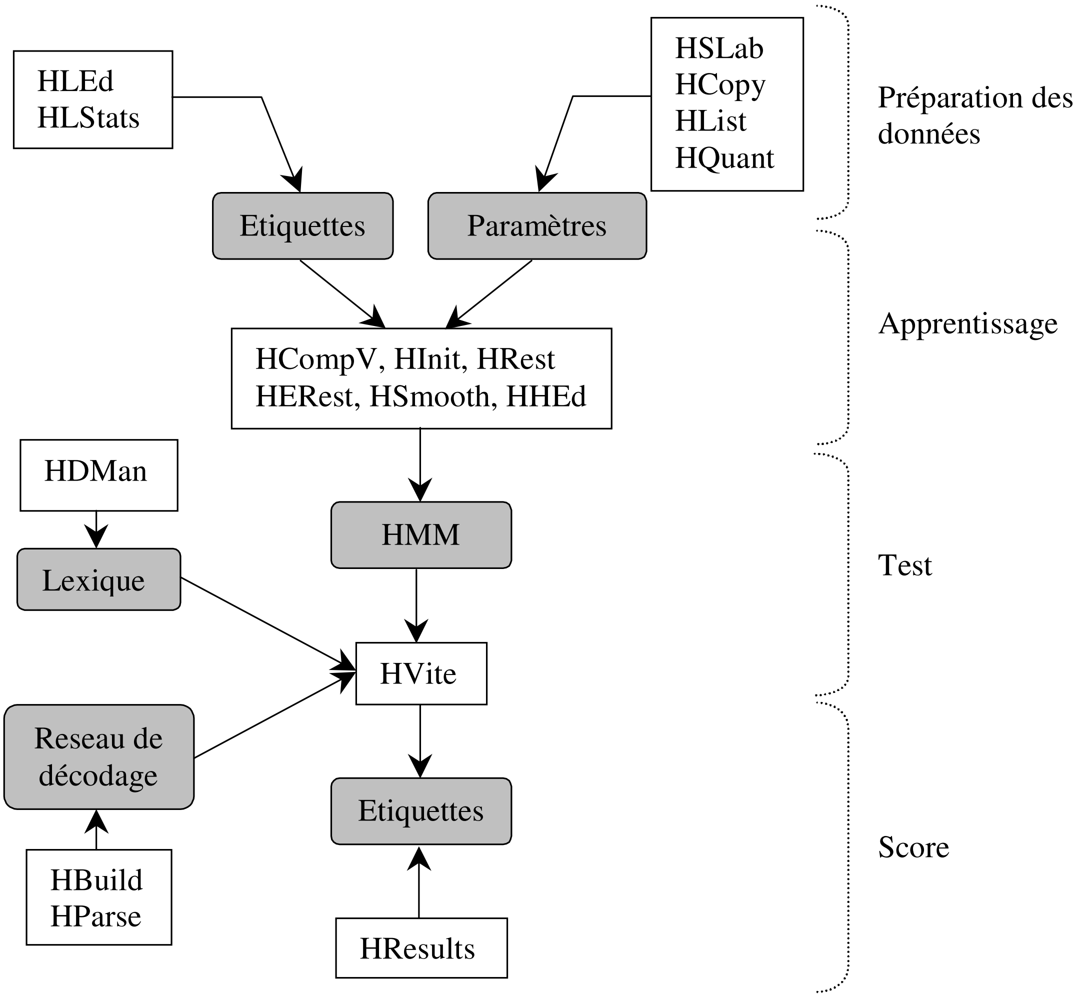


Figure C-1 Organigramme d'un système de reconnaissance avec HTK 2.0.

C.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 imposé que tous les modèles utilisent le même prototype. Les probabilités d'émission sont associées aux états et sont décrites par le tableau des probabilités d'un dictionnaire de prototype dans le cas discret ou par un GMM. 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.

C.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 réestime 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 réestimés par HRest ou HERest. HTK offre aussi des facilités pour travailler avec des modèles contextuels, dépendants des contextes phonétiques gauches ou droits.

C.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.

C.3 Adaptation de HTK à l'estimateur des K-ppv

HTK offrant un environnement complet et performant pour le développement de systèmes de reconnaissance de la parole, son utilisation pour la mise en œuvre du système HMM/K-ppv nous a semblé un bon choix. Le principe qui a guidé l'adaptation a été de conserver le plus d'outils inchangés que possible.

Les principales modifications ont concerné :

- la *définition des modèles* : elles sont maintenant contenues dans des fichiers binaires. Un fichier définit l'ensemble des modèles du système. Pour chaque modèle, les informations suivantes sont archivées : son nom, le nombre d'états, la liste des ces états (les états sont indexés globalement à tous les modèles pour permettre le partage de paramètres), et sa matrice de transition. Deux fichiers binaires permettent la mémorisation des coefficients d'apprentissage pour l'ensemble des vecteurs de référence du système. Le premier fichier donne pour chaque vecteur la liste des états sur lesquels il possède un coefficient d'appartenance non nul. Le second fichier possède la même structure mais donne la valeur du coefficient. Ces deux fichiers sont répétés pour chaque flux de données.

  L'initialisation des fichiers de modèles est réalisée à partir d'une version modifiée de HHEd.
- l'*apprentissage des modèles* : les outils permettant l'apprentissage des modèles (HInit, HRest et HERest) ont été adaptés au nouveau mode de définition des modèles. Ils intègrent de plus les mécanismes particuliers à l'estimation par K-ppv (la gestion des probabilités nulles par élimination de trames par exemple, voir §3.2.2).
- le *décodage* : seul HVite a été modifié afin d'intégrer le nouveau mode de calcul des probabilités d'émissions d'états.

Annexe D LE SYSTEME DE DICTEE VOCALE D-DAL

Nous présentons dans cette annexe le système de dictée vocale du groupe Parole du LIP6. Cette présentation a pour but de fournir un exemple de mise en œuvre des techniques markoviennes introduites dans le chapitre 1.

Le principe de l'évaluation d'un tel système est ensuite abordé au travers de la présentation des campagnes d'évaluation française de l'AUPELF-UREF et anglo-saxonne de DARPA. Finalement, les exemples de résultats de D-DAL nous permettent de discuter certaines limitations du système.

D.1 Présentation de D-DAL

D-DAL est le système du LIP6 participant à la première campagne de test de dictée vocale de l'AUPELF-UREF [Caraty, 1999]. Pour les systèmes "sans contrainte de vocabulaire", la tâche a porté sur 655 phrases (structurées en paragraphes) du journal *Le Monde* lues par vingt locuteurs ; soit 2 heures de parole. Les ressources utilisées pour l'apprentissage du système comprenaient pour la partie acoustique le corpus BREF-TOTAL (100 heures de parole) et pour la partie linguistique un corpus textuel composé des archives électroniques des dix dernières années du journal *Le Monde* (du 01/01/87 au 02/11/96).

D-DAL est un système de reconnaissance de parole continue, sans contrainte de vocabulaire (au-delà d'une centaine de milliers de mots) et indépendant du locuteur. Les modèles acoustiques sont des modèles de Markov cachés dont les densités de probabilité sont estimées par des GMM. Ces modèles sont appris en contexte. Ils sont ensuite spécialisés par genre (masculin/féminin). Les modèles de langage statistiques des n-grammes et n-classes utilisés par le système sont les bigrammes, les trigrammes, les biclasses et les triclasses.

Dans sa version 96, D-DAL est organisé en quatre phases de reconnaissance :

- Typologie par *genre du locuteur* de test et sélection des modèles acoustiques correspondants.
- Sélection d'un *lexique dynamique* de 40-k formes (lexicalement, acoustiquement et grammaticalement différentes) par DAP du signal de test et analyse du treillis phonétique de profondeur 3.
- *Décodage* par l'algorithme du Passage de Jetons engendrant un graphe de mots centi-seconde de profondeur 60 utilisant le modèle de langage des bigrammes et des biclasses avec le traitement des liaisons et des pauses. Le graphe de mots comporte

en moyenne 600 000 arcs et permet d'engendrer potentiellement 600^{60} hypothèses différentes de phrases pour une durée type de 10 secondes.

- *Fouille asynchrone* du graphe de mots en utilisant le modèle de langage des trigrammes et des triclasses pour donner l'hypothèse textuelle de D-DAL.

Pour chaque phrase de test à décoder, un sous-lexique de reconnaissance, appelé lexique dynamique, est sélectionné (passe 2) à partir du DP-L (Dictionnaire Phonétisé du LIP6) constitué de plus d'un million de formes lexicalement et syntaxiquement différentes [Montacié, Caraty *et al.*, 1997]. Nos expériences sur les taux de couverture des tests de la dictée vocale de l'AUPELF-UREF ont montré qu'un lexique dynamique de 40k-mots correspond à un lexique de 100k-mots les plus fréquents dans le corpus textuel d'apprentissage [Caraty, Montacié *et al.*, 1997]. Ainsi, il est possible de réaliser une tâche de type sans contrainte de vocabulaire (centaines de milliers de mots) à l'aide d'un moteur de reconnaissance grand vocabulaire. Cette approche a permis de développer notre système sur des machines grand public. Dans ces conditions, la durée de décodage d'une phrase de dix secondes peut être estimée à 2h30 sur un PC Pentium-Pro à 200 MHz.

D.2 Intégration d'un modèle de langage

Le développement de D-DAL a nécessité un effort particulier pour l'introduction d'un modèle de langage dans le moteur de reconnaissance du système [Lefèvre, 1995]. Les modèles des bigrammes et des biclasses sont introduits lors du décodage. Les modèles des trigrammes et des triclasses sont intégrés au travers d'une fouille asynchrone du graphe de mot obtenu par le décodage.

D.2.5 Intégration synchrone

Le moteur de reconnaissance grand vocabulaire de D-DAL a été conçu à partir de la plate-forme HTK. L'algorithme de décodage implémenté au sein de HTK est l'algorithme du passage de jeton. Notre participation a donc consisté à permettre l'intégration de modèles de langage lors du décodage et pour cela a nécessité diverses optimisations : la gestion dynamique de la structure de recherche et l'optimisation des accès au modèle de langage.

Dans la version distribuée de HTK, l'utilisation des bigrammes et des biclasses supposait la création d'un réseau de décodage comprenant un arc valué entre tous les mots. Dans le cas d'un lexique 40k-mots, cela représente 1,6 milliard d'arcs. Or toutes les suites de mots ne sont pas possibles. De plus, cette solution nécessite une capacité mémoire qui n'est pas envisageable sur les machines utilisées. Pour y remédier, nous avons modifié l'algorithme afin qu'il value les arcs entre mots dynamiquement. La structure de recherche comporte une boucle sur l'ensemble des mots du vocabulaire qui passe par deux nœuds communs (Figure D-1). Lorsqu'un jeton est propagé entre ces deux nœuds, il est automatiquement prolongé vers l'ensemble des nœuds de début de mot. L'algorithme value alors l'arc de la valeur du bigramme correspondant au dernier mot franchi par le jeton suivi du mot vers lequel il est prolongé. Cette méthode réduit la place mémoire nécessaire pour le décodage mais augmente le coût de calcul de l'algorithme.

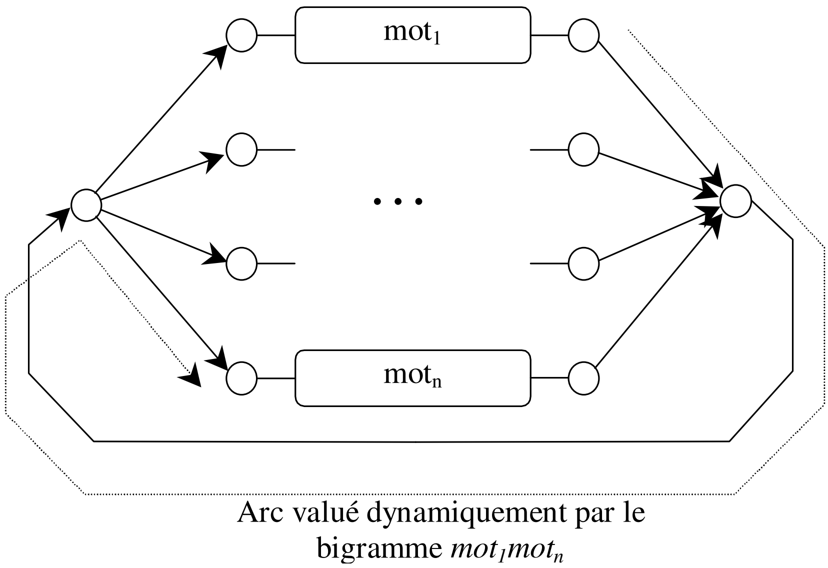


Figure D-1 Structure de recherche pour la valuation dynamique du bigramme.

Le coût de la valuation dynamique des arcs est diminué en deux étapes. Une étape préalable évite la recherche des bigrammes dont on peut prévoir qu'ils ne propageront pas de jetons maximaux. Le principe est de déterminer le seuil minimal de la probabilité des chemins prolongés. Ce seuil est estimé par le prolongement du meilleur jeton en fin de mot par son plus mauvais bigramme. Tous les jetons qui, associés à leur meilleur bigramme, ont une probabilité inférieure à ce seuil sont d'emblée écartés. La seconde étape consiste pour les jetons restants à comparer leur probabilité en fin de mot au meilleur jeton déjà enregistré pour le début de mot. La recherche du bigramme concerné est alors évitée. Enfin, le recours à des tables de hachages permet d'accélérer la recherche des valeurs des bigrammes et des biclasses. Ces optimisations ont permis de gagner un facteur 100 en temps de calcul.

D.2.6 Intégration asynchrone

L'intégration dans le système de modèles de langage plus évolués que les bigrammes et biclasses nécessite la réduction du nombre de solutions étudiées. Pour ce faire, on a recours à un graphe d'hypothèses constitué des N meilleures solutions du décodage pour chaque trame de la phrase-test. Nous avons développé l'algorithme des probabilités conjointes acoustico-linguistiques afin de réévaluer ce graphe.

Génération de graphes de mots

L'intégration d'un modèle de langage plus évolué que le bigramme (le trigramme, en l'occurrence) ne peut être envisagée dans l'algorithme de propagation de jetons sans une augmentation considérable du coût de calcul. Toutefois, le bigramme est un modèle de langage assez succinct dans le sens où sa portée, limitée à deux mots, l'empêche de modéliser les dépendances à plus long terme entre les mots de la phrases. Les modèles de langages tenant compte des modélisations à long terme (voir §1.4.1.3) présentant des mises en œuvre assez complexe, nous avons décidé dans un premier temps d'avoir recours au modèle des trigrammes et des triclasses.

L'intégration du modèle des trigrammes dans l'algorithme de décodage est rendue possible par le biais d'un graphe de mots dont la profondeur N peut être choisie indépendamment de la taille du lexique. Ce graphe est construit lors du décodage par l'algorithme PJ (§1.4.4) en mémorisant, à chaque instant *t* du décodage, les N meilleurs jetons de fin de mot, représentant les N meilleurs sous-chemins (Figure D-2). Il est ensuite réévalué de manière asynchrone par l'algorithme d'intégration des probabilités conjointes acoustico-linguistiques.

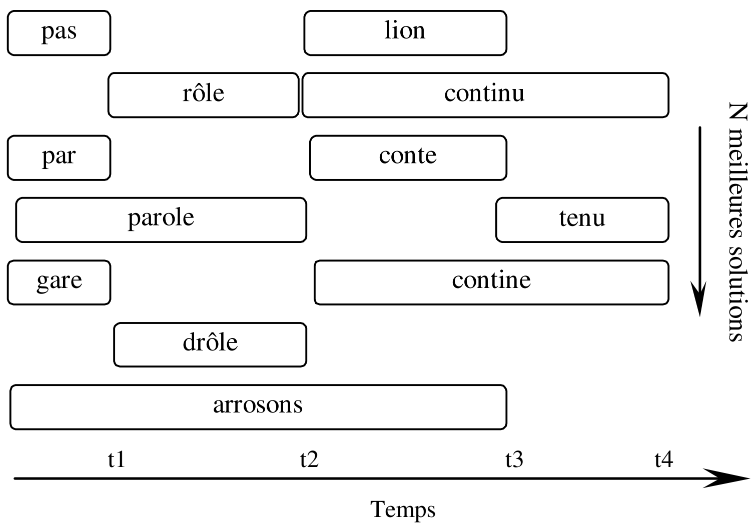


Figure D-2 Exemple de graphe de mots.

Algorithme d'intégration des probabilités conjointes

Les modèles de langage sont utilisés pour réévaluer de manière asynchrone les graphes de mots. L'algorithme décrit ici pour les trigrammes et les triclasses du langage est généralisable aux n-grammes et n-classes, avec une augmentation en conséquence de sa complexité. Le graphe de mots est composé de N mots par trame. Les mots ont été mémorisés sur leur trame finale et possèdent un pointeur sur leur trame initiale (Figure D-3). A la trame *t*, pour chaque mot on peut associer les N "meilleurs" mots l'ayant précédé : ce sont les N meilleurs mots mémorisés sur la trame initiale du mot courant. Ainsi, pour chaque trame, on construit ainsi un tableau de N^2 couples de mots.

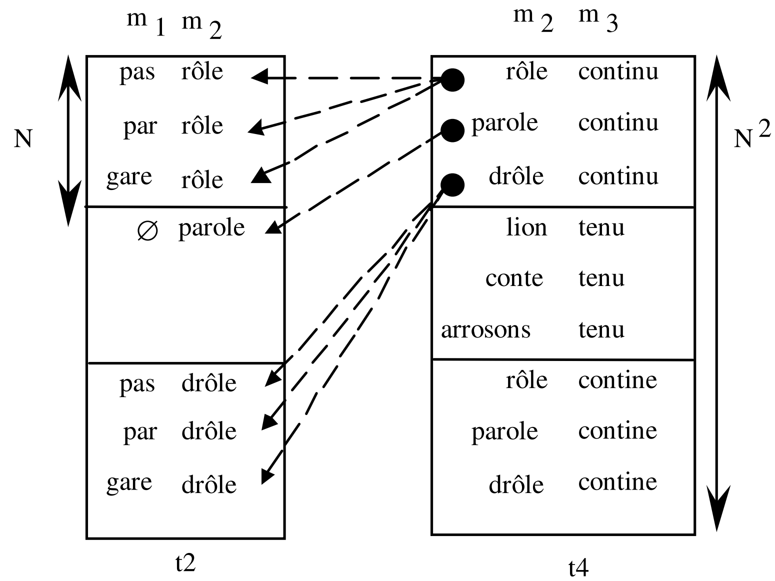


Figure D-3 Illustration de l'algorithme de calcul des probabilités conjointes.

La Figure D-3 donne les tableaux ainsi construits pour les trames correspondant aux temps t_2 et t_4 de l'exemple de graphe de mots illustré à la Figure D-2. Considérons le tableau des couples de mots m_2m_3 (temps t_4), pour chaque couple m_2m_3 on estime le meilleur mot précédent parmi les N mots m_1 possibles avant m_2 (Figure D-3). Le "meilleur" mot est obtenu par comparaison des probabilités conjointes acoustique-modèle de langage. Les probabilités conjointes sont calculées comme le produit de quatre facteurs : la probabilité conjointe jusqu'à m_2, la probabilité acoustique associée à m_3, la probabilité du trigramme $m_1m_2m_3$ et la probabilité du triclasse $C(m_1m_2m_3)$:

$$P_c(m_2m_3) = \max_{m_1} P_c(m_1m_2) \times P_a(m_3) \times P_{ML}(m_1m_2m_3) \times P_{ML}(C(m_1m_2m_3))$$

Un lien est alors crée du couple m_2m_3 vers le couple m_1m_2. On itère ce processus de la première à la dernière trame par programmation dynamique. La meilleure phrase décodée est alors reconstruite à partir du couple de la dernière trame ayant la plus forte probabilité conjointe. La complexité de cet algorithme est en $O(N^3)$.

D.3 Evaluation du système de dictée vocale D-DAL

La campagne d'évaluation de systèmes de dictée vocale de l'AUPELF-UREF de 1997 fut la première menée en langue française. Les évaluations sponsorisées par l'agence DARPA permettent depuis une dizaine d'années de mesurer l'évolution des performances des systèmes en langue anglaise. Après la description des paradigmes d'évaluation, un certain nombre de résultats du système D-DAL permettent de présenter les différents types d'erreurs rencontrés.

D.3.7 Paradigmes d'évaluation

De nombreux progrès ont été le résultat de l'établissement de standards pour l'évaluation des performances des systèmes de reconnaissance de la parole continue grand vocabulaire. Il y a seulement une décennie, les chercheurs développaient et testaient leurs systèmes sur des données collectées localement, et ne délimitaient pas toujours clairement les ensembles de développement et de test. En conséquence, il était difficile de comparer les performances des différents systèmes. La mise à la disposition de gros corpus de données couplée à la spécification de standards d'évaluation a permis d'uniformiser la présentation des résultats.

Campagnes d'évaluation anglo-saxonnes

Le paradigme d'évaluation de reconnaissance de la parole continue (*Continuous Speech Recognition Evaluation*) a été expérimenté pour la première fois aux Etats-Unis dans le cadre des projets de l'agence DARPA. Les premières campagnes débutèrent en 1971 et se poursuivent jusqu'à maintenant. Si l'on s'intéresse aux dix dernières années, trois périodes peuvent être distinguées :

- Période 1987-1992 : *Ressource Management*
  Durant cette période, les spécifications peuvent être résumées par : reconnaissance de parole lue indépendante du locuteur avec un vocabulaire fermé de 1000 mots (*i.e.* seuls ces mots apparaissent dans le test). Les données utilisées pour l'apprentissage, le développement et le test étaient issues de la base de données *Ressource Management*.
  L'utilisation des ressources était contrainte. Le corpus acoustique d'apprentissage, le lexique de reconnaissance et le modèle de langage étaient imposés. L'objectif était de focaliser le développement des systèmes sur la modélisation acoustique. Le taux d'erreur en mots chuta de 20 % en 1987 à 4% en 1992. Ce résultat fût obtenu à l'aide d'une grammaire de paire de mots qui réduit la perplexité de la tâche à 60. Les HMM deviennent la base des systèmes de l'état de l'art.
- Période 1992-1996 : *Wall Street Journal*
  Différents corpus seront utilisés durant cette période : *The Wall Street Journal* (1992-1993), puis les *North American Buisiness News* (1994-1995, le WSJ plus d'autres journaux économiques). De 1992 à 1993, la taille du lexique s'accrut de 5k à 20k mots ; l'utilisation des ressources était contrainte pour permettre des tests

contrastifs.

Depuis cette période, les évaluations sont en partie sans contraintes, le lexique a atteint 65k-mots, la complexité de la tâche a augmenté (*i.e.*, téléphone, microphone connu ou inconnu). L'élément principal de l'évaluation est le hub-test H1 sur lequel tous les participants doivent essayer leur système. Ce test est découpé en 2 parties principales : H1-C1 qui fixe les données d'apprentissage et le modèle de langage (trigramme de 20k mots appris sur des nouvelles représentant 227 millions de mots) ; et H1-P0 qui laisse libre l'ensemble des ressources. De plus, pour H1-C1, les phrases doivent être reconnues indépendamment tandis que les changements de locuteurs sont indiqués pour H1-P0, permettant ainsi une adaptation au locuteur incrémentale non-supervisée.

Le meilleur système obtenait 10,5% d'erreurs sur H1-C1 et 7,2% sur H1-P0 avec un lexique 65k, un modèle 4-grammes et une adaptation au locuteur.

- Période 1996-1999 : *Broadcast News*

  Cette évaluation consiste à transcrire des nouvelles radio-télédiffusées. Son intérêt est de combiner un ensemble significatif d'éléments réduisant les performances des systèmes de RAP : bruits de fonds, qualité téléphonique, juxtaposition de musique, parole plus spontanée… La prise en compte de ces éléments a conduit à la définition de 7 conditions (Tableau D-1) permettant de catégoriser l'ensemble des segments de parole contenus dans la base de données.

  La tâche a été séparée en deux types d'évaluation : Partionned and Unpartitionned Evaluation (PE & UE). Pour UE, les données consistent en des épisodes bruts, pour PE les épisodes sont préalablement découpés en segments homogènes (même locuteur, même conditions d'enregistrement…). Ces segments varient d'1 seconde à plusieurs minutes, ce qui obligera beaucoup des participants à re-segmenter ceux-ci pour satisfaire les contraintes d'entrée-sortie de leur système (par exemple <30s pour HTK).

  Les données acoustiques d'apprentissage représentent 35 heures d'enregistrement, inégalement réparties entre les conditions (12 h pour F0, 16mn pour F5). Les données textuelles mises à disposition par le LDC représentent 132 millions de mots. Les données d'évaluation représentent 4 demi-heures de programme (2 de télévision, 2 de radio). Le taux de mots hors vocabulaire est estimé à 0,7% et la perplexité à 172 par l'équipe d'IBM (ces éléments dépendent du lexique et du modèle de langage utilisés).

  L'approche la plus courante consiste à segmenter le flux de données en segments dont la condition est connue ; puis des techniques de modélisation pour chacune des classes sont utilisées. Par exemple, pour UE, les segments de musique détectés sont tout simplement écartés, non décodés, les segments identifiés comme de la parole de qualité téléphonique sont décodés par un système entraîné sur de la parole à bande limitée téléphonique. Les différences principales entre systèmes résident dans la façon dont ils modélisent chacune des conditions. Seule l'équipe de BBN propose un système qui ne tient pas compte des conditions et subit une adaptation non-supervisée. Le taux d'erreurs moyen est de l'ordre de 20%.

Conditions	Description	% test
F0	Parole de référence (propre)	29,7
F1	Parole spontanée (propre)	32,7
F2	Parole mauvaise qualité	8,7
F3	Parole sur fond musical	7,0
F4	Parole en conditions acoustiques dégradées	9,1
F5	Locuteurs d'autres langues	1,5
FX	Toutes les autres	11,4

Tableau D-1 Pourcentage de l'ensemble de test du Hub 4 de 1997 pour chacune des conditions.

Campagne d'évaluation française

La première campagne d'évaluation de dictée vocale a eu lieu en 1997. Elle s'inscrivait dans le cadre plus large des ARC (*Actions de Recherches Concertées*) du Réseau FRANCIL (*Réseau Francophone de l'Ingénierie de la langue)* de l'*AUPELF-UREF* qui concernaient aussi bien le traitement du langage écrit qu'oral, dans des domaines aussi variés que l'accès en langage naturel à l'information textuelle, l'alignement de corpus multilingues, la création automatique de base de données terminologiques, la compréhension de messages, la dictée vocale, le dialogue oral ou la synthèse orale de texte [Mariani, 1998]. L'évaluation a porté sur le problème spécifique de la reconnaissance de parole continue provenant de textes de journaux lus, en mode indépendant du locuteur.

Cinq laboratoires ont participé aux tests : le CRIM (Montréal), le LORIA (ex-CRIN de Nancy), l'INRS (Montréal), le LIMSI (Orsay) et le LIP6 (Paris). Deux catégories de systèmes ont été définies, reposant sur la taille du lexique utilisé : les systèmes 20 k-mots (trois premiers laboratoires) et "sans contraintes de vocabulaire" (LIMSI et LIP6). Un certain nombre de ressources ont été distribuées aux participants : deux corpus acoustiques (BREF-TOTAL, corpus d'apprentissage et ILORB1_DEV, corpus de développement), le corpus textuel de deux années (87 et 88) du journal *Le Monde* et les modèles de langage (bigrammes et trigrammes) associés et, finalement, un lexique phonétisé des 20 k-mots les plus fréquents dans le corpus textuel précédent.

Tous les systèmes étaient fondés sur une recherche multi-passes [Adda, Adda-Decker *et al.*, 1997; Caraty, Montacié *et al.*, 1997; Fohr, Haton *et al.*, 1997; Savariaux, Farhat *et al.*, 1997]. Au minimum, une première passe produit un graphe de mots à l'aide de modèles acoustiques et linguistiques simples ; alors qu'une ou plusieurs passes successives permettent l'application de modèles plus élaborés. Parmi ces passes : l'adaptation non supervisée, l'utilisation d'un modèle grammatical formel ou la sélection d'un lexique dynamique. Deux algorithmes de décodage sont utilisés pour générer un graphe de mots : algorithme du Viterbi-Bloc (génération de proposition à la fin d'un bloc et non à la fin de la phrase) et algorithme de type *One-Pass* (cf. §1.4.3.2). Les modèles acoustiques sont tous fondés sur une modélisation markovienne d'ordre 1 ou 2. L'élaboration principale des modèles acoustiques se fait par la prise en compte du contexte ; ainsi des monophones aux triphones, en passant par les biphones, pour finir par des triphones dépendant de leur position dans la phrase. Les modèles de langages évoluent des bigrammes ou bigrammes/biclasses vers les trigrammes ou trigrammes/triclasses.

D.3.8 Exemples de résultats

L'évaluation d'un système de reconnaissance est réalisée en comparant le décodage obtenu sur les phrases de test avec l'étiquetage de référence de ces phrases. La précision de cette estimation dépend du nombre de tests réalisés. Le taux d'identification du système D-DAL sur les 655 phrases de test est de 69% avec un taux d'insertions de 8%. Nous donnons en exemple quelques phrases décodées lors de la campagne de test.

La première phrase constitue la référence, la seconde l'hypothèse textuelle de D-DAL. Elles donnent un échantillon des erreurs commises par le système. Il est difficile, voire impossible, de déterminer avec certitude le niveau du système responsable de l'erreur. Ceci est du à l'utilisation d'un processus de décision globale qui utilise simultanément les niveaux acoustiques, lexicaux et syntaxiques. Il est toutefois possible, dans certains cas, de déterminer le niveau qui a la plus grande part de responsabilité dans l'erreur.

<u>Erreurs acoustiques</u>

REF :	les	deux	victimes		travaillaient	pour	la	cause	tamoule
HYP :	les	deux	villes	qui	travaillait	pour	la	grosse	dame

Voici un exemple de phrase pour laquelle l'acoustique est mal reconnue. Si l'on décomposait les mots en phonèmes, on dénombrerait des d'insertions (/l/ entre /i/ et /k/ dans *victimes* qui conduit à reconnaître *villes qui*), des omissions (/m/ et /t/ de *victimes*, /u/ et /t/ de *tamoule*) et des substitutions (/s/ de *grosse* pour /z/ de *cause* ou /d/ de *dame* pour /t/ de *tamoule*). Lorsque, le taux de reconnaissance en décodage acoustico-phonétique devient trop faible, les contraintes lexicales (réseaux de modèles acoustiques) ne suffisent plus à compenser les erreurs et peuvent même les amplifier.

<u>Erreurs n-grammes</u>

REF :	Louis	Viannet		poursuit	une	stratégie
HYP :	Louis	Viannet	et	poursuit	une	stratégie

Dans cette phrase, tous les phonèmes sont corrects, l'acoustique est respectée. C'est l'intervention du modèle de langage des n-grammes qui produit l'erreur. Le trigramme *Louis Viannet poursuit* n'existe pas. Le système a donc recours au bigramme *Viannet poursuit*. Or celui-ci est plus faible que le produit des bigrammes *Viannet et* et *et poursuit*. Le système commet alors une erreur d'insertion.

REF :	son	premier	film	en	mille	neuf	cent	cinquante
HYP :	son	premier	film	en	mille	neuf	cent	cinquante
REF :	huit	après	avoir	rejoint		la	Nikkatsu	
HYP :	huit	après	avoir	rejoint	année	quatre	sous	

La raison de cette erreur est plus simple que la précédente : *Nikkatsu* est un mot hors vocabulaire. On vérifie alors que la méconnaissance d'un mot ne se limite pas à une erreur mais a tendance à se propager (en moyenne, un mot hors vocabulaire, lorsqu'il n'est pas détecté, conduit à 2,5 erreurs).

REF :	selon	Peter	Kessler	du	Haut	Commisariat	des	Nations
HYP :	selon	Peter	Kessler	du	haut	commissariat	des	nations
REF :	Unies	pour	les	réfugiés				
HYP :	unies	pour	les	réfugiés				

Il s'agit d'une erreur d'emphase dans ce cas les suites de mots *haut commissariat* et *nations unies* doivent être considérées comme des expressions et mises en majuscule.

<u>Erreurs n-classes</u>

REF :	une	action			solitaire	serait	un		échec
HYP :	une	action	sur	les	affaires	seraient	à	l'	échec

Cet exemple illustre l'effet de la propagation des erreurs. Le système fait une première erreur acoustique en reconnaissant *sur les affaires* au lieu de *solitaire*. Le décodage acoustico-phonétique est ensuite correct puisque *serait* et *seraient* sont homophones (*i.e.*, ils sont composés des mêmes phonèmes). Mais, dans ce cas, c'est le modèle de langage des n-classes qui induit le système en erreur en réalisant l'accord pluriel avec *affaires*.

Erreurs sémantiques

REF :	un	état		d'esprit	plutôt	
HYP :	un	état	des	esprits	plus	tôt

Ce dernier exemple est révélateur des limitations du système. En effet, ne prenant pas en compte de niveau sémantique, aucune information ne lui permet de se décider entre *plutôt* et *plus tôt*.

REF :	les	nationalistes	eux-mêmes	disent	leur	inquiétude
HYP :	les	nationalistes	eux-mêmes	disent	leurs	inquiétudes
REF :	face	à	cette	situation		
HYP :	face	à	cette	situation		

Cet exemple illustre les limitations des capacités humaines : sans plus d'informations sur la nature du discours des *nationalistes* il n'est pas possible de déterminer s'ils présentent une *inquiétude* (généralité) ou des *inquiétudes* (particulières). Cette information aurait pu être tirer de l'absence de liaison entre *leur* et *inquiétude* si celles-ci étaient obligatoires en français ; ce qui n'est pas le cas.

Annexe E BASE DE DONNEES TIMIT

E.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. Un sous-ensemble, appelé noyau, a été extrait de l'ensemble de test original. Il contient 24 locuteurs : deux hommes et une femme par "dialecte". Les phrases de calibration étant exclues, ces 24 locuteurs prononcent 192 phrases. Le noyau de test comprend 7215 phonèmes. Sa taille réduite a permis de multiplier les expériences tout en conservant une estimation assez fiable des taux de reconnaissances réels (cf. Annexe B).

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 E-1 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	m̩	bottom	ih	ɪ	bit
kcl k	k	key	n	n	noon	ix	ɨ	debit
bcl b	b	bee	nx	ɾ̃	winner	eh	ɛ	bet
dcl d	d	day	en	n̩	button	ae	æ	bat
gcl g	g	gay	ng	ŋ	sing	aa	ɑ	bott
dx	ɾ	muddy	eng	ŋ̍	washington	ao	ɔ	bought
q	ʔ	bat	Liquides:			uh	ʊ	book
Affriquées:			l	l	lay	uw	u	boot
dcl jh	ǰ	joke	el	l̩	bottle	ux	ü	toot
tcl ch	č	choke	r	r	ray	ax	ə	about
Fricatives:			Semi-voyelles:			ax-h	ə̥	suspect
f	f	fin	w	w	way	ah	ʌ	but
th	θ	thin	y	y	yacht	er	ɝ	bird
s	s	sea	Fricative glottale:			axr	ɚ	butter
sh	š	she	hh	h	hay	Diphtongues:		
v	v	van	hv	ɦ	ahead	Ey	e^y	bait
dh	ð	then	Silences:			Ay	$ɑ^y$	bite
z	z	zone	h#			oy	$ɔ^y$	boy
zh	ž	azure	pau			aw	$ɑ^w$	bout
			epi			ow	o^w	boat

Tableau E-1 Etiquetage de TIMIT, code API correspondant et exemple de mot anglais contenant le phonème.

E.2 Regroupements 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]. 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.

Le Tableau E-2 donne 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.

Etiquette	Nombre	Durée (ms)	Etiquette	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 E-2 Statistiques sur le nombre de représentants et la durée moyenne des 48 classes phonétiques (les confusions autorisées au décodage sont désignées par une accolade).

BIBLIOGRAPHIE

Note : les publications disponibles en ligne sur Internet au moment de la rédaction de ce manuscrit sont signalées par la mention *web*. Les URL étant soumise à de fortes variations, elles ne sont pas données (une recherche sur le nom de l'auteur par un moteur de recherche de type *AltaVista* se révélera souvent suffisante).

Adda, G., M. Adda-Decker, J.-L. Gauvain & L. Lamel (1997). "Le système de dictée du LIMSI pour l'évaluation AUPELF'97". *1ères Journées Scientifiques et Techniques AUPELF-UREF*, Avignon, pp. 35-40. *web*

Aikawa, K., H. Singer, H. Kawahara & Y. Tohkura (1993). "A Dynamic Cepstrum Incorporating Time-frequency Masking and its Application to Continuous Speech Recognition". *ICASSP*, Minneapolis, pp. II.668-671.

André-Obrecht, R. (1988). "A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals" *IEEE Transactions on Acoustics, Speech and Signal Processing* **36**(1):29-40.

Austin, S., R. Schwartz & P. Placeway (1991). "The Forward-Backward Search Algorithm". *ICASSP*, Toronto, pp. II.697-700.

Austin, S., G. Zavaliagkos, J. Makhoul & R. Schwartz (1992). "Speech Recognition Using Segmental Neural Nets". *ICASSP*, San Fransisco, pp. I.625-628.

Bahl, L. (1990). *Apparatus and Method of Grouping Utterances of a Phoneme into Context-dependent Categories Based on Sound Similarity for ASR* Patent of IBM, USA.

Bahl, L., P. Brown & P. de Souza (1993). "Estimating Hidden Markov Model Parameters so as to Maximize Speech Recognition Accuracy" *IEEE Transactions on Speech and Audio Processing* **1**(1):77-83.

Bahl, L., P. Brown, P. de Souza & R. Mercer (1987). "Speech Recognition with Continuous-parameter Hidden Markov Models" *Computer Speech and Language* **2**:219-234.

Bahl, L., P. Brown, P. de Souza & R. Mercer (1988). "A New Algorithm for the Estimation of Hidden Markov Models Parameters". *ICASSP*, New-York, pp. .

Bahl, L., P. Brown, P. de Souza & R. Mercer (1989). "A Tree-based Statistical Language Model for Natural Language Speech Recognition" *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**(7):1001-1008.

Bahl, L., P. Brown, P. de Souza, R. Mercer & M. Picheny (1993). "A Method for the Construction of Acoustic Markov Models for Words" *IEEE Transactions on Speech and Audio Processing* **1**(4):443-452.

Bahl, L., P. de Souza, P. Gopalakrishnan, D. Nahamoo & M. Picheny (1991). "Decision Trees for Phonological Rules in Continuous Speech". *ICASSP*, Toronto, pp. I.185-188.

Baker, J. (1975). "The DRAGON System-An Overview" *IEEE Transactions on Acoustics, Speech and Signal Processing* **23**(1):24-29.

Bakis, R. (1976). "Continuous Speech Recognition via Centisecond Acoustic States". *91'th Meeting of the Acoustical Society of America*, pp. .

Barras, C. (1996). *Reconnaissance de la parole continue : adaptation au locuteur et contrôle temporel dans les modèles de Markov cachés*. Thèse de Doctorat Université Pierre et Marie Curie, Paris. *web*

Baum, L. (1970). "A Maximization Technique Occuring in the Statistical Analysis of Probabilistic Functions of Markov Chains" *Annals of Mathematical Statistics* **41**:164-171.

Baum, L. (1972). “An Inequality and Association Maximization Technique in Statistical Estimation for probabilistic Functions of Markov Processes” *Inequalities* **3**:1-8.

Bellagarda, J. & D. Nahamoo (1990). “Tied Mixture Continuous Parameter Modeling for Speech Recognition” *IEEE Transactions on Acoustics, Speech and Signal Processing* **38**(12):2033-2045.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.

Belluardo, E. (1997). *Data mining et inférence de topologie*. Rapport de stage Université Pierre et Marie Curie, Paris.

Bengio, Y., R. de Mori & G. Flammia (1992). “Global Optimization of Neural Network-Hidden Markov Model Hybrid” *IEEE Transactions on Neural Networks* **3**(2):252-259.

Bengio, Y., R. de Mori, G. Flammia & R. Kampe (1991). "A Comparative Study on Hybrid Acoustic Phonetic Decoders based on Artifical Neural Networks". *Eurospeech*, Gênes, pp. II.1007-1010.

Black, E. (1993). *Satistically-Driven Computer Grammars of English : The IBM Lancaster Approch*. Rodopi, Amesterdam-Atlanta.

Black, E., F. Jelinek, J. Lafferty, D. Magerman & S. Roucos (1992). "Towards History-based Grammars : Using Richer Models for Probabilistic Parsing". *DARPA Spoken Language Systems Workshop*, pp. .

Bocchieri, E. & J. Wilpon (1993). “Discriminative Feature Selection for Speech Recognition” *Computer Speech and Language* **7**:229-246.

Bourlard, H., N. Morgan & S. Renals (1992). “Neural Nets and Hidden Markov Models : Review and Generalizations” *Speech Communication* **11**(2/3):237-246.

Bourlard, H. & C. Wellekens (1989). “Speech Pattern Discrimination and Multilayer Perceptrons” *Computer Speech and Language* **3**:1-19.

Bourlard, H. & C. Wellekens (1990). “Links Between Markov models and Multilayer Perceptrons” *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(12):1167-1178.

Bridle, J. (1990). “Alpha-nets : a Reccurrent 'Neural' Network Architecture with a Hidden Markov Model Interpretation” *Speech Communication* **9**(1).

Bridle, J., M. Brown & R. Chamberlain (1982). "An Algorithm for Connected Word Recognition". *ICASSP*, Paris, pp. II.899-902.

Broder, A. (1990). “Strategies for Efficient Incremental Nearest Neighbor Search” *Pattern Recognition Letters* **23**(1/2):171-178.

Brugnara, F., R. de Mori, D. Giuliani & M. Omologo (1992). "Improved Connected Digit Recognition Using Spectral Variation Functions". *ICSLP*, Banff, pp. I.627-630.

Bunke, H., M. Roth & E. G. Schukat-Talamazzini (1995). “Off-line cursive handwriting recognition using hidden Markov models” *Pattern Recognition* **28**(9):1399-1413.

Cacoullos, T. (1966). “Estimation of a Multivariate Density” *Annals of Inst. of Stat. Math.* **18**:179-189.

Calliope (1989). *La parole et son traitement automatique*. Masson et CNET-ENST, Paris.

Caraty, M.-J. (1999). *La reconnaissance vocale et son mentor : l'évaluation*. Habilitation à diriger des recherches Université Pierre et Marie Curie, Paris. *web*

Caraty, M.-J., C. Barras, F. Lefèvre & C. Montacié (1996). "D-DAL : un système de dictée vocale développé sous l'environnement HTK". *XXI Journées d'Etude sur la Parole*, Avignon, pp. 289-292. *web*

Caraty, M.-J., C. Montacié & F. Lefèvre (1997). "D-DAL, le Dictaphone-Dactylographe du LIP6". *Atelier de dictée vocale de l'AUPELF-UREF*, Avignon, pp. .

Caraty, M.-J., C. Montacié & F. Lefèvre (1997). "Dynamic Lexicon for a Very Large Vocabulary Vocal Dictation System". *Eurospeech*, Rhodes, pp. V.2691-2694.

Caraty, M.-J. & X. Rodet (1985). "Distance interspectrale à critères perceptifs". *XIV Journées d'Etudes sur la Parole*, Paris, pp. 87-90.

Casacuberta, F., E. Vidal, B. Mas & H. Rulot (1990). "Learning the Structure of HMM's through Grammatical Inference Techniques". *ICASSP*, Albuquerque, pp. I.717-720.

Clarkson, P. & A. Robinson (1998). "The Applicablity of Adaptative Language Modelling for the Braodcast News Task". *ICSLP*, Sydney, pp. V.1699-1702. *web*

Cook, G., A. Robinson & J. Christie (1998). "Real-Time Recognition of Broadcast News". *ICSLP*, Sydney, pp. V.1319-1322. *web*

Cooley, J. & J. Tukey (1965). "An Algorithm for machine calculation of Complex Fourier Series" *Mathematical Computer* **19**:297-301.

Cover, T. & P. Hart (1967). "Nearest Neighbor Pattern Classification" *IEEE Transactions on Information Theory* **13**:21-27.

Crystal, T. & A. House (1982). "Segmental Durations in Connected Speech Signals : Preliminary Results" *Journal of the Acoustical Society of America* **72**:705-716.

DARPA (1997). *1997 DARPA Speech Recognition Workshop*, Chantilly, Virginia. *web*

DARPA (1998). *DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowe, Virginia. *web*

Davies, E. (1988). "Training Sets and a priori Probabilities with the Nearest Neighbour Method of Pattern Recognition" *Pattern Recognition Letters* **8**:11-13.

Davis, S. & P. Mermelstein (1980). "Comparaison of Parametric Representations for Monosyllabic World Recognition in Continuously Spoken Sentences" *IEEE Transactions on Acoustics, Speech and Signal Processing* **28**(4):357-366.

de Mori, R. (1998). *Spoken Dialogues with Computers*. Academic Press, London - San Diego.

de Mori, R., M. Galler & F. Brugnara (1995). "Search and Learning Strategies for Improving Hidden Markov Models" *Computer Speech and Language* **9**:107-121.

Deligne, S. (1996). *Modèles de séquences de longueurs variables : application au traitement du langage écrit et de la parole*. Thèse de Doctorat ENST, Paris.

Dempster, A., N. Laird & D. Rubin (1977). "Maximum-likelihood from incomplete data via the EM algorithm" *Journal of the Royal Statistical Society* **39**:1-38.

Denoeux, T. (1995). "A K-nearest Neighbor Classification Rule Based on Dempster-Shafer Theory" *IEEE Transactions on Systems, Man and Cybernetics* **25**(5):1-28. *web*

Deroo, O. (1998). *Modèles dépendants du contexte et méthodes de fusion de données appliqués à la reconnaissance de la parole par modèles hybrides HMM/MLP*. Thèse de Doctorat Faculté Polytechnique de Mons, Mons. *web*

Derouault, A.-M. (1987). "Context-dependent Phonetic Markov Models for Large Vocabulary Speech Recognition". *ICASSP*, Dallas, pp. I.360-363.

Devijver, P. & J. Kittler (1982). *Pattern Recognition : A Statistical Approach*. Prentice-Hall, London.

Devillers, L. (1992). *Reconnaissance de la parole continue avec un système hybride neuronal et markovien*. Thèse de Doctorat Université Paris XI, Orsay.

Doddington, G. (1989). "Phonetically Sensitive Discriminants for Improved Speech Recognition". *ICASSP*, Glasgow, pp. I.556-559.

Dubuisson, B. (1990). *Diagnostic et reconnaissance des formes*. Hermes, Paris.

Duchateau, J. (1998). *HMM-based Acoustic Modelling in Large Vocabulary Speech Recognition*. PhD Thesis Université Catholique, Leuven. *web*

Duda, R. & P. Hart (1973). *Pattern Classification and Scene Analysis*. Wiley.

Dudani, S. (1976). "The Distance-Weighted k-Nearest-Neighbor Rule" *IEEE Transactions on Systems, Man and Cybernetics* **6**(4):325-327.

Dugast, C., L. Devillers & X. Aubert (1994). "Combining TDNN and HMM in an Hybrid System for Improved Continuous-Speech Recognition" *IEEE Transactions on Speech and Audio Processing* **2**(1):217-223.

Dupont, P. & L. Miclet (1998). *Inférence grammaticale régulière : fondements théoriques et principaux algorithmes*. Rapport interne IRISA, Rennes. *web*

Ephraïm, Y., A. Dembo & L. Rabiner (1989). "A Minimum Discrimination Information Approach for Hidden Markov Modeling" *IEEE Transactions on Information Theory* **35**(5):1000-1013.

Ephraïm, Y. & L. Rabiner (1990). "On the Relations Between Modeling Approaches for Speech Recognition" *IEEE Transactions on Information Theory* **36**(2):372-380.

Fix, E. & J. Hodges (1951). *Discriminatory Analysis : Nonparametric Discrimination : Consistency Proprieties*. Report n°4 USAF School of Aviation Medecine, .

Fohr, D., J.-P. Haton, J.-F. Mari, K. Smaïli & I. Zitouni (1997). "MAUD : un prototype de machine à dicter vocale". *1ères Journées Scientifiques et Techniques AUPELF-UREF*, Avignon, pp. 25-30.

Forney, G. (1973). "The Viterbi Algorithm" *Proceedings of the IEEE* **61**(3):268-278.

Franzini, M. & A. Waibel (1991). "Recent Work in Continuous Speech Recognition Using the Connectionist Viterbi Training Procedure". *ICASSP*, Toronto, pp. II.1213-1216.

Friedman, J., J. Bentley & R. Finkel (1975). "An Algorithm for Finding Nearest Neighbors" *IEEE Transactions on Computers* **october**:1000-1006.

Fukunaga, K. (1972). *Introduction to Statistical Pattern Recognition*. Academic Press, New-York.

Fukunaga, K. & P. Narendra (1975). "A Branch and Bound Algorithm for Computing K-Nearest Neighbors" *IEEE Transactions on Computers* **october**:750-753.

Furui, S. (1986). "On the Role of Spectral Transition for Speech Perception" *Journal of the Acoustical Society of America* **80**(4):1016-1025.

Furui, S. (1986). "Speaker-independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum" *IEEE Transactions on Acoustics, Speech and Signal Processing* **34**(1):52-60.

Gauvain, J.-L., L. Lamel, G. Adda & M. Adda-Decker (1994). "Speaker-Independent Continuous Speech Dictation" *Speech Communication* **15**:21-37. *web*

Gauvain, J.-L. & C.-H. Lee (1994). "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains" *IEEE Transactions on Speech and Audio Processing* **2**(2):291-298.

Godin, C. & P. Lockwood (1989). "DTW Schemes for Continuous Speech Recognition : a Unified View" *Computer Speech and Language* **3**:169-198.

Gopalakrishnan, P., D. Kanevsky, A. Nàdas & D. Nahamoo (1991). "An Inequality for Rational Functions with Applications to Some Statistical Estimation Problems" *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**(1):107-113.

Gopalakrishnan, P., D. Kanevsky, A. Nadàs, D. Nahamoo & M. Picheny (1988). "Decoder Selection Based on Cross-entropies". *ICASSP*, New-York, pp. .

Hart, P. (1968). "The Condensed Nearest Neighbor rule" *IEEE Transactions on Information Theory* **14**:515-516.

Hauptmann, A. (1995). "Speech Recognition in the Informedia Digital Video Library : Uses and Limitation". *7th IEEE International Conference on Tools with AI*, Washington DC, pp. .

Hermansky, H. (1990). "Perceptual Linear Predictive (PLP) Analysis" *Journal of the Acoustical Society of America* **87**(4):1738-1752.

Hermansky, H. & N. Morgan (1994). "Rasta Processing of Speech" *IEEE Transactions on Speech and Audio Processing* **2**:578-589.

Huang, X. (1992). "Phoneme Classification Using Semicontinuous Hidden Markov Models" *IEEE Transactions on Signal Processing* **40**(5):1062-1067.

Huang, X. & M. Jack (1989). "Semi-continuous Hidden Markov Models for Speech Signals" *Computer Speech and Language* **3**:239-251.

Hunt, J., M. Lenning & P. Mermelstein (1980). "Experiments in syllabe-based recognition of continuous speech". *ICASSP*, Denver, pp. I.880-883.

Hwang, M.-Y. & X. Huang (1993). "Shared-distribution Hidden Markov Models for Speech Recognition" *IEEE Transactions on Speech and Audio Processing* **1**(4):414-420.

Hwang, M.-Y., X. Huang & F. Alleva (1996). "Predicting Unseen Triphones with Senones" *IEEE Transactions on Speech and Audio Processing* **4**(6):412-419.

Itakura, F. (1975). “Minimum Prediction Residual Principle Applied To Speech Recognition” *IEEE Transactions on Acoustics, Speech and Signal Processing* **23**:62-72.

Iwamida, H., S. Katagiri & E. McDermott (1991). "Speaker-independent Large Vocabulary Word Recognition Using a LVQ/HMM Hybrid Algorithm". *ICASSP*, Toronto, pp. .

Jelinek, F. (1976). “Continuous Speech Recognition by Statistical Methods” *Proceedings of the IEEE* **64**(4):532-556.

Jelinek, F. (1989). "Self-Organized Language Modeling for Speech Recognition". *Readings in Speech Recognition*. A. Waibel & K.-F. Lee, Eds, Morgan Kaufman, pp. 450-506.

Jelinek, F. (1997). *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, Massachusetts.

Jouvet, D. (1995). "Modèles de Markov pour la reconnaissance de la parole". *Ecole Thématique "Fondements et Perspectives en Traitement Automatique de la Parole"*, Marseille, pp. I.99-108.

Jouvet, D., M. Dautremont & A. Gossart (1994). "Comparaison de multimodèles et des densités multigaussiennes pour la reconnaissance de la parole par modèles de Markov". *XX Journées d'Etude sur la Parole*, Trégastel, pp. I.159-164.

Jouvet, D., L. Mauuary & J. Monné (1991). "Automatic Adjustments of the Structure of Markov Models for Speech Recognition Applications". *Eurospeech*, Gênes, pp. I.927-930.

Juang, B.-H. & S. Katagiri (1992). “Discriminative learning for Minimum Error Classification” *IEEE Transactions on Signal Processing* **40**(12):3043-3054.

Juang, B.-H. & L. Rabiner (1985). “Mixture Autoregressive Hidden Markov Models for Speech Signals” *IEEE Transactions on Acoustics, Speech and Signal Processing* **33**(6):1404-1413.

Juang, B.-H. & L. Rabiner (1985). “A Probabilistic Distance Measure for Hidden Markov Models” *Bell System Technical Journal* **64**(2):391-408.

Juang, B.-H., L. Rabiner & J. Wilpon (1987). “On the Use of Bandpass liftering in Speech Recognition” *IEEE Transactions on Acoustics, Speech and Signal Processing* **35**(7):275-284.

Junqua, J.-C. & L. Vassallo (1996). "Context Modeling and Clustering in Continuous Speech Recognition". *ICSLP*, Philadelphia, pp. IV.2262-2265.

Katagiri, S., C.-H. Lee & B.-H. Juang (1991). “New Discriminative Training Algorithm Based on the Generalized Descent” *IEEE Workshop on Neural Networks for Signal Processing*:309-318.

Katz, S. (1987). “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recogniser” *IEEE Transactions on Acoustics, Speech and Signal Processing* **35**(3):400-401.

Keller, J., M. Gray & J. Givens (1985). “A Fuzzy K-NN Neighbor Algorithm” *IEEE Transactions on Systems, Man and Cybernetics* **15**(4):580-585.

Kim, B. & S. Park (1986). “A Fast k Nearest Neighbor Finding Algorithm Based on the Ordered Partition” *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8**(6):761-766.

Kittler, J. (1978). “A Method for Determining K-Nearest Neighbours” *Kibernetics* **7**:313-315.

Konig, Y. (1996). *REMAP : Recursive Estimation and Maximisation of A Posteriori in Transition-based Speech Recognition*. PhD Thesis University of California, Berkeley. *web*

Kuhn, R. & R. de Mori (1990). “A Cache-based Natural Language Model for Speech Recognition” *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(6):570-583.

Kullback, S. & R. Leibler (1951). “On Information and Sufficiency” *Annals of Mathematical Statistics* **22**:79-86.

Lamel, L. & J.-L. Gauvain (1993). "High Performance Speaker-independent Phone Recognition Using CDHMM". *Eurospeech*, Berlin, pp. I.121-124. *web*

Lau, R., R. Rosenfeld & S. Roucos (1993). "Trigger-based Language Models : a Maximum Entropy Approach". *ICASSP*, Minneapolis, pp. II.45-48.

Lee, C.-H., E. Giachin, L. Rabiner, R. Pieraccini & A. Rosenberg (1992). “Improved Acoustic Modeling for Large Vocabulary Continuous Speech Recognition” *Computer Speech and Language* **6**:103-127.

Lee, K.-F. (1988). *Large-Vocabulary Speaker-Independant Continuous Speech Recognition : the SPHINX System*. PhD Thesis Carnegie Mellon, Pittsburgh.

Lee, K.-F. (1990). “Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition” *IEEE Transactions on Pattern Analysis and Machine Intelligence* **38**(4):599-609.

Lee, K.-F. & H.-W. Hon (1989). “Speaker-independent Phone Recognition Using Hidden Markov Models” *IEEE Transactions on Acoustics, Speech and Signal Processing* **17**(11):1641-1648.

Lee, K.-F., H.-W. Hon, M.-Y. Hwang & S. Mahajan (1990). “Recent Progress and Future Outlook of the SPHINX Speech Recognition System” *Computer Speech and Language* **4**:57-69.

Lee, K.-F. & S. Mahajan (1990). “Corrective and Reinforcement Learning for Speaker-Independent Continuous Speech Recognition” *Computer Speech and Language* **4**:231-245.

Lee, L. & R. Rose (1996). "Speaker Normalisation Using Efficient Frequency Warping Procedures". *ICASSP*, Atlanta, pp. I.353-356.

Lefèvre, F. (1995). *Contribution au système de dictée vocale D-DAL : N-Best et modèles de langage*. Rapport de DEA Université Pierre et Marie Curie, Paris.

Lefèvre, F., C. Montacié & M.-J. Caraty (1997). "K-NN Estimator in a HMM-Based Recogntion System". *Computational Models of Speech Pattern Processing*. K. Ponting, Ed., St Hélier, Jersey, Springer, pp. 96-101.

Lefèvre, F., C. Montacié & M.-J. Caraty (1998). "On the Influence of the Delta Coefficients on a HMM-based Recognition System". *ICSLP/SST*, Sydney, pp. .

Lefèvre, F., C. Montacié & M.-J. Caraty (1999). "A MLE Algorithm for the K-nn/HMM System". *Eurospeech*, Budapest, pp. VI.2733-2736.

Leggetter, C. & P. Woodland (1995). “Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models” *Computer Speech and Language* **9**:171-185. *web*

Linde, Y., A. Buzo & R. Gray (1980). “An Algorithm for Vector Quantizer Design” *IEEE Transactions on Communications* **28**(1):84-95.

Liporace, L. (1982). “Maximum Likelihood Estimation for Multivariate Observations of Markov Sources” *IEEE Transactions on Information Theory* **28**(5):729-734.

Lippmann, R. (1987). “An Introduction to Computing with Neural Nets” *IEEE Transactions on Acoustics, Speech and Signal Processing* **3**:4-22.

Ljolje, A. (1994). “High Accuracy Phone recogntion Using Context Clustering and Quasi-triphones Models” *Computer Speech and Language* **8**:129-151.

Lockwood, P. & M. Blanchet (1993). "An Algorithm for the Dynamic Inference of Hidden Markov Models (DIHMM)". *ICASSP*, Minneapolis, pp. II.251-255.

Loftsgaarden, D. & C. Quesenbury (1965). “A Nonparametric Estimate of a Multivariate Density Function” *Annals of Mathematical Statistics* **36**:1049-1151.

Mari, J.-F., J.-P. Haton & A. Kriouile (1997). “Automatic Word Recognition Based on Second-order Hidden Markov Models” *IEEE Transactions on Speech and Audio Processing* **5**(1):22-25.

Mariani, J. (1998). “Some Evaluation-based Language Engineering Actions for French” *Computer Speech and Language* **12**:307-316.

Markel, J. & A. Gray (1976). *Linear Prediction of Speech*. Springer-Verlag, Berlin Heidenberg New-York.

Miclet, L. (1990). "Grammatical Inference". *Syntactic and Structural Pattern Recognition : Theory and Application*. H. Bunke & A. Sanfeliu, Eds, Singapore, World Scientific, pp. 7.237-290.

Miclet, L. & M. Dabouz (1983). “Approximative Fast Nearest-Neighbor Recognition” *Pattern Recognition Letters* **I**:277-285.

Montacié, C. (1999). *Indexation et interface Hmme-Machine. Reconnaissance d'un signal vocal*. Habilitation à diriger des recherches Université Pierre et Marie Curie, Paris. *web*

Montacié, C. & M.-J. Caraty (1998). "A Silence/Noise/Music/Speech Splitting Algorithm". *ICSLP*, Sydney, pp. IV.1579-1582.

Montacié, C., M.-J. Caraty & C. Barras (1996). "Mixture Splitting Technic and Temporal Control in a HMM-Based Recognition System". *ICSLP*, Philadelphie, pp. III.977-980. *web*

Montacié, C., M.-J. Caraty & F. Lefèvre (1997). "KNN versus Gaussian in a HMM-Based System". *Eurospeech*, Rhodes, pp. II.529-533. *web*

Montacié, C., M.-J. Caraty & F. Lefèvre (1997). "Lexique Dynamique dans un système de reconnaissance de la parole". *VIème Congrès Français d'Acoustique*, Marseille, pp. I.413-416. *web*

Montacié, C. & G. Chollet (1987). "Systèmes de référence pour l'évaluation d'applications et la caractérisation de bases de données en reconnaissance automatique de la parole". *XVI Journées d'Etude sur la Parole*, Hammamet, pp. 323-326.

Morgan, N., H. Bourlard & S. Renals (1993). “Hybrid Neural Network/Hidden Markov Model Systems for Continuous Speech Recognition” *International Journal of Pattern Recognition and Artifical Intelligence* **7**(4):899-916.

Myers, C. & L. Rabiner (1981). “A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition” *IEEE Transactions on Acoustics, Speech and Signal Processing* **29**(2):284-296.

Nadàs, A. (1983). “A Decision Theoric Formulation of a Training Problem in Speech Recognition and a Comparaison of Training by Unconditional Versus Conditional Maximum Likelihood” *IEEE Transactions on Acoustics, Speech and Signal Processing* **31**(4).

Nadàs, A., R. Mercer, L. Bahl, R. Bakis, P. Cohen, A. Cole, F. Jelinek & B. Lewis (1981). "Continuous Speech Recognition with Automatically Selected Acoustic Prototypes Obtained by Either Bootstrapping or Clustering". *ICASSP*, Atlanta, pp. .

Neimann, H. & G. Goppert (1988). “An Efficient Branch-and-Bound Nearest Neighbour Classifier” *Pattern Recognition Letters* **7**:67-72.

Ney, H., R. Haeb-Umbach, B.-H. Tran & M. Oerder (1992). "Improvement in Beam Search for 10000-Word Continuous Speech Recognition". *ICASSP*, San Fransisco, pp. I.9-12.

Ney, H. & A. Noll (1988). "Phoneme Modelling Using Continuous Mixtures Densities". *ICASSP*, New-York, pp. .

Nguyen, L. & R. Schwartz (1997). "Efficient 2-Pass N-Best Decoder". *Eurospeech*, Rhodes, pp. I.167-170.

Niles, L. (1991). "TIMIT Phoneme Recognition Using an HMM-derived Recurrent Neural Network". *Eurospeech*, Gênes, pp. II.559-562.

Nilsson, N. (1971). *Problem Solving Methods in Artifical Intelligence*. McGraw-Hill, New-York.

NIST (1990). *The DARPA TIMIT Acoustic-phonetic Continuous Speech Recognition Database*. Fourni avec le CDROM .

Normandin, Y., R. Cardin & R. de Mori (1994). “High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation” *IEEE Transactions on Speech and Audio Processing* **2**(2):299-311.

Ortmanns, S., A. Eiden & H. Ney (1998). "Improved Lexical Tree Search for Large Vocabulary Speech Recognition". *ICASSP*, Seattle, pp. II.817-820.

Ortmanns, S., T. Firzlaff & H. Ney (1997). "Fast Likelihood Computation Methods for Continuous Mixture Densities in Large Vocabulary Speech Recognition". *Eurospeech*, Rhodes, pp. I.139-142.

Ortmanns, S., H. Ney, F. Seide & I. Lindam (1996). "A Comparaison of Time Conditioned and Word Conditioned Search Techniques for Large Vocabulary Speech Recognition". *ICSLP*, Philadelphie, pp. 2091-2094.

Ostendorf, M., V. Digalakis & O. Kimball (1996). “From HMM's to Segment Models : a Unified View of Stochastic Modeling for Speech Recognition” *IEEE Transactions on Speech and Audio Processing* **4**(5):360-378.

Oswald, J. (1986). *Théorie de l'information et analyse chaotique des systèmes*. Masson, Paris.

Parzen, E. (1962). “On Estimation of a Probability Density Function and Mode” *Annals of Mathematical Statistics* **33**:1065-1076.

Paul, D. & E. Martin (1988). "Speaker Stress-Resistant Continuous Speech Recognition". *ICASSP*, New-York, pp. .

Picone, J. (1993). “Signal Modeling Techniques in Speech Recognition” *Proceedings of the IEEE* **81**(9):1215-1247.

Rabiner, L. & R. Schafer (1978). *Digital Processing of Speech Signals*. Prentice-Hall.

Rabiner, L., J. Wilpon & F. Soong (1989). “High Performance Connected Digit Recognition Using Hidden Markov Models” *IEEE Transactions on Acoustics, Speech and Signal Processing* **17**(8):1214-1225.

Ravishankar, M. (1996). *Efficient Algorithms for Speech Recognition*. PhD Thesis Carnegie Melon University, Pittsburgh. *web*

Redner, R. & H. Walker (1984). “Mixture Densities, Maximum Likelihood and the EM Algorithm” *SIAM Review* **26**(2):195-239.

Renals, S. & M. Hochberg (1995). "Efficient Search Using Posterior Phone Probability Estimates". *ICASSP*, Detroit, pp. I.596-599.

Renals, S., N. Morgan, M. Cohen & H. Franco (1992). "Connectionist Probability Estimation in the DECIPHER Speech Recogntion System". *ICASSP*, San-Francisco, pp. I.601-604.

Richetin, M., G. Rives & M. Naranjo (1980). “Algorithme rapide pour la détermination des k plus proches voisins” *RAIRO Informatique* **14**(4):369-378.

Richter, A. (1986). "Modeling of Continuous Speech Observations". *Advances in Speech Processing Conference*, IBM Europe Institute, pp. .

Rigoll, G. (1994). “Maximum Mutual Information Neural Networks for Hybrid Connectionist-HMM Speech Recognition Systems” *IEEE Transactions on Speech and Audio Processing* **2**(1):175-184. *web*

Robinson, A. (1994). “An Application of Reccurent Nets to Phone Probability Estimation” *IEEE Transactions on Neural Networks* **5**(2):298-305. *web*

Rumelhart, D., G. Hinton & R. Williams (1986). "Learning Internal Representations by Error Propragation". *Parallel Distributed Processing Exploration of the Microstructure of Cognition*. D. Rumelhart & J. McClelland, Eds, MIT Press, pp. 1.318-362.

Ruske, G. (1982). "Auditory Perception and its Application to Computer Analysis of Speech". *Computer Analysis and Perception*. C. Y. Suen & R. de Mori, Eds, Baco Raton, CRC Press, pp. II.

Sakoe, H. (1979). “Two-level DP-Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition” *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**:588-595.

Sakoe, H. & S. Chiba (1978). “Dynamic Programming Algorithm Optimization for Spoken Word Recognition” *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**(1):43-49.

Savariaux, C., A. Farhat, M. Héon, D. O'Shaughnessy & C. Lee (1997). "Nouvelles avancées en reconnaissance de la parole continue grand vocabulaire du Français basées sur le système de reconnaissance de l'INRS-Télécommunications". *1ères Journées Scientifiques et Techniques AUPELF-UREF*, Avignon, pp. 31-34.

Schwartz, R. & S. Austin (1991). "A Comparaison of Several Approximate Algorithms for Finding Multiple (N-Best) Sentence Hypothesis". *ICASSP*, Toronto, pp. I.701-704.

Schwartz, R. & Y. Chow (1990). "The N-Best Algorithm : An Efficient and Exact Procedure for Finding the N Most Likely Sentence Hypotheses". *ICASSP*, Albuquerque, pp. I.81-84.

Schwartz, R., J. Klovstad, J. Makhoul & J. Sorensen (1980). "A Preliminary Design of a Phonetic Vocoder Based on a Diphone Model". *ICASSP*, Denver, pp. I.32-35.

Schwartz, R. & F. Kubala (1992). "Hidden Markov Models and Speaker Adaptation". *Speech Recognition and Understanding*. P. Laface & R. de Mori, Eds, Heidelberg, Springer-Verlag, pp. 31-57.

Seneff, S. & V. Zue (1988). *Transcription and Alignement of the TIMIT Database*. NIST.

Shikano, K., K.-F. Lee & D. Reddy (1986). "Speaker Adaptation Through Vector Quantization". *ICASSP*, Tokyo, pp. .

Simon, J.-C. (1984). *Reconnaissance des formes par algorithmes*. Masson, Paris.

Soudoplatoff, S. (1986). "Markov Modeling of Continuous Parameters in Speech Recogntion". *ICASSP*, Tokyo, pp. I.45-48.

Takami, J. & M. Sagayama (1992). "A Successive State Splitting Algorithm for Efficient Allophone Modelling". *ICASSP*, San Fransisco, pp. II.573-576.

Takara, T., Y. Iha & I. Nagayama (1998). "Selection of the Optimal Structure fo the Continuous HMM Using the Genetic Algorithm". *ICSLP*, Sydney, pp. III.751-754.

Valtchev, V. (1995). *Discriminative Methods in HMM-based Speech Recognition*. PhD Thesis University of Cambridge, Cambridge. *web*

Valtchev, V., J. Odell, P. Woodland & S. Young (1994). "A Dynamic Network Decoder Design for Large Vocabulary Speech Recognition". *ICSLP*, Yokohama, pp. III.1351-1354. *web*

Valtchev, V., J. Odell, P. Woodland & S. Young (1996). "Lattice-Based Discriminative Training for Large Vocabulary Speech Recognition". *ICASSP*, Atlanta, pp. I.605-608. *web*

Van Compernolle, D., J. Smolders, P. Jaspers & T. Hellemans (1991). "Speaker Clustering for Dialectic Robustness in Speaker Independent Recognition". *Eurospeech*, Gênes, pp. II.723-726.

Vidal, E. (1994). “New Formulation and Improvements of the NN ASEA” *Pattern Recognition Letters* **15**(1):1-7.

Vintsyuk, T. (1971). “Element-wise Recogntion of Continuous Speech Composed of Words from a Specified Dictionary” *Cybernetics* **7**(2):133-143.

Viterbi, A. (1967). “Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm” *IEEE Transactions on Information Theory* **13**(2):260-269.

Waast, C., L. Bahl & M. El Bèze (1995). "Fast Match on Decision Trees". *Eurospeech*, Madrid, pp. II.909-912.

Waegner, N. & S. Young (1992). "A Treillis-based Language Model for Speech Recognition". *ICSLP*, Banff, pp. II.245-248.

Wagner, R. & M. Fisher (1974). “The String to String Correction Problem” *Journal of ACM*.

White, G. & P. Fong (1975). “K-Nearest-Neighbour Decision Rule Performance in a Speech Recognition System” *IEEE Transactions on Systems, Man and Cybernetics* **5**:389.

Wilpon, J., C.-H. Lee & L. Rabiner (1993). “Connected Digit Recognition Based on Improved Acoustic Resolution” *Computer Speech and Language* **7**:15-26.

Wilson, D. (1972). “Asymptotic Proprieties of Nearest Neighbor Rules Using Edited Data” *IEEE Transactions on Systems, Man and Cybernetics* **2**(3):408-421.

Woodland, P. & D. Cole (1991). "Optimising Hidden Markov Models Using Discriminative Output Distributions". *ICASSP*, Toronto, pp. I.545-548. *web*

Woodland, P., C. Leggetter, J. Odell, V. Valtchev & S. Young (1995). "The 1994 HTK Large Vocabulary Speech Recognition System". *ICASSP*, Detroit, pp. I.73-76.

Young, S. (1992). "The General Use of Tying in Phoneme-based HMM Speech Recognisers". *ICASSP*, San-Francisco, pp. I.569-572.

Young, S., J. Jansen, J. Odell, D. Ollason & P. Woodland (1995). *The HTK Book*Cambridge University, Cambridge. *web*

Young, S., N. Russel & J. Thornton (1989). *Token Passing : a Simple Conceptual Model for Connected Speech Recognition Systems*. Technical Report CUED-Speech Group, Cambridge. *web*

Young, S. & P. Woodland (1994). “State Clustering in Hidden Markov Model-based Continuous Speech Recognition” *Computer Speech and Language* **8**:369-383. *web*

Zadeh, L. (1965). “Fuzzy Sets” *Information Control* **8**:338-353.

Zouhal, L. & T. Denoeux (1997). "Generalizing the Evidence-theoretic k-NN Rule to Fuzzy Pattern Recognition". *Second International Symposium on Fuzzy Logic and Applications*, Zurich, pp. 294-300. *web*

TABLE DES NOTATIONS

$\alpha_t(i)$, $\beta_t(i)$	variables directe et rétrograde de Baum
$A = \{a_{ij}\}$	matrice des probabilités de transition entre états d'un HMM
$B = \{b_j\}$	vecteur des fdp associés aux états d'un HMM
C_k	k^{ième} coefficient cepstral
D_k	k^{ième} coefficient différentiel (delta)
δ	symbole de Kronecker
d	dimension des vecteurs de paramètres
d_E, d_M, d_∞	distance euclidienne, de Manhattan, max
$D(x)$	règle de décision pour le vecteur x, renvoie l'index de la classe reconnue
Ex	ensemble d'exemples (par extension, le cardinal de cet ensemble)
$E[\ ]$	espérance mathématique
$\Phi(t,i)$	probabilité partielle du chemin optimal au temps t sur l'état i
f	indice de flux de données
γ_f	exposant de pondération du flux de données f
G	nombre de gaussiennes
$H(\ )$	mesure de l'entropie
h	échelle du noyau de Parzen
h_i	paramètres du H-critère
$I(\ )$	mesure de la quantité d'information
$K(\ )$	noyau de Parzen
$KL(\ )$	divergence de Kullback-Leibler
k_i	nombre d'éléments de la classe i parmi k
λ	modèle de Markov caché (par extension, les paramètres du modèle)
L	nombre de symboles d'une source d'information
m	mot
M	suite de mots
n	nombre d'éléments d'un ensemble
n_m	nombre de mots
n_V	nombre de symboles
N	nombre d'états d'un HMM
$\mathcal{N}(o,\mu,\Sigma)$	loi normale de moyenne μ et de matrice de covariance Σ
$O = (o_1 \ldots o_T)$	suite d'observations acoustiques
o	observation acoustique, vecteur de paramètres
$P = (p_1 \ldots p_T)$	séquence de ppv

$P(\,)$	probabilité
$\hat{P}()$	estimation de la probabilité
p( )	densité de probabilité
p(.|c)	densité de probabilité de la classe c
$Q = (q_1 \ldots q_T)$	séquence d'états dans un modèle
$\Re$	ensemble des réels
R	région de l'espace de représentation
s	facteur d'échelle du bigrammes
S	nombre total d'états du système
t, T	instant, longueur d'une suite d'observations
u_i^s, u_i^d	coefficient d'appartenance associé aux coefficients statiques, delta
V	ensemble des vecteurs de référence
Vol	volume d'une région de l'espace
Vol_n	volume d'une région de l'espace à n échantillons
X,Y	variables aléatoires de réalisation x,y
w	poids relatifs des coefficients delta

TABLE DES ABREVIATIONS

ACP	Analyse en Composantes Principales
ALD	Analyse Linéaire Discriminante
BBN	Bolt, Beranek & Newman
CMLE	Conditional Maximum Likelihood Estimation (*estimation par maximum de vraisemblance conditionelle*)
CMU	Carnegie Mellon University
DAP	Décodage Acoustico-Phonétique
D-DAL	Dictaphone-Dactylographe du LIP6
DTW	Dynamic Time Warping (*alignement temporel dynamique*)
EM	Expectation-Maximisation
fdp	fonction de densité de probabilité
GPD	Generalized Probabilistic Descent (*descente de gradient généralisée*)
HMM	Hidden Markov Models (*modèles de Markov cachés*)
HTK	HMM Toolkit (*boîte à outils pour modèles de Markov cachés*)
ICASSP	International Conference on Acoustics, Speech and Signal Processing
IEEE	Institute of Electrical and Electronics Engineers
K-ppv	K-plus proches voisins
LB	Level Building (*algorithme de décodage par construction de niveaux*)
LIP6	Laboratoire d'Informatique de Paris 6
LPCC	Linear Predictive Cepstral Coefficient
MAP	Maximum *A posteriori*
MCE	Minimum Classification Error (*erreur de classification minimale*)
MFCC	Mel Frequency Cepstral Coefficient (*coefficient cepstral en échelle Mel*)
MFCC+E	, augmentés de l'énergie
MFCC+E+D	, augmentés de l'énergie et des coefficients delta
MFCC+E_D	Sous vecteur de MFCC+E+D composé des coefficients delta
MDI	Minimum Discrimination Information
ML	Modèle de langage
MLE	Maximum Likelihood Estimation (*estimation par maximum de vraisemblance*)

MLLR	Maximum Likelihood Linear Regression (*régression linéaire par maximum de vraisemblance*)
MLP	Multi-Layer Perceptron (*perceptron multi-couches*)
MMI	Maximum Mutual Information (*information mutuelle maximale*)
OP	One Pass (*algorithme de décodage une passe*)
QV	Quantification Vectorielle
RAP	Reconnaissance Automatique de la Parole
REMAPS	Recursive Estimation and Maximisation of *A posteriori* Probabilities
RN	Reccurent Network (*réseau de neurones récurrent*)
TDNN	Time Delay Neural Network (*réseau de neurones à délai*)
TL	Two Level (*algorithme de décodage à deux niveaux*)
TP	Token Passing (*algorithme de décodage par passage de jetons*)

TABLE DES FIGURES

TABLE DES TABLEAUX

TABLE DES MATIERES