Introduction de l'énrgie dans un modèle de reconnaissance automatique de la parole

Abdellah Yousfi & Abdelouafi Meziane

Départment de Mathématiques
Université Mohamed Premier,
Faculté des Sciences, Oujda, Maroc
Mail: {yousfi.abdellah,meziane}@sciences.univ-oujda.ac.ma

ABSTRACT

A major deficiency of standard Hidden Markov Models (HMM) is that both the spectral and the prosodic feature are uniformly processed. To combine more efficiently the prosodic cues with the acoustic ones, a segmental two Level Hidden Markov Model has been recently studied by suaudeau [Suaudeau 94].
In this paper, we present an adapted version of this model in wich the segmental processing is replaced by the classical centisecond processing. This new model is called Two Level Hidden Semi Markov Centisecond Model (TLHSMCM)). Our approach retains the traditionnal hierarchical structure of an HMM, and facilitate the introduction of others prosodic parameters (in particular the energy) in the phonetic level.
Experiments on a french database composed of 20 numbers show that this model reduces the recognition error rates.

Keywords :
Hidden Markov Model(HMM), Energy, Two Level Hidden Semi Markov Centisecond Model(TLHSMCM).

1. INTRODUCTION

Le semi-modèle de Markov caché à deux niveaux centiseconde (SMMCDNC) est un modèle inspiré des deux modèles suivants :
le modèle de Markov caché à deux niveaux segmental (MMCDN) [Suaudeau 94] et le modèle de Markov caché à deux niveaux centiseconde (MMCDNC) [Meziane 97].
Le modèle MMCDN développé par suaudeau est très difficile à utiliser dans le pratique[1]. De ce fait nous avons développé ce moèle pour faciliter l'introduction de nouveaux paramètres prosodiques (autre que la durée, en particulier l'énergie) au niveau phonétique.
L'application de ce modèle nécessite d'abord, comme le modèle de Markov caché standard (MMC) [Rabiner 89, Jouvet 88], la construction d'un réseau probabiliste suivant une structure hiérarchique. Chaque mot du vocabulaire est représenté par une transcription d'unités phonétiques, puis chaque unité est associée à un modèle acoustique.
Le SMMCDNC suppose que le long d'un chemin et à chaque changement d'unité phonétique on émet un vecteur d'observations phonétiques de dimension supérieur à un suivant une loi de probabilité.
Au cours de ce papier, nous allons définir ce modèle, ensuite donner des application de ce dernier.

[1]Dans la formule du maximum de vraisemblance, il intervient un facteur qui ne peut pas être calculé facilement

L'analyse acoustique est faite sur des tarmes de longueurs fixe (32ms), pour obtenir deux types de vecteurs d'observations, un vecteur d'observations acoustiques (noté y) qui contient les paramètres spectraux (par exemple les MFCC), et un autre que nous appelons vecteur d'observations pseudo-phonétiques noté z[2], il contient des paramètres qui sont utilisés pour le calcul du vecteur d'observations phonétiques. Les vecteurs d'observations acoustiques et pseudo-phonétiques sont modélisés au niveau acoustique et les vecteurs d'observations phonétiques sont modélisés au niveau phonétique.

2. DÉFINITION DU MODÈLE

Un semi-modèle de Markov caché à deux niveaux centiseconde, est un modèle défini à partir de cinq processus stochastiques : $(X_t)_{t\geq 1}$, $(Y_t)_{t\geq 1}$, $(Z_t)_{t\geq 1}$, $(\Lambda_\tau)_{\tau\geq 1}$, $(P_\tau)_{\tau\geq 1}$.

• $(Y_t)_{t\geq 1}$ est un processus observable représentant les observations acoustiques, à valeurs dans un ensemble mesurable Y.

• $(\wedge_\tau)_{1\leq\tau\leq\epsilon} = \{(\phi_{k_i}, \theta_i) \mid i = 1,..,\epsilon\}$ est une fonction déterministe de $(X_t)_{t\geq 1}$ à valeurs dans un ensemble de dimension deux :

$$\wedge_\tau = (\Phi_\tau, \Theta_\tau)$$

- ϵ représente le nombre total d'unités phonétiques traversées lorsqu'on parcourt la suite d'états $(X_t)_{1\leq t\leq T}$.
- $\Phi_\tau = \phi_k$ à valeurs dans l'ensemble fini Σ des unités phonétiques élémentaires, $\Sigma = \{\phi_1, ..., \phi_k\}$, Φ_τ représente la τ^{eme} unité phonétique traversée lorsque nous empruntons la suite d'états $(X_t)_{t\geq 1}$ dans le sens des indices temporels croissants.
- $\Theta_\tau = \theta_\tau$ représente l'indice temporel du 1^{er} état de la suite $(X_t)_{t\geq 1}$ issu de la τ^{eme} unité. Cet indice correspond à un instant de changement de modèles acoustiques élémentaires.
- les valeurs Φ_τ et Θ_τ sont liées avec la suite d'états $(X_t)_{t\geq 1}$ par la relation R :

$$R(X_{\theta_\tau - 1}) \neq R(X_{\theta_\tau})$$
$$R(X_{\theta_\tau}) = ... = R(X_{\theta_{\tau+1}-1}) = \Phi_\tau = \phi_k$$

• $(Z_t)_{t\geq 1}$ est un processus observable à valeurs dans un ensemble mesurable, il représente les observations pseudo-phonétiques. Le rôle principal de $(Z_t)_{t\geq 1}$ est le calcul du vecteur d'observations phonétiques P_τ.

[2]Dans le cas du traitement de l'énergie, l'observation pseudo-phonétique Z_t est l'énergie de la trame.

Nous supposons que le processus Z_t vérifie l'hypothèse suivante :

H) $Pr(Z_t = z_t \mid Z_{t-1} = z_{t-1}, ..., Z_1 = z_1, Y_1 = y_1, ..., Y_t = y_t, X_1 = q_{i_1}, ..., X_t = q_{i_t}) = Pr(Z_t = z_t \mid X_t = q_{i_t}) = g_{i_t}(z_t)$

• $(P_\tau)_{\tau \geq 1}$ est un vecteur aléatoire de dimension q, il représente les observations phonétiques liées à l'unité phonétique $\Phi_\tau = \phi_k$. Ce vecteur est associé à une loi de probabilité notée f_k. Les observations pseudo-phonétiques émisent dans l'unité ϕ_k sont $Z_{\theta_\tau}, ..., Z_{\theta_{\tau+1}-1}$. Il existe une application F tel que P_τ est donné par :

$$P_\tau = F(Z_{\theta_\tau}, ..., Z_{\theta_{\tau+1}-1})$$

• X_t est un processus semi-markovien à valeurs dans l'ensemble des états $Q = \{q_1, .., q_N\}$, et $q_{i_1}, ..., q_{i_T}$ un chemin du résau global sur lequel les observations $y_1, .., y_T$ et $z_1, .., z_T$ peuvent être émises. On suppose que les transitions entre deux états successives, appartenants à ce chemin, vérifient :
$Pr(\ X_{t+1} = q_{i_{t+1}} \mid \ X_t = q_{i_t}, ..., X_1 = q_{i_1}, R(q_{i_{t+1}}) = R(q_{i_t})) = a_{i_t i_{t+1}}$
$Pr(\ X_{t+1} = q_{i_{t+1}} \mid \ X_t = q_{i_t}, ..., X_1 = q_{i_1}, R(q_{i_{t+1}}) \neq R(q_{i_t})) = a_{i_t i_{t+1}}.f_k(p)$
avec :
- $R(q_{i_t}) = \phi_k$, dans la suite nous utilisons les deux notations suivantes : $f_k = f_{\phi_k}$.
- p est la valeur mesurée du vecteur d'observations phonétiques P_τ sur l'unité ϕ_k, elle est donnée par :
$p = F(z_{\theta_\tau}, ..., z_t)$, avec θ_τ est le premièr instant d'entrer dans l'unité phonétique ϕ_k et $t = \theta_{\tau+1} - 1$ (le premièr instant d'entrer dans l'unité phonétique $R(q_{i_{t+1}})$).
- Pour toute $i = 1, ..., N$ on a :

$$Pr(X_1 = q_i) = \pi_i$$

Conclusion :
Ce nouveau modèle est défini entièrement par un vecteur de paramètres noté
$\lambda = (\Pi, A, B, G, \Delta)$.
• $\Pi = \{\pi_1, ..., \pi_N\}$ l'ensemble des probabilités initiales.
• $A = (a_{ij})_{1 \leq i,j \leq N}$ la matrice des probabilités de transitions entre les états du modèle.
• $B = \{b_i \mid 1 \leq i \leq N\}$ l'ensemble des lois de probabilités des observations acoustiques.
• $G = \{g_i \mid 1 \leq i \leq N\}$ l'ensemble des lois de probabilités des observations pseudo-phonétiques.
• $\Delta = \{f_k \mid 1 \leq k \leq K\}$ l'ensemble des lois de probabilités des observations phonétiques.

3. VRAISEMBLANCE D'UNE SUITE D'OBSERVATIONS

Soient $Y = y_1, .., y_T$ et $Z = z_1, ..., z_T$ deux suites d'observations acoustiques et pseudo-phonétiques générées par le modèle SMMCDNC et associées à la suite phonétique $(\wedge_\tau)_{1 \leq \tau \leq \epsilon}$, la probabilité d'émettre les suites d'observations Y et Z suivant un chemin $q_{i_1}, ..., q_{i_T}$, est donnée par :

$$\begin{aligned}
&Pr(y_1, .., y_T, z_1, ..., z_T, q_{i_1}, ..., q_{i_T}) = \\
&\pi_{i_1} \times b_{i_1}(y_1) \times g_{i_1}(z_1) \times \\
&\times \prod_{n=2}^{\theta_2} a_{i_{n-1} i_n} b_{i_n}(y_n) \times g_{i_n}(z_n) \times f_{k_1}(p_1) \times \\
&\times \prod_{n=\theta_2+1}^{\theta_3} a_{i_{n-1} i_n} b_{i_n}(y_n) \times g_{i_n}(z_n) \times f_{k_2}(p_2) \times \\
&\qquad \times ... \times \\
&\qquad \times ... \times \\
&\times \prod_{n=\theta_\epsilon+1}^{T} a_{i_{n-1} i_n} b_{i_n}(y_n) \times g_{i_n}(z_n) \times f_{k_\epsilon}(p_\epsilon)
\end{aligned}$$

avec :

$$p_i = F(z_{\theta_i}, ..., z_{\theta_{i+1}-1}) \qquad i = 1, ..., \epsilon$$

La vraisemblance conjointe de Y et Z est donnée par la formule suivante :
$Pr(y_1, ..., y_T, z_1, ..., z_T) = \sum_{q_{i_1}, ..., q_{i_T}} Pr(y_1, .., y_T, z_1, ..., z_T, q_{i_1}, ..., q_{i_T})$

4. APPRENTISSAGE

Le but principal de l'apprentissage, est d'optimiser le vecteur des paramètres du modèle $\lambda = (\Pi, A, B, G, \Delta)$.

Pour calculer l'estimateur de ce modèle, deux méthodes peuvent être utilisées, la procédure de Baum-Welch, ou celle de Viterbi. En général on utilise la méthode de viterbi car elle est la plus simple, moins coûteuse en mémoire et en calcul et donne des résultats comparables à ceux de Baum-Welch.

Etant donné un ensemble d'apprentissage $W = \{w_1, ..., w_R\}$ constitué de R prononciations, chaque w_i est associée aux suites d'observations acoustiques $y_1^i, ..., y_{T_i}^i$ et pseudo-phonétiques $z_1^i, ..., z_{T_i}^i$ et au chemin optimal $\xi^*(i)$. Nous avons utilisé la fonction auxilaire $Q(\lambda, \lambda')$ pour estimer ces paramètres :

$$\begin{aligned}
&Q(\lambda, \lambda') = \sum_n \sum_\xi \delta(\xi - \xi^*(n)) \times \\
&Pr_\lambda(\xi \mid y_1^n, ..., y_{T_n}^n, z_1^n, ..., z_{T_n}^n) \times \\
&\ln Pr_{\lambda'}(\ y_1^n, ..., y_{T_n}^n, z_1^n, ..., z_{T_n}^n, \xi\ ) \\
&= \sum_n Pr_\lambda(\xi^*(n) \mid \ y_1^n, ..., y_{T_n}^n, z_1^n, ..., z_{T_n}^n\ ) \times \\
&\ln Pr'_\lambda(y_1^n, ..., y_{T_n}^n, z_1^n, ..., z_{T_n}^n, \xi^*(n))
\end{aligned}$$

ξ est un chemin de longueur T_n.

L'avantage de cette fonction est qu'elle est décomposable en une somme de quatre termes indépendants :

$$Q(\lambda, \lambda') = Q_{A'}(\lambda, \lambda') + Q_{B'}(\lambda, \lambda') + Q_{G'}(\lambda, \lambda') + Q_{\Delta'}(\lambda, \lambda')$$

avec :

$$Q_{\Delta'}(\lambda, \lambda') = \sum_{n=1}^{R} \sum_{k=1}^{K} \delta(k, \xi^*(n), p_\tau^n).\ln(f'_k(p_\tau^n))$$

$$Q_{G'}(\lambda, \lambda') = \sum_{n=1}^{R} \sum_t \sum_i \delta(q_i, t, \xi^*(n)).\ln(g'_i(z_t^n))$$

avec :

$$\delta(q_i, t, \xi^*(n)) = \begin{cases} 1 & \text{si } q_i \in \xi^*(n) \text{ et l'observation } z_t^n \\ & \text{est émise de l'état } q_i \\ 0 & \text{sinon} \end{cases}$$

Une maximisation séparée de chacun de ces quatre termes conduit aux formules de réestimation des paramètres du modèle SMMCDNC.
Pour les formules des estimateurs des paramètres Π, A, B

elles sont inchangeables.

Si la loi de P_τ, associée à ϕ_{k_τ}, est gaussienne de moyenne μ_{k_τ} et de matrice de covariance C_{k_τ}, alors les estimateurs de ces paramètres sont donnés par [3]:

$$\overline{\mu}_{k_\tau} = \frac{\sum_{i=1}^{R}\sum_{l=1}^{K} p.\delta(\xi_i^*, l, \tau, p)}{\sum_{i=1}^{R}\sum_{l=1}^{K}\delta(\xi_i^*, l, \tau, p)}$$

$$\overline{C}_{k_\tau} = \frac{\sum_{i=1}^{R}\sum_{l=1}^{K}(p-\overline{\mu}_{k_\tau}).(p-\overline{\mu}_{k_\tau})'.\delta(\xi_i^*, l, \tau, p)}{\sum_{i=1}^{R}\sum_{l=1}^{K}\delta(\xi_i^*, l, \tau, p)}$$

avec :
$\delta(\xi_i^*, l, \tau, p) = 1$ si $\phi_l = \phi_{k_\tau}$ et ϕ_l est allignée sur le chemin ξ_i^*, de plus la valeur mesurée du vecteur phonétique de ϕ_k est p.
$\delta(\xi_i^*, l, \tau, p) = 0$ sinon.
Si g_i est une gaussienne de moyenne m_i et de matrice de covariance H_i alors les formules d'estimation de ces paramètres sont données par :

$$\overline{m_i} = \frac{\sum_{n=1}^{R}\sum_{t=1}^{T} z_t^n.\delta(q_i, t, \xi_n^*)}{\sum_{n=1}^{R}\sum_{t=1}^{T}\delta(q_i, t, \xi_n^*)}$$

$$\overline{H}_i = \frac{\sum_{n=1}^{R}\sum_{t=1}^{T}(z_t^n-\overline{m_i}).(z_t^n-\overline{m_i})'.\delta(q_i, t, \xi_n^*)}{\sum_{n=1}^{R}\sum_{t=1}^{T}\delta(q_i, t, \xi_n^*)}$$

5. APPLICATION

On note par :
D_τ : variable aléatoire représentant la durée de séjour dans l'unité phonétique ϕ_{k_τ}.
E_τ : variable aléatoire représentant l'énergie émise lors de l'émission de l'unité phonétique ϕ_{k_τ}.
Ces deux variables sont corrélées entre eux.
Pour appliquer ce nouveau modèle nous avons étudié les trois cas suivants :

• $P_\tau = D_\tau$, dans ce cas nous avons Z_t est constant et égale à la longueur de la trame (car le modèle est centiseconde). P_τ devient le nombre d'observations émmisent dans l'unité ϕ_{k_τ}.
On note ici que dans ce cas le SMMCDNC coincide avec le modèle de Markov caché à deux niveaux centiseconde [Meziane 97].

• $P_\tau = E_\tau$, dans ce cas Z_t est l'énergie produite lors de l'émission de l'observation y_t.

$$P_\tau = \sum_{t=\theta_\tau}^{\theta_{\tau+1}-1} Z_t$$

• $P_\tau = (D_\tau, E_\tau)$ vecteur de dimension supérieur à 1.
La loi de probabilité de P_τ, associée à l'unité phonétique ϕ_{k_τ} est notée f_k ou $f_{\phi_{k_\tau}}$, elle est supposée gaussienne de vecteur moyen μ_{k_τ}, et de matrice de covariance C_{k_τ}.

5.1. *Présentation de l'outil informatique*

Pour tester le modèle SMMCDNC, nous avons adapté les programmes développés à l'IRIT (Institut de Recherche en Informatique Toulouse) au sein de l'équipe ART.ps, par Jacob [Jacob 95] et par Meziane [Meziane 1997], en faisant les modifications nécessaires et en ajoutant de nouvelles fonctions.

[3] ces estimateurs sont obtenus par annulation de la dérivé de Q_Δ par rapport aux paramètres μ_{k_τ}, et C_{k_τ}.

5.2. *La mise en œuvre du modèle d'énergie et de la durée*

Nous avons effectué une analyse acoustique pour obtenir deux types de vecteurs d'observations :
- vecteur d'observations acoustiques constitué des 8 coefficients MFCC.
- vecteur d'observations pseudo-phonétiques contenant l'énergie des demi trames.

La fréquence d'échantillonnage est $f_e = 16KHZ$ et la longueur des trames est 32ms.

5.3. *Expérimentations*

Nous avons fait nos expériences sur un vocabulaire qui se compose de 20 nombres de 0 à 19, extrait de la base de données (BDSONS). Chaque nombre est prononcé une seule fois par 20 locuteurs (13 masculins et 7 féminins) pour former un ensemble d'apprentissage et une autre fois pour former un ensemble test.

Le modèle SMMCDNC est construit de manière hiérarchique comme dans le cas du modèle MMCDNC. Nous avons utilisé comme unité de base le pseudo-diphone [4] en plus des silences de début et de fin du mot.
- Les lois d'observations acoustiques sont supposées gaussiennes de matrice de covariance diagonale.
- La loi du vecteur d'observations pseudo-phonétiques est supposée gaussienne.
- La loi du vecteur (dureé, énergie) est supposée gausienne de matrice de covariance non diagonale.
- Pour les structures topologiques des modèles acoustiques associés aux pseudo-diphones nous avons utilisé ceux de Jacob [Jacob 95].

Résultats :

	App	Test
Le MMC seul	6.39%	7.78 %
Le SMMCDNC (la durée)	3.06%	4.17%
Le SMMCDNC (énergie)	2.5%	5.56 %
Le SMMCDNC (durée, énergie)	1.94%	5.28 %

Tableau 1 : Les taux d'erreur sur l'ensemble d'apprentissage et de test en utilisant le SMMCDNC.

Dans ces expériences nous avons traité trois cas du modèle SMMCDNC :
• Le vecteur phonétique égale la durée, dans ce cas nous obtenons les mêmes résultats que le modèle TLHMM centiseconde développé par Meziane [Meziane 97], car dans ce cas le modèle SMMCDNC coincide avec ce modèle.
• Le vecteur phonétique = l'énergie, dans ce cas nous remarquons qu'il y a une réduction du taux d'erreur par rapport au cas standard (MMC seul). Nous avons une réduction de 60.87% pour l'ensemble d'apprentissage et de 28.53% pour l'ensemble test.
• Le vecteur phonétique = (durée, énergie), dans ce cas nous obtenons une réduction du taux d'erreur de

[4] le pseudo-diphone caractérise deux types d'entité : la partie stable d'un phonème et la partie transitoire entre deux phonèmes. L'atout principal de ces unités est la modélisation du phénomène de coarticulation. Parmi les travaux utilisants ce type d'unité on cite par exemple : [André Obrecht 90], [Jovet 88], [Meziane 97].

69.64% pour l'ensemble d'apprentissage et de 32.13% pour l'ensemble test par rapport au modèle MMC standard. Ce nouveau modèle apporte une réduction du taux d'erreur (par rapport au cas du TLHMM centiseconde) de 36.60% pour l'ensemble d'apprentissage et une dégradation de 21.02% pour l'ensemble test. Nous pensons que cette dégradation du taux d'erreur est dûe à la variabilité de l'énergie qu'est très sensible aux conditions d'enregistrement :
- la variabilité de la distance entre le microphone et le locuteur,
- le changement du matériel d'enregistrement entre les ensembles d'apprentissage et de test (le changement du microphone par exemple),
- la différence de l'état du locuteur entre les deux enregistrements.
D'autre études sur ce facteur mettera surement en évidence l'intérêt de l'introduction de ce paramètre, au niveau phonétique, dans les modèle de reconnaissance automatique de la parole.

BIBLIOGRAPHIE

[André 90] R. André-Obrecht : *"reconnaissance automatique de la parole à partir de segments acoustiques et de modèles de Markov cachés "*, XVIIIémes JEP, Montréal, Mai 1990.

[Jacob 1995] B. Jacob : *"Un outil informatique de gestion de modèles de Markov cachés expérimentations en reconnaissance automatique de la parole "*, Thèse de doctorat de 3 $\circ$ cycle, universié Paul Sabatier Toulouse III, 1995.

[Jouvet 1988] D. Jouvet :*"Reconnaissance de mots connectés indépendamment do locuteur par méthodes statistiques "*, Thèse du 3 cycle, Juin 1988.

[Meziane 1997] A. Meziane : *"Introduction de la durée des sons dans un modèle de Markov caché au niveau supra-segmental "*, Thèse du 3 cycle, 19 Juillet 1997.

[Suaudeau 94] Suaudeau N., André-Obrecht R., An efficient combination of acoustic and supra-segmental informations in a speech recongnition system. ICASSP 94, April 94, pp 65-68, Vol1.

[Rabiner 1989] L. R. Rabiner : *"A tutorial on hidden Markov models and selected applications in speech recognition "* , IEEE Trans. ASSP, Februray 1989, vol. 77, no. 2, pp. 257-286.