CHAPITRE IX
Post-traitement segmental statistique

1 Introduction

La figure IX.1 présente le post-traitement segmental appliqué aux N meilleures solutions proposées par le module markovien. A la sortie de ce module markovien, on récupère le score, l'alignement et la séquence de mots de chaque solution. Pour chaque solution, l'alignement fournit un découpage en segments de parole. Un segment de parole est défini comme un ensemble de trames acoustiques et identifié par une étiquette qui représente un phonème. Afin de calculer des scores segmentaux pour chaque solution, un modèle statistique est associé à chaque étiquette. Ainsi on parle de post-traitement segmental statistique.

Le traitement segmental consiste à calculer un score de vraisemblance pour chaque segment de chacune des solutions. Le score global d'une solution est égal au produit des scores des segments la constituant. Le score final d'une solution est obtenu par une combinaison linéaire du type $(\alpha, 1-\alpha)$, de son score segmental et de son score markovien. La solution reconnue parmi les N développées est celle ayant le score final le plus élevé.

La phase d'apprentissage consiste à déterminer les paramètres utilisés dans le traitement segmental (paramètres des modèles statistiques) et à calculer le paramètre α utilisé pour la combinaison linéaire des scores.

Le module markovien a été développé dans les chapitres précédents, aussi nous allons nous intéresser dans ce chapitre uniquement au module de post-traitement segmental statistique.

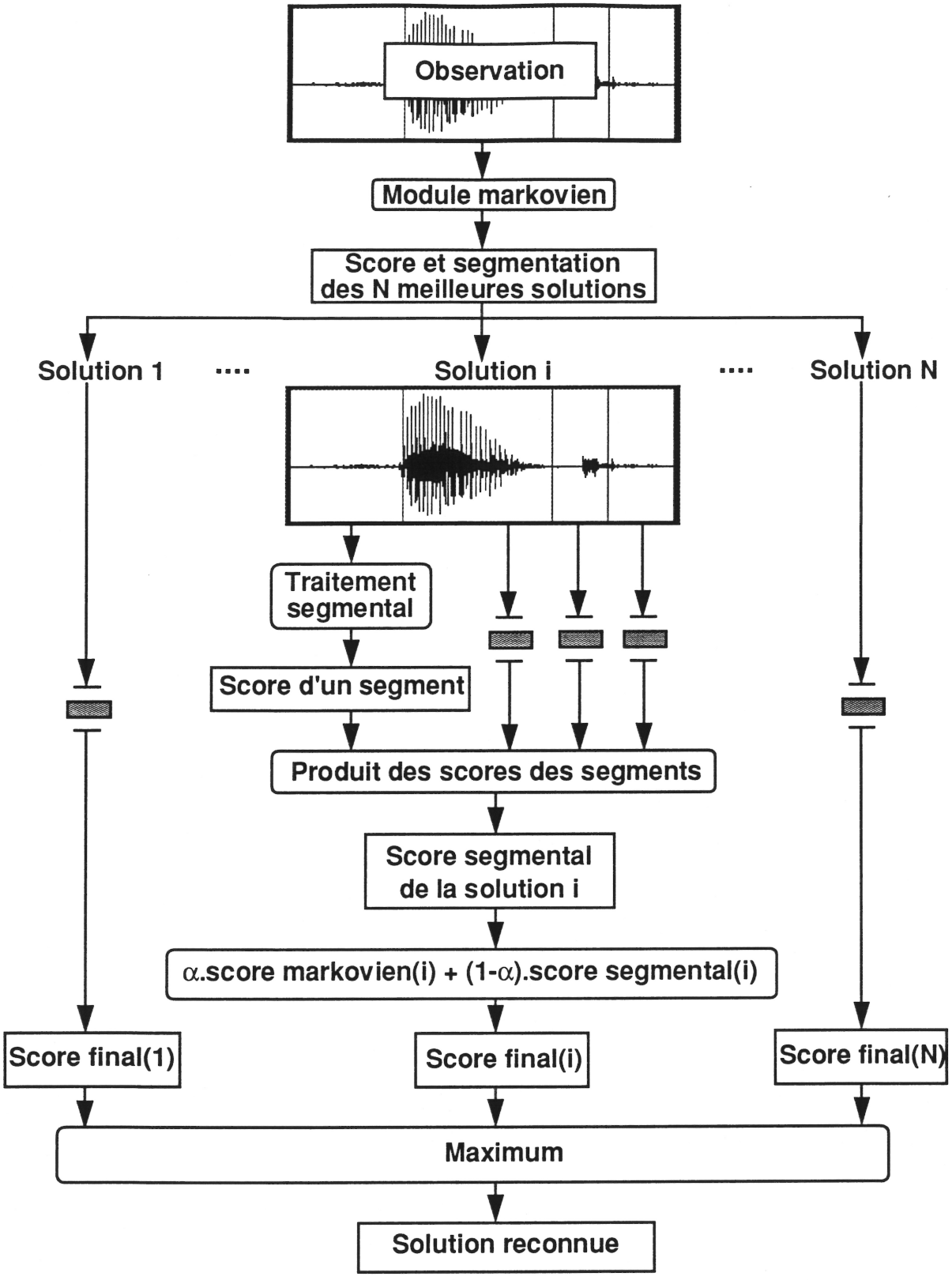

Figure IX.1 Recherche des *N* meilleures solutions
et post-traitement segmental statistique

2 Description de l'observation

Pour une observation inconnue X, on développe les N meilleures solutions. L'alignement obtenu pour chaque solution par le module markovien définit les différents segments utilisés pour décrire la solution. La figure IX.2 donne un exemple des 3 meilleures solutions pour le mot prononcé "5" et illustre les différents segments mis en jeu pour chaque solution. L'observation correspondant à la k^{ième} meilleure solution est notée X_{i_k}.

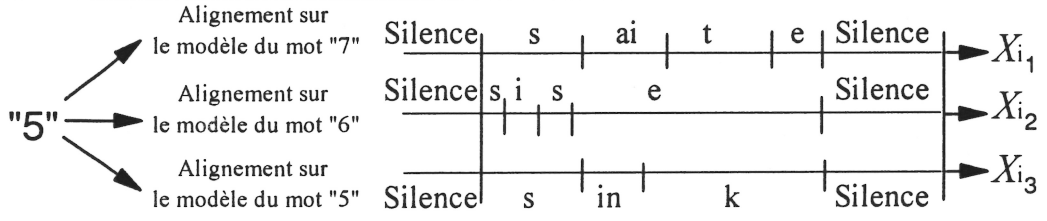

Figure IX.2 Exemple d'alignements acoustiques obtenus en prononçant le mot "5".

Ainsi, on remarque qu'on n'a plus une observation unique mais N observations correspondant aux N meilleures solutions fournies par le premier module. Afin de ne pas privilégier une segmentation plus qu'une autre, on définit l'observation globale $\tilde{X}$ comme étant la concaténation des observations associées à chacune des solutions proposées :

$$\tilde{X} = \left\{ X_{i_1}, X_{i_2}, X_{i_3}, \ldots, X_{i_N} \right\} \qquad \text{(IX.1)}$$

où N est le nombre de solutions développées.

L'indice de i sert à distinguer l'ordre dans lequel les solutions sont développées.

3 Modélisation statistique d'un segment

Avant de définir les différents paramètres segmentaux utilisés, nous allons introduire dans cette section différentes notations utilisées pour définir un segment et tout son environnement.

w est le mot prononcé, X est l'observation obtenue et K le nombre de segments contenus dans l'observation. Chaque observation est représentée par un ensemble de segments et d'étiquettes, obtenus à partir de l'alignement. Les modèles statistiques à utiliser sont choisis à partir des étiquettes de l'observation.

3.1 Segments

Une observation X_{i_k} des N meilleures solutions est définie par la concaténation des segments $s_{i_k,j}$.

$$X_{i_k} = \left\{ s_{i_k,1}, s_{i_k,2}, \ldots, s_{i_k,K_{i_k}} \right\} \qquad \text{(IX.2)}$$

où $s_{i_k,j}$ est le j^{ième} segment de l'observation X_{i_k} et K_{i_k} le nombre de segments dans l'observation X_{i_k}.

Chaque segment de parole est défini par un ensemble de trames acoustiques. Les marqueurs de début et de fin de segments (première et dernière trame de chaque segment) sont déterminés à partir de l'alignement du mot prononcé sur un modèle de Markov.

Pour l'exemple de la figure IX.2, l'observation X_{i_3} correspond à une segmentation correcte. De ce fait les segments associés à cette solution sont qualifiés de segments corrects. Pour les autres observations (ici X_{i_1} et X_{i_2}) les segments observés sont des segments incorrects.

Pour certains corpus de données, il existe fréquemment des similitudes entre les différentes phrases possibles, par exemple 80 dans 80, 81, 82, 83, etc. Quand 80 est prononcé et que 81 est reconnu, il n'est pas judicieux de considérer tous les segments relatifs à la solution 81 comme incorrects pour la phase d'apprentissage (estimation des paramètres des modèles). Pour cela, avant de déclarer un segment comme incorrect, on commence par rechercher s'il a des liens avec le mot correctement reconnu. On identifie ainsi les segments qui diffèrent des segments corrects. Ces segments servent pour estimer les paramètres des modèles statistiques "incorrects" associés. Les segments des alignements incorrects qui correspondent à des segments corrects ne sont pas pris en compte pour l'apprentissage des paramètres des modèles statistiques.

Pour l'exemple présenté sur la figure IX.2, on ne tient pas compte du segment /s/ appartenant à la solution incorrecte X_{i_1}, car il a un recouvrement important avec le segment /s/ de la solution correcte X_{i_3}. Ainsi on va réaliser un apprentissage discriminant sur les segments.

Le recouvrement de deux segments $s_{i_k,j}$ et $s_{i_c,l}$ est défini comme le rapport de leur intersection sur leur union :

$$recouvrement = \frac{durée\left(s_{i_k,j} \cap s_{i_c,l}\right)}{durée\left(s_{i_k,j} \cup s_{i_c,l}\right)} \qquad \text{(IX.3)}$$

Dans les expériences effectuées et décrites dans le chapitre X, on a exclu les segments des alignements incorrects ayant un recouvrement supérieur à 75% avec un segment correct de même nom.

3.2 Etiquettes

Chaque segment $s_{i_k,j}$ est identifié soit par une étiquette contextuelle $E_{Cs_{i_k,j}}$ ou par une étiquette hors contexte $E_{HCs_{i_k,j}}$. Dans notre application l'étiquetage est défini à partir du découpage en phonèmes, que réalise la modélisation markovienne. Dans le cas hors contexte, l'étiquette prise en compte pour le segment $s_{i_k,j}$ correspond au nom du phonème mis en correspondance avec ce segment. Dans le cas contextuel, l'étiquette prend en considération les noms des phonèmes qui précédent et suivent le phonème considéré.

L'exemple présenté ci-dessous (figure IX.3) est l'alignement d'un mot prononcé "5" sur le modèle de Markov de ce même mot.

La première colonne représente les différents modèles (unités de base) utilisés pour la description du mot. La deuxième colonne représente les différentes fonctions de densité de probabilité utilisées par chaque modèle (en fait, uniquement celles apparaissant sur le chemin optimal). Le reste du tableau représente les trames acoustiques (symbolisées par des *) mises en correspondance avec chacune des fonctions gaussiennes lors de l'alignement.

Les fonctions gaussiennes sont identifiées par un nom du type "**Pho.xxx**" ; **Pho** correspond au nom de l'unité de base et **.xxx** est un indice à l'intérieur de cette unité.

Modèles	Fonctions	Trames acoustiques
##	##.f3 ##.g3 ##.f8	********** * *
S	s.fe_pau s.fh1 s.fh2 s.fv3 s.fs_ann	* ** ***** * **
IN	in.fex_alv in.fv1 in.fv3 in.fs_vel	** * *** **
K	k.fe_ann k.fex_ann k.fn1 k.fn2 k.fv3 k.fsx_aam k.fs_aam	** ** ** **** ** ** *
E	e.fex_vel e.fsx_pau	*** **
$$	$$.f2 $$.g2 $$.g5 $$.g7	** *** ** **********
Etiquettes →		< ## > < S > <IN> < K > <E> < $$ >

Figure IX.3 Alignement du mot "5" sur le modèle du "5".

Chaque segment est obtenu en regroupant l'ensemble des trames mises en correspondance avec des gaussiennes appartenant à une même unité (phonème ou allophone). Ici, l'alignement a permis de définir les six segments résultant de la reconnaissance du mot prononcé "5". Les étiquettes correspondant à ces segments ont été identifiées à partir du nom des fonctions de densité associées. Ces étiquettes sont, pour le mot 5 traité ici :

$$"5" : \#\# \,.\, S \,.\, IN \,.\, K \,.\, E \,.\, \$\$.$$

Puisqu'on a réalisé l'alignement du mot prononcé "5" sur le modèle du mot "5", les étiquettes obtenues correspondent à des segments corrects. On distingue aussi deux sortes d'étiquettes : les étiquettes $E_{correct\, s_{i_k},j}$ des segments corrects et les étiquettes $E_{incorrect\, s_{i_k},j}$ des segments incorrects. On retrouve ces étiquettes dans les cas hors contextes et contextuels.

3.3 Modèles statistiques

Un modèle statistique M_E est associé à chaque étiquette E. Le modèle statistique correspondant à l'étiquette correcte résulte des statistiques effectuées sur des segments corrects, donc appartenant à des solutions correctes. Le modèle statistique correspondant à l'étiquette incorrecte résulte des statistiques effectuées sur les segments incorrects.

Chaque modèle statistique est représenté par une fonction de densité de probabilité multi-gaussiennes à matrices de covariance diagonales. Le nombre NG de gaussiennes est déterminé dans la phase d'apprentissage. Les vecteurs des moyennes et les matrices de covariance diagonales des gaussiennes sont estimés sur des données d'apprentissage. Les vecteurs acoustiques du segment s, obtenus sur le corpus d'apprentissage, seront répartis entre les différentes gaussiennes du modèle statistique associé au segment s. Pour cette répartition un classificateur probabiliste est utilisé. Ce qui va être détaillé par la suite.

4 Paramètres segmentaux

4.1 Introduction

Pour profiter des informations dynamiques qui peuvent être véhiculées dans un segment s, ce dernier est divisé en trois parties [Leung, 88] : la partie début s_d, la partie milieu s_m et la partie finale s_f. Ces différentes parties sont déterminées par l'alignement sur un modèle de Markov ou bien par une division arbitraire en trois parties du segment courant.

L'exemple suivant présente les différentes parties du segment "IN" obtenues à partir de l'alignement sur le modèle de Markov.

L'étiquette du segment est donnée par le nom de l'unité alors que les différentes parties du segment sont déterminées à partir des indices des fonctions à l'intérieur de cette unité. On distingue trois sortes d'indices : l'indice de début de phonème **fexxx** (exemple fex_alv), l'indice de fin de phonème **fsxxx** (exemple fs_vel) et l'indice de milieu de phonème **fxxx** (exemple fv1, fv2 ou fv3).

Modèle	Fonctions	Trames acoustiques		
IN	in.fex_alv	**		
	in.fv1		*	
	in.fv3		***	
	in.fs_vel			**
Etiquettes →		< IN_d >	< IN_m >	< IN_f >

Figure IX.4 Les différentes parties du segment IN

4.2 Vecteurs des coefficients

Le vecteur global de coefficients V, pour un segment donné s est défini par ($\oplus$ désigne la concaténation de vecteurs) :

$$V = V_c \oplus V_d \tag{IX.4}$$

où V_c est le vecteur de coefficients acoustiques et V_d le vecteur des coefficients temporels.

4.2.1 Coefficients temporels

Pour profiter des informations temporelles véhiculées par un segment, on introduit dans le vecteur des paramètres des informations d'ordre temporel [Levinson, 86], [Suaudeau, 92], [André-Obrecht, 93]. Les paramètres temporels choisis sont la durée globale D_S d'un segment s (exprimée en nombre de trames) et la durée relative (exprimée en pourcentage) de chaque partie constituant le segment (D_d pour la durée de s_d, D_m pour la durée de s_m et D_f pour la durée de s_f) par rapport à la durée globale du segment.

$$D_k = \frac{\textit{durée de la partie k du segment}}{D_s} \tag{IX.5}$$

$$k \in \{\textit{début}, \textit{milieu}, \textit{fin}\}$$

Le vecteur de durée pour un segment donné, est composé de :

$$V_d = (D_s, D_d, D_m, D_f) \tag{IX.6}$$

4.2.2 Coefficients acoustiques

On calcule les vecteurs moyens de chacune des parties du segment. Nous obtenons ainsi 3 vecteurs moyens de 9 coefficients (l'énergie et les 8 coefficients cepstraux) par segment. Nous avons choisi de modéliser la probabilité d'émission d'un segment par des lois multigaussiennes à matrices de covariance diagonales. Ce choix implique que les coefficients d'entrée soient non corrélés entre eux (ou le moins possible). Il est donc nécessaire d'appliquer une transformation simple sur ces trois vecteurs pour décorréler les données sans toutefois perdre de l'information. Le nouveau vecteur de 3x9 coefficients V_c est alors constitué de la manière suivante :

Si C_d désigne le vecteur moyen calculé sur le début du segment s_d, C_m désigne le vecteur moyen calculé sur le milieu du segment s_m, et C_f désigne le vecteur moyen calculé sur la fin du segment s_f alors ($\oplus$ désigne la concaténation de vecteurs) :

$$V_c = C_m \oplus (C_f - C_d) \oplus (C_f + C_d - 2 * C_m) \qquad \text{(IX.6)}$$

Sous l'hypothèse que les vecteurs moyens (C_d, C_m et C_f) suivent la même loi statistique, les espérances des produits croisés des vecteurs (C_m), ($C_f - C_d$) et ($C_f + C_d - 2 * C_m$) sont nulles. Les coefficients du vecteur V_c, sont donc décorrélés au premier ordre.

($C_f - C_d$) représente une approximation de la dérivée première et ($C_f + C_d - 2* C_m$) une approximation de la dérivée seconde.

5 Modélisation d'un mot

Le mot (ou la séquence de mots) est défini comme une concaténation de segments. Le modèle global d'un mot peut être représenté par un modèle de Markov caché. Les paramètres du modèle sont :

NQ : le nombre d'états, il sera égal au nombre de segments dans le mot plus 1 (K+1).

$\{q_i ; i=1 .... NQ\}$ l'ensemble des états du modèle.

$\{t_{i,j} ; i,j= 1 ..... NQ\}$ l'ensemble des transitions du modèle. On n'autorise que les transitions correspondant aux différentes prononciations possibles.

La figure suivante IX.5 illustre ce type de modèle dans le cas d'une seule prononciation.

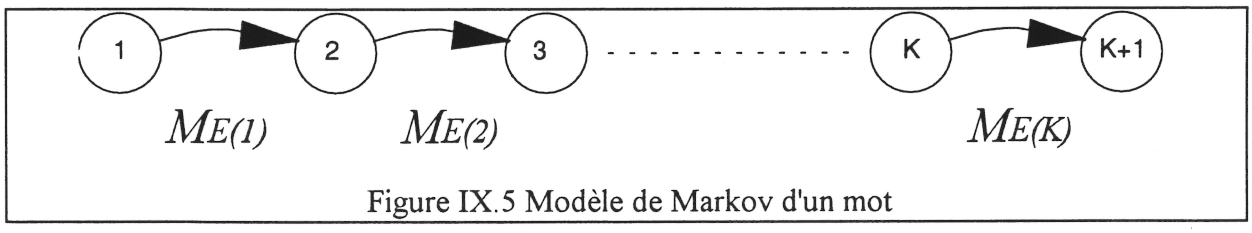

Figure IX.5 Modèle de Markov d'un mot

A titre d'exemple, la figure IX.6 présente le modèle de Markov du mot "5" et ses différentes prononciations :

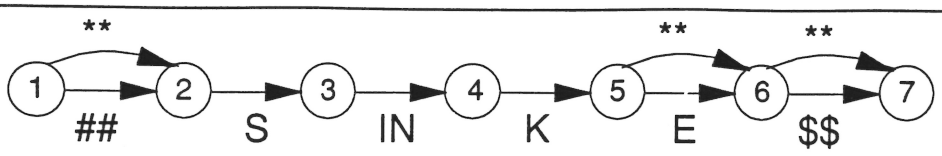

Figure IX.6 Modèle de Markov du mot "5"

Les transitions étiquetées par ** sont des transitions vides. Elles permettent l'omission du segment associé à la transition non vide reliant les deux mêmes états. Ici les segments "##", "E" et "$$" pouront donc être omis.

Comme on l'a décrit précédemment pour les segments, on associe à chaque mot deux modèles statistiques :

* Un modèle correct ($M_{E\ correct\ w_j}$) représentant les statistiques du mot prononcé w_j quand il est correctement reconnu.

* Un modèle incorrect ($M_{E\ incorrect\ w_j}$) représentant les statistiques de tous les mots prononcés w_{i_k} $(i_k \neq j)$ quand ils sont incorrectement reconnus comme w_j.

Nous avons modélisé toutes les observations qui correspondent à une reconnaissance incorrecte de w_j (w_j est la réponse incorrecte) dans un seul modèle incorrect.

6 Calcul des probabilités des solutions

6.1 Calcul de la probabilité d'émission d'une solution

Nous cherchons à calculer la probabilité $P(w_j\ prononcé\ /\ X)$ du j^{ième} mot w_j prononcé connaissant l'observation X. La règle de Bayes permet d'écrire :

$$P(w_j\ prononcé\ /\ X) = \frac{P(X\ /\ w_j\ prononcé) * P(w_j\ prononcé)}{P(X)} \qquad \text{(IX.7)}$$

où $P(X\ /\ w_j\ prononcé)$ est la probabilité d'avoir X sachant que le mot w_j est prononcé, $P(w_j\ prononcé)$ est la probabilité a priori d'avoir prononcé le mot w_j et $P(X)$ la probabilité a priori d'avoir l'observation X.

En supposant que les mots sont équiprobables et puisque $P(X)$ est commun à tous les mots du vocabulaire, on a :

$$\underset{j}{ArgMax}\, \mathrm{P}\left(w_j \ prononcé \ / \ X\right) = \underset{j}{ArgMax}\, \mathrm{P}\left(X \ / \ w_j \ prononcé\right) \qquad \text{(IX.8)}$$

Cependant X n'est pas une observation unique mais un ensemble d'observations $\tilde{X}$ (i.e. l'équation IX.1), d'où :

$$\mathrm{P}\left(\tilde{X} \ / \ w_j \ prononcé\right) = \mathrm{P}\left(X_{i_1}, X_{i_2}, X_{i_3}, \ldots, X_{i_N} \ / \ w_j \ prononcé\right) \qquad \text{(IX.9)}$$

avec $j = \{i_1, i_2, \ldots, i_N\}$.

Vu que les composantes de l'observation $\tilde{X}$ sont supposées indépendantes, on obtient :

$$\mathrm{P}\left(\tilde{X} \ / \ w_j \ prononcé\right) = \mathrm{P}\left(X_{i_1} \ / \ w_j \ prononcé\right) * \mathrm{P}\left(X_{i_2} \ / \ w_j \ prononcé\right) * \qquad \text{(IX.10)}$$
$$\ldots\ldots \mathrm{P}\left(X_{i_N} \ / \ w_j \ prononcé\right)$$

Sur le tableau IX.1, nous présentons un exemple d'association des modèles corrects et des modèles incorrects à chaque solution. On a développé pour cela les 3 meilleures solutions.

	3-meilleures solutions		
Solutions	Solution 1	Solution 2	Solution 3
hypothèses	$w_1 \rightarrow X_{i_1}$	$w_2 \rightarrow X_{i_2}$	$w_3 \rightarrow X_{i_3}$
$w_j \overset{?}{=} w_1$	$M_{E_{correct\ w_1}}$	$M_{E_{incorrect\ w_2}}$	$M_{E_{incorrect\ w_3}}$
$w_j \overset{?}{=} w_2$	$M_{E_{incorrect\ w_1}}$	$M_{E_{correct\ w_2}}$	$M_{E_{incorrect\ w_3}}$
$w_j \overset{?}{=} w_3$	$M_{E_{incorrect\ w_1}}$	$M_{E_{incorrect\ w_2}}$	$M_{E_{correct\ w_3}}$

Tableau IX.1 Modèles des corrects et des incorrects

Si les premières solutions correspondent aux mots w_1, w_2 et w_3, l'ensemble des observations obtenues à la sortie du module markovien est $\tilde{X} = \{X_{i_1}, X_{i_2}, X_{i_3}\}$. On se place de ce fait dans trois hypothèses, pour calculer $\mathrm{P}\left(\tilde{X} \ / \ w_j \ prononcé\right)$:

La première solution est correcte, on calcule alors la probabilité :

$$\mathrm{P}\left(\tilde{X} \ / \ w_1 \ prononcé\right) = \mathrm{P}\left(X_{i_1} \ / \ M_{E_{correct\ w_1}}\right) * \mathrm{P}\left(X_{i_2} \ / \ M_{E_{incorrect\ w_2}}\right) * \mathrm{P}\left(X_{i_3} \ / \ M_{E_{incorrect\ w_3}}\right) \qquad \text{(IX.11)}$$

Par contre, si la seconde est correcte, on calcule alors :

$$\mathrm{P}\left(\tilde{X} \ / \ w_2 \ prononcé\right) = \mathrm{P}\left(X_{i_1} \ / \ M_{E_{incorrect\ w_1}}\right) * \mathrm{P}\left(X_{i_2} \ / \ M_{E_{correct\ w_2}}\right) * \mathrm{P}\left(X_{i_3} \ / \ M_{E_{incorrect\ w_3}}\right) \qquad \text{(IX.12)}$$

Et finalement si la solution correcte correspond à la troisième réponse, on calcule alors :

$$P\left(\tilde{X} / w_3 \text{ prononcé}\right) = P\left(X_{i_1} / M_{E_{incorrect\ w_1}}\right) * P\left(X_{i_2} / M_{E_{incorrect\ w_2}}\right) * P\left(X_{i_3} / M_{E_{correct\ w_3}}\right) \quad \text{(IX.13)}$$

Comme dans la phase de reconnaissance, on ne connaît pas forcément la solution correcte, on est ramené à calculer les équations IX.11 à IX.13. La solution correcte est celle ayant le score de probabilité le plus élevé.

On peut étendre ce principe pour un nombre N de solutions. Pour chaque solution, on va calculer le produit de la probabilité d'avoir cette solution sachant son modèle correct par les probabilités d'avoir les autres solutions par leurs modèles incorrects respectifs.

Pour simplifier les calculs, on divise les équations IX.11 à IX.13 par :

$$P\left(X_{i_1} / M_{E_{incorrect\ w_1}}\right) * P\left(X_{i_2} / M_{E_{incorrect\ w_2}}\right) * P\left(X_{i_3} / M_{E_{incorrect\ w_3}}\right) \quad \text{(IX.14)}$$

Le calcul de score de chaque solution se réduit alors à :

$$R_{i_j} = Log \frac{P\left(X_{i_j} / M_{E_{correct\ w_j}}\right)}{P\left(X_{i_j} / M_{E_{incorrect\ w_j}}\right)} \quad \text{(IX.15)}$$

R_{i_j} représente le logarithme du rapport (rapport de vraisemblance) de la probabilité d'émission de l'observation X_{i_j} par le modèle correct sur la probabilité d'émission de X_{i_j} par le modèle incorrect.

6.2 Probabilité d'émission d'un segment

Nous avons décrit dans le paragraphe 5, un modèle de Markov d'un mot donné en fonction de ses segments. Nous nous intéressons ici à la détermination de la probabilité d'émission d'un segment. Nous développons les calculs pour un segment étiqueté correct. Pour un segment étiqueté incorrect, les calculs sont identiques.

Nous nous intéressons à la détermination de la probabilité que le k^{ième} vecteur de coefficients $V_{i_j}(k)$ associé au segment $s_{i_j,k}$ de l'observation X_{i_j}, soit émis par le modèle statistique des corrects $M_{E_{correct\ s_{i_j,k}}}$.

Cette probabilité est calculée en recherchant la gaussienne du modèle qui fournit la probabilité maximale d'observer le vecteur des coefficients $V_{i_j}(k)$ du segment $s_{i_j,k}$ connaissant son modèle statistique $M_{E_{correct\ s_{i_j,k}}}$, sur l'ensemble des gaussiennes appartenant au modèle statistique :

$$P\left(V_{i_j}(k) / M_{E_{correct\ s_{i_j,k}}}\right) = \max_{f=1}^{NG} \lambda_f * g\left(V_{i_j}(k), \mu_f, \Sigma_f\right) \quad \text{(IX.16)}$$

NG est le nombre de gaussiennes dans le modèle $M_{E_{correct\ s_{i_j,k}}}$, λ_f le poids de la gaussienne *f*, μ_f et Σ_f sont respectivement le vecteur des moyennes et la matrice de covariance diagonale de la *fième* gaussienne du modèle statistique $M_{E_{correct\ s_{i_j,k}}}$ associé à l'étiquette $E_{correct\ s_{i_j,k}}$.

La valeur de la *fième* gaussienne est déterminée par :

$$g_{correct\ s_{i_j,k}}\left(V_{i_j}(k),\mu_f,\Sigma_f\right) = \frac{1}{\left|\Sigma_f\right|^{1/2}(2*\pi)^{n/2}} * e^{-\frac{1}{2}\left(\left(V_{i_j}(k)-\mu_f\right)^T \Sigma_f^{-1}\left(V_{i_j}(k)-\mu_f\right)\right)} \qquad \text{(IX.17)}$$

où n représente la dimension du vecteur des coefficients $V_{i_j}(k)$.

La pondération de chaque gaussienne est déterminée lors de la phase d'apprentissage à partir de sa fréquence d'utilisation.

6.3 Score segmental d'une solution

A partir du score des segments et des modèles de mots, on calcule le score segmental des solutions. Ce score prend en compte ici les différents segments constituant cette solution.

Pour la modélisation décrite par la figure IX.2, on a :

$$\mathrm{P}\left(X_{i_j} / M_{E\ correct\ w_j}\right) = \mathrm{P}\left(V_{i_j}(1), V_{i_j}(2), \ldots\ldots, V_{i_j}\left(K_{i_j}\right) / M_{E\ correct\ w_j}\right) \qquad \text{(IX.18)}$$

Sous l'hypothèse de l'indépendance entre les segments d'un même mot et en ignorant les probabilités de transition entre les états du modèle, on peut écrire :

$$\mathrm{P}\left(X_{i_j} / M_{E_{correct\ w_j}}\right) = \prod_{k=1}^{K_{i_j}} \mathrm{P}\left(V_{i_j}(k) / M_{E_{correct\ s_{i_j,k}}}\right) \qquad \text{(IX.19)}$$

où $M_{E_{correct\ s_{i_j,k}}}$ représente le modèle des statistiques correctes du k^{ième} segment de la j^{ième} solution.

En introduisant les statistiques des modèles corrects et incorrects dans l'équation IX.15, on obtient :

$$R_{i_j} = Log\left(\frac{\prod_{k=1}^{K_{i_j}} \mathrm{P}\left(V_{i_j}(k) / M_{E\ correct\ s_{i_j,k}}\right)}{\prod_{k=1}^{K_{i_j}} \mathrm{P}\left(V_{i_j}(k) / M_{E\ incorrect\ s_{i_j,k}}\right)}\right) \qquad \text{(IX.20)}$$

Ce qui peut encore se formuler par :

$$R_{i_j} = \sum_{k=1}^{K_{i_j}} Log\left(\frac{P\left(V_{i_j}(k) / M_{E\ correct\ s_{i_j,k}}\right)}{P\left(V_{i_j}(k) / M_{E\ incorrect\ s_{i_j,k}}\right)}\right) \quad \text{(IX.21)}$$

7 Caractéristiques de l'application

Lors de la phase d'apprentissage, les caractéristiques de la modélisation doivent être spécifiées. Ainsi, on peut choisir de tenir compte ou non du silence, du contexte et de la durée d'un enregistrement. Nous allons voir ces caractéristiques plus en détail dans ce paragraphe.

7.1 Silence

La modélisation des mots du vocabulaire inclut un silence de début et de fin d'enregistrement (i.e. précédent et suivant le mot). On s'intéresse dans ce paragraphe à la prise en compte ou non de ces segments de silence. La figure suivante (IX.7) donne les portions du signal prises en compte dans les deux cas (silence ou hors silence) pour le mot prononcé "5". La segmentation a été déterminée par un modèle de Markov.

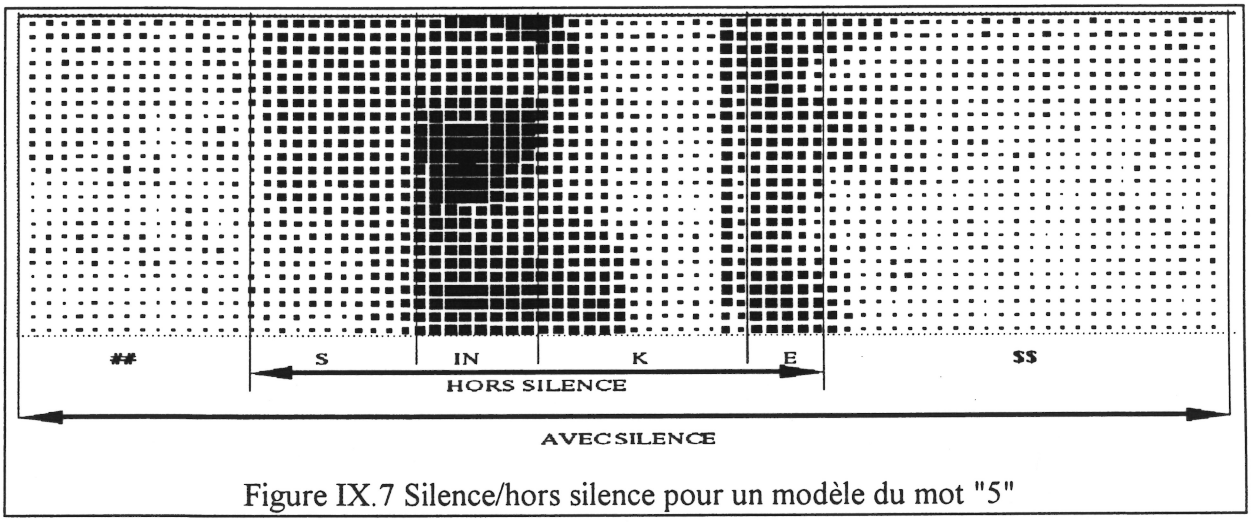

Figure IX.7 Silence/hors silence pour un modèle du mot "5"

Dans le cas hors silence l'enregistrement sera écourté de ses trames de silence identifiées par les étiquettes silence de début et silence de fin. De ce fait les modèles statistiques associés aux silences ne seront pas pris en compte, nous aurons ainsi pour l'exemple de la figure IX.7 les segments suivants :

"5"	"S"."IN"."K"."E"	hors silence
"5"	"##"."S"."IN"."K"."E"."$$"	avec silence

7.2 Contexte

On s'intéresse à la prise en compte (contexte dépendant) ou non (contexte indépendant) des phonèmes qui précèdent et qui suivent le phonème courant.

L'exemple suivant présente les différentes étiquettes retenues pour le calcul statistique. Pour chaque étiquette du mot "5" est définie sa correspondance sans la prise en compte du contexte puis en tenant compte des contextes gauche et droit (phonème de gauche - phonème courant - phonème de droite). Le symbole (**) représente l'absence d'unité à cet endroit.

Segments	Etiquettes hors contexte	Etiquettes contextuelles
##	**Silence de début**	** --**Silence**-- S
S	**S**	Silence-- **S** -- IN
IN	**IN**	S -- **IN** -- K
K	**K**	IN -- **K** -- E
E	**E**	K -- **E** --Silence
$$	**Silence de fin**	E --**Silence**-- **

Tableau IX.2 Contexte/hors contexte, modèle du mot 5

Dans le cas contextuel, en phase d'apprentissage, les paramètres des modèles sont calculés à partir des statistiques réalisées avec et sans prise en compte du contexte. En effet, à cause des variantes de prononciation, certains contextes peuvent être présents sur le corpus de test sans l'être sur le corpus d'apprentissage. Ces unités "fantômes" seront alors remplacées par le modèle associé au phonème courant hors contexte.

Dans le cas hors silence, on ne tient pas compte des étiquettes "Silence de début" et "Silence de fin" dans le cas hors contexte, et "**-Silence-S" et "E-Silence-**" dans une modélisation contextuelle, pour l'exemple traité dans le tableau IX.2.

7.3 Normalisation de la durée

On calcule des statistiques soit sur la durée relative soit sur la durée absolue d'un segment.

La durée relative est définie comme le rapport de la durée du segment sur la durée totale du mot auquel appartient le segment. La durée absolue (non normalisée) n'est autre que la durée effective du segment.

8 Recherche de la combinaison optimale

Le module markovien a permis de déterminer, pour un mot prononcé donné, *N* solutions avec leurs scores respectifs. Pour chaque solution, nous avons déterminé un score segmental. La prise en compte du score de post-traitement segmental seul ne permettant pas d'obtenir de meilleures performances que celles obtenues par le module markovien, cela nous a amené à combiner le score du module markovien avec les scores obtenus par le post-traitement segmental.

Nous recherchons une combinaison efficace entre ces deux scores pour avoir la bonne solution en première place. La combinaison retenue est du type linéaire :

score final de la $i^{ième}$ solution = α * score markovien (i) + (1-α) score segmental (i)

La solution déclarée reconnue est la solution ayant le score final le plus élevé. La valeur optimale du paramètre α est celle qui permet d'obtenir le taux d'erreur minimum sur le corpus d'apprentissage.

Nous décrivons ci-dessous, à travers un exemple, la méthode de combinaison des scores de probabilité obtenus à la sortie des deux modules du système de reconnaissance. Cet exemple ne s'applique que si le nombre de solutions *N* est égal à 2.

Pour cet exemple, le module markovien fournit les 2 meilleures solutions. Et on fait l'hypothèse que la bonne solution est toujours parmi les 2 développées. Deux cas se présentent :

Premier cas : la première solution est la solution correcte ; dans ce cas le module markovien a raison.

Deuxième cas : la solution correcte n'est pas la première solution, mais la seconde ; dans ce cas le module markovien a tort.

Les mêmes hypothèses s'appliquent au post-traitement segmental, ce qui nous donne 4 situations possibles représentées sur la figure IX.8. La signification des axes est la suivante :

*** Axe des X ; module markovien**

Δ_m = score de la bonne réponse - score de la mauvaise réponse

*** Axe des Y ; module post-traitement**

Δ_p = score de la bonne réponse - score de la mauvaise réponse

La signification des 4 zones de la figure IX.8 est la suivante :

Zone I, les deux modules ont raison et dans tous les cas on obtient la bonne solution. Dans ce cas la bonne solution est en première position, Δ_m et Δ_p sont tous les deux positifs. Quelque soit la valeur de α, la bonne solution n'est pas perdue.

Zone II, le module markovien a raison mais le post-traitement a tort. Ce cas correspond à Δ_m positif et Δ_p négatif. Il faut bien choisir la valeur de α pour que le post-traitement segmental n'influe pas sur la décision markovienne.

Zone III, les deux modules ont tort et dans ce cas la bonne solution ne sera jamais récupérée. Δ_m et Δ_p sont tous les deux négatifs, quelque soit α, on ne peut avoir la bonne solution.

Zone IV, le module markovien a tort et le post-traitement a raison. C'est l'inverse du cas de la zone II, c'est là qu'on doit voir l'intérêt d'utiliser un post-traitement.

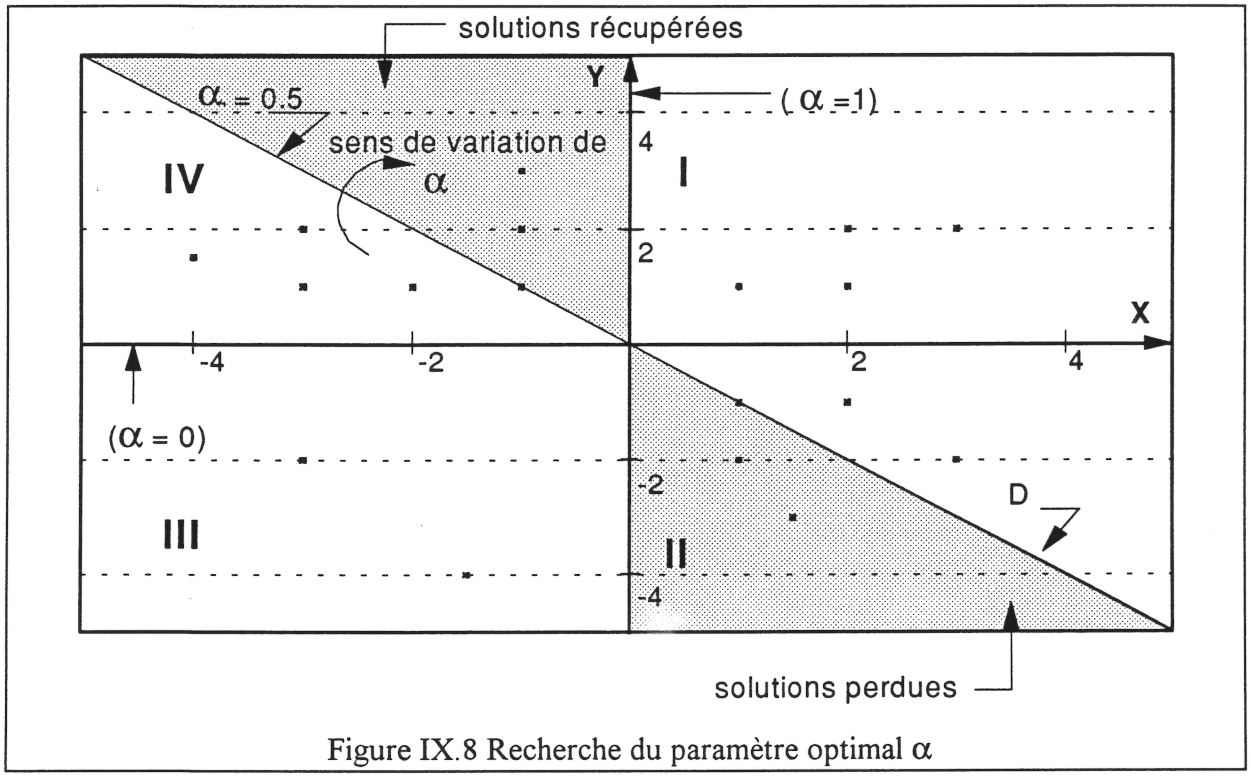

Figure IX.8 Recherche du paramètre optimal α

La zone I est exclue car les deux modules trouvent toujours la bonne solution. La zone III, correspond à une partie de la figure où l'on ne peut rien faire. Le module markovien et le module de post-traitement ont tous les deux tort. Ils ne peuvent récupérer de solutions séparément ou ensemble. Cette zone III est automatiquement exclue elle aussi. La méthode de combinaison intervient finalement sur les zones II et IV. Le principe de cette méthode est de définir à partir d'une droite D, une surface ayant le minimum de solutions dans la zone II hachurée (solutions perdues), et le maximum de solutions dans la zone IV hachurée (solutions récupérées).

Ces deux surfaces sont déterminées comme suit à partir du score final des solutions 1 et 2 :

$$Sf(1) = \alpha * score\ markovien(1) + (1-\alpha) * score\ segmental(1)$$
$$Sf(2) = \alpha * score\ markovien(2) + (1-\alpha) * score\ segmental(2)$$

où *Sf(1)* est le score final de la première solution et *Sf(2)* celui de la seconde solution.

On va examiner les deux cas de figure suivants :

1/ Lorsque la bonne solution est en deuxième position, on veut que son score final soit plus élevé que le score de la première solution. Pour cela, on doit trouver une valeur de α permettant de la passer en première place.

La différence de score entre la bonne solution et la mauvaise solution est définie par :

$$\Delta_{score} = Sf(2) - Sf(1) = \alpha * X + (1 - \alpha) * Y$$

avec $X = \Delta_m = score\ markovien(2) - score\ markovien(1)$

et $Y = \Delta_p = score\ segmental(2) - score\ segmental(1)$

Puisque la bonne est en seconde position alors $X < 0$, donc on se situe sur le cadrant III ou IV. Il reste à connaître la valeur de Y. Si la valeur de Y est négative alors, on se retrouve dans la zone III, et la combinaison ne sert à rien. Le module markovien a tort et le post-traitement de même. Dans le cas où Y est positif, on va se situer dans la zone IV, et là tout dépend de la combinaison. Dans ce cas de figure, pour que la bonne solution soit récupérée il faut que $\Delta_{score} \geq 0$ (surface supérieure par rapport à la droite D), ce qui nous ramène à l'équation :

$$Y > \frac{\alpha}{\alpha - 1} * X \qquad \text{(IX.17)}$$

$\frac{\alpha}{\alpha - 1}$ est la pente de la droite D.

2/ Lorsque la bonne solution est en première position, on veut que son score final soit plus élevé que le score de la deuxième solution. Pour cela, on doit trouver une valeur de α permettant de la garder en première place.

La différence de score entre la bonne solution et la mauvaise solution est définie par :

$$\Delta_{score} = Sf(1) - Sf(2) = \alpha * X + (1 - \alpha) * Y$$

avec $X = \Delta_m = score\ markovien(1) - score\ markovien(2)$

et $Y = \Delta_p = score\ segmental(1) - score\ segmental(2)$

Puisque la bonne est en première position alors $X > 0$, donc on se situe sur le cadrant I ou II. Il reste à connaître la valeur de Y. Si la valeur de Y est positive alors, on se retrouve dans la zone I, on n'a donc plus besoin de la combinaison, puisque le module markovien a raison et le post-traitement de même. Dans le deuxième cas de figure où la valeur de Y est négative. Le module markovien a raison et le post-traitement a tort. Il faut donc trouver la valeur de α qui permette de garder la première solution, pour cela il faut que $\Delta_{score} \geq 0$ (surface supérieure par rapport à la droite D), ce qui nous ramène à l'équation :

$$Y > \frac{\alpha}{\alpha - 1} * X \qquad \text{(IX.18)}$$

En phase de reconnaissance on ne sait pas où se trouve la bonne solution, il nous faut donc rechercher la meilleure combinaison pour garder la bonne solution si elle est à la première place et prendre en compte la seconde solution si cette dernière est la bonne. Pour cela, on recherche la position de la droite D sur les zones II et IV qui permet de récupérer le maximum de solutions sur l'ensemble d'apprentissage, tout en perdant le moins possible.

($\alpha = 0$) correspond à la prise en compte du module de post-traitement seul (Axe X sur la figure IX.8) et ($\alpha = 1$) correspond à la prise en compte du module markovien seul (Axe Y).

On a présenté l'exemple où ($\alpha = 0.5$), en indiquant les deux zones de récupération et de perte de solutions.

La valeur optimale[1] du paramètre α (déterminée sur l'ensemble d'apprentissage) est utilisée sur le corpus de test en phase de reconnaissance.

9 Conclusion

Nous avons décrit dans ce chapitre notre approche du post-traitement segmental statistique. Cette approche a permis de trouver une solution théorique au problème des différentes observations associées à un mot prononcé donné (la segmentation propre à chaque solution proposée). Nous avons mis en évidence l'apport théorique des modèles corrects et incorrects pour le calcul final d'une solution, nous allons décrire dans le chapitre suivant le coté pratique, en présentant les résultats obtenus pour diverses configurations des paramètres de la modélisation.

---

[1] Pratiquement, on fait varier la valeur de α entre 0 et 1 avec un pas déterminé (par exemple 0.001). Pour chaque valeur de α, on calcule le taux d'erreur obtenu sur le corpus d'apprentissage. La valeur de α utilisée sur le corpus de test est celle ayant permis d'obtenir le taux d'erreur le plus bas sur le corpus d'apprentissage.