Modélisation multi-bandes de la parole par champ de Markov

Guillaume Gravier, Marc Sigelle et Gérard Chollet

ENST-TSI & CNRS-URA 820
46, rue Barrault 75634 Paris Cedex 13
ggravier@infres.enst.fr, sigelle@tsi.enst.fr, chollet@tsi.enst.fr

Abstract

In this paper, an extension of the multi-band model that includes inter-band control of time asynchrony is described. The proposed model is based on the framework of Markov random fields. The law of the speech process is given by a parametric Gibbs distribution and a maximum likelihood parameter estimation algorithm is developed. This random field model is applied to isolated word recognition. It is shown that similar performances are obtained with the proposed model and with standard HMM techniques in the mono-band case. In a multi-band approach, results show that inter-band synchrony is an important parameter to take into account, in particular when dealing with noisy test signals.

1. Introduction

Les modèles de Markov cachés (MMC), utilisés dans la plupart des systèmes existants de reconnaissance de la parole (cf. par exemple [Jel98], chap. 2), donnent de bonnes performances. Cependant, cette modélisation souffre d'un certain nombre de limitations. En particulier, les performances des MMC se dégradent en présence de bruits additifs ou convolutifs et de distorsions, comme la réverbération. Pour éliminer les bruits convolutifs à variations lentes, comme les distorsions liées au canal téléphonique, on utilise en général la soustraction cepstrale ou le filtrage RASTA. Pour le traitement des bruits additifs, deux familles de solutions sont envisageables. Il est possible de rechercher une représentation du signal moins sensible au bruit que le représentation cepstrale ou, d'autre part, d'avoir un modèle statistique qui permette de traiter ce problème de manière efficace. Dans le premier cas, plusieurs représentations ont été proposées ces dernières années : spectre de modulation, TRAPS, analyse LDA, soustraction spectrale, etc. Dans le cadre de la modélisation statistique, des approches par MMC multi-bandes [HPT96] ont été proposées pour traiter le problème du bruit additif.

Dans l'approche multi-bandes, le signal est divisé en sous-bandes, chaque bande étant modélisée de manière indépendante à l'aide d'un MMC. Les scores partiels obtenus dans chaque bande sont ensuite recombinés. Entre deux points de recombinaison des scores, ce modèle repose sur une modélisation asynchrones des bandes. Par ailleurs, Tomlinson *et al.* ont montré l'intérêt de l'asynchronie entre les sous-bandes dans [TRM+97]. Cependant, l'hypothèse d'indépendance entre les sous-bandes dans l'approche multi-bandes ne parait pas réaliste. De plus, si les sous-bandes ne sont pas totalement synchrones, il semble cependant peu probable qu'elles soient totalement asynchrones. Notons également que l'asynchronie entre les bandes peut en partie être due au canal de transmission et ne présenter aucun intérêt pour la reconnaissance de la parole.

Nous proposons donc d'introduire des interactions entre les sous-bandes pour modéliser la synchronie spectrale en étudiant une modélisation de la parole basée sur les champs de Markov. Un tel modèle a précédemment été utilisé dans [GSC98b] pour modéliser la sortie d'un banc de filtre, les résultats obtenus étant décevants à cause de la trop grande variabilité de la représentation, et aussi en raison de l'absence d'algorithme d'estimation des paramètres. Dans cet article, nous proposons d'appliquer cette modélisation à une approche multi-bandes avec une représentation cepstrale du signal dans les bandes. Après avoir introduit le formalisme des champs Markoviens, nous décrivons dans un premier temps le modèle proposé. Nous rappelons également, section 3, l'algorithme d'estimation des paramètres et les stratégies de décodage en reconnaissance de mots isolés précédemment introduits dans [GSC98a]. Nous présentons finalement des résultats pour différentes architectures en sous-bandes et étudions le modèle en présence de bruit additif, avant de conclure.

2. Modélisation par champ Markovien

2.1. Champ de Markov et distribution de Gibbs

Un champ aléatoire X, défini sur un ensemble de sites (ou treillis) S, est Markovien si la probabilité d'observer une valeur en un site s ne dépend que d'un nombre fini de sites voisins V_s. L'ensemble des voisins est donné par un système de voisinage noté V. De manière formelle, la propriété Markovienne est donnée par $P[x_s|x_r\ \forall r \neq s] = P[x_s|x_r\ \forall r \in V_s]$.A un système de voisinage donné correspond un ensemble de cliques, une clique étant un ensemble de points du treillis mutuellement voisins. Le théorème de Hammersley-Clifford [Bes74] permet d'établir une correspondance entre un champ de Markov et un champ de Gibbs lorsqu'aucune réalisation de X n'est de probabilité nulle. La loi du champ X est alors

donnée par la distribution de Gibbs

$$P[x] = \frac{1}{Z} \exp\left(- \sum_{c \in \mathcal{C}} U_c(x) \right) \quad , \qquad (1)$$

où $U_c(x)$ correspond à un potentiel associé à la clique c, $\mathcal{C}$ désignant l'ensemble des cliques sur S pour V. La constante Z, appelée fonction de partition, assure que l'équation (1) définit une mesure de probabilité. Elle est donnée par une somme sur toutes les configurations x possibles. Les définitions précédentes montrent que la probabilité d'une configuration dépend d'un ensemble d'interactions locales, *i.e.* au niveau des cliques. On note aussi que plus l'énergie totale $U(x) = \sum U_c(x)$ est grande, moins la configuration x est probable.

Nous pouvons donc définir un modèle basé sur les champs de Markov en déterminant un système de voisinage et en définissant les potentiels pour chaque clique associé au système de voisinage choisi.

2.2. *Définition des potentiels*

Dans le MMC multi-bandes, la loi du processus (ou champ) caché X est donnée par les différents MMC en parallèle, la variable aléatoire $X_{t,k}$ ne dépendant que de $X_{t-1,k}$. On montre que cette relation a une équivalence bilatérale [Geo88] dans laquelle le voisinage du point (t,k) est donné par $\{(t-1,k),(t+1,k)\}$. Pour introduire une interaction entre les chaînes de Markov correspondant à chacune des bandes, on considérera le voisinage donné par

$$V_{t,k} = \{(t-1,k),(t+1,k),(t,l) \quad \forall l \neq k\} \ .$$

Un tel système de voisinage met en jeu deux types de cliques, *horizontales* et *verticales*, pour lesquelles nous allons définir un potentiel.

A partir d'études montrant qu'une chaîne de Markov est une distribution de Gibbs particulière [ZAZ91], nous définissons le potentiel associé aux cliques du type $\{(t-1,k),(t,k)\}$ par

$$U^{(h)}_{t,k} = \sum_{i,j} a^{(k)}_{ij} \ \delta(x_{t-1,k} = i)\, \delta(x_{t,k} = j) \quad , \qquad (2)$$

$\delta(x_{t-1,k} = i)$ prenant la valeur 1 si l'égalité est vérifiée et 0 sinon. Le paramètre $a^{(k)}_{ij}$, appelé poids de transition, est en fait homogène à $-\ln P^k(i,j)$, P^k étant la matrice de transition de la chaîne de Markov associée à la bande k.

En numérotant de 1 à N les états dans chaque bande et si l'on considère que deux bandes sont synchrones lorsque les changements d'états sont observés aux mêmes instants dans les deux bandes, alors la synchronie entre les bandes k et l peut-être modélisée en associant aux cliques $\{(t,k),(t,l)\}$ le potentiel défini par

$$U^{(v)}_{k,l} = f_{kl} \ |x_{t,k} - x_{t,l}| \ . \qquad (3)$$

En effet, lorsque le poids de synchronisation f_{kl} est grand, on favorise un comportement synchrone des bandes puisqu'alors $|x_{t,k} - x_{t,l}|$ est petit pour les configurations vraisemblables.

2.3. *Lois a priori et a posteriori*

En considérant les potentiels définis par les équations (2) et (3) et dans le cas d'un modèle à K bandes avec N états par bandes, la loi *a priori* de X est une distribution de Gibbs d'énergie totale

$$U(x) = \sum_{k=1}^{K} \sum_{i,j=1}^{N} a^{(k)}_{ij} \ \varphi^{(k)}_{ij}(x) + \sum_{k,l>k}^{K} f_{kl} \ \psi_{kl}(x) \ . \quad (4)$$

La fonction $\varphi^{(k)}_{ij}(x)$ compte le nombre de transitions de l'état i vers l'état j dans la bande k tandis que la fonction $\psi_{kl}(x)$ est l'écart cumulé entre les bandes k et l donné par

$$\psi_{kl}(x) = \sum_t |x_{t,k} - x_{t,l}| \ .$$

En associant à chaque état i dans chaque bande k une densité gaussienne, notée $g(\ .\ ; \mu^{(k)}_i, \sigma^{(k)}_i)$, et sous l'hypothèse classique d'indépendance conditionnelle des données, la loi d'une observation $Y = y$ conditionnellement à $X = x$ est une distribution de Gibbs d'énergie

$$U(y|x) = - \sum_{t,k} \sum_{i} \left(\ln(g(y_{t,k}; \mu^{(k)}_i, \sigma^{(k)}_i)) \ \delta(x_{t,k} = i) \right) \ .$$

On montre également que la distribution *a posteriori* de X connaissant $Y = y$ est donnée par une distribution de Gibbs d'énergie $U(x|y) = U(x) + U(y|x)$. Il est donc possible d'appliquer sur X les algorithmes classiques des champs de Markov pour les lois *a priori* et *a posteriori*.

3. Estimation des paramètres et algorithmes de décodage

3.1. *Estimation des paramètres*

Dans le cas de K sous-bandes, le modèle de champ de Markov est défini par l'ensemble des K matrices $(N \times N)$ de poids de transitions, $A^{(k)}$, la matrice $(K \times K)$ des poids de synchronisation, F, et les moyennes $\mu^{(k)}_i$ et variances $\sigma^{(k)}_i$ des gaussiennes. Nous proposons d'estimer simultanément l'ensemble de ces paramètres, noté θ, selon un critère du maximum de vraisemblance, en utilisant une procédure EM couplée à un algorithme de gradient pour l'étape de maximisation [Lan93].

Dans le cas d'une unique observation d'apprentissage, la fonction auxiliaire de l'algorithme EM est donnée par

$$\begin{aligned} Q(\theta, \theta^{(n)}) \quad = \quad & - \sum_k \sum_{i,j} a^{(k)}_{ij} E_{\theta^{(n)}}[\varphi^{(k)}_{ij}(x)|y] - \\ & \sum_{k,l>k} f_{kl} E_{\theta^{(n)}}[\psi_{kl}(x)|y] - \ln Z_\theta - \\ & \sum_{t,k} \gamma_{t,k}(i) \ \ln g(y_{t,k}; \mu^{(k)}_i, \sigma^{(k)}_i) \quad (5) \end{aligned}$$

où Z_θ est la fonction de partition associée à (4), $\theta^{(n)}$ l'estimation courante des paramètres et $\gamma_{t,k}(i) = P_{\theta^{(n)}}[X_{t,k} = i | Y = y]$. En dérivant (5) par rapport à $a^{(k)}_{ij}$, on obtient l'équation de maximisation suivante

$$E_{\theta^{(n+1)}}[\varphi^{(k)}_{ij}(x)] - E_{\theta^{(n)}}[\varphi^{(k)}_{ij}(x)|y] = 0 \qquad (6)$$

pour le poids de transition $a_{ij}^{(k)}$. L'équation (6) n'admettant pas de solution analytique, nous proposons de maximiser $Q(\theta, \theta^{(n)})$ en appliquant un pas de gradient pour obtenir une nouvelle estimation $\theta^{(n+1)}$ du paramètre, ce qui donne alors

$$a_{ij}^{(k)} \leftarrow a_{ij}^{(k)} + \frac{E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)] - E_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)|y]}{V_{\theta^{(n)}}[\varphi_{ij}^{(k)}(x)]} \quad . \quad (7)$$

Les espérances mises en jeu dans cette équation n'étant pas explicitement calculable, on les approxime à partir d'échantillons du champ X tirés suivants les lois *a priori* et *a posteriori*. Enfin, les paramètres $a_{ij}^{(k)}$ n'étant pas indépendants, cette méthode est appliquée en pratique à un vecteur regroupant les paramètres dépendants [GSC98a], la dénominateur dans (7) étant alors une matrice de covariance.

La réestimation des poids de synchronisation suit le même schéma que précédemment, ces paramètres étant supposés indépendants. Les formules de réestimation des paramètres des gaussiennes sont les mêmes que dans l'algorithme de Baum-Welch, les probabilités d'occupation des états, $\gamma_{t,k}(i)$, étant estimées à partir d'échantillons de X selon la loi *a posteriori*.

Les paramètres initiaux $\theta^{(0)}$, à l'exception des poids de synchronisation qui ne sont pas initialisés, sont obtenus à l'aide d'une stratégie itérative. Cette stratégie se base sur des estimateurs empiriques appliqués aux données complètes (x^*, y), x^* étant la configuration la plus vraisemblable *a posteriori*. Nous avons pu montré par des simulations que la procédure d'estimation des paramètres proposée donne de bons estimateurs.

3.2. Stratégie de décodage

La reconnaissance de mots isolés se base sur le calcul du score $p_w(y)$ pour chaque mot w du vocabulaire. Comme dans le cas des MMC, il est nécessaire de recourir à des approximations dans le calcul de ce score. Nous approximons $p_w(y)$ par le score sur les données complètes donné par

$$p_w(y) = pl_w(x^*) p_w(y|x^*) \quad . \quad (8)$$

Dans cette équation, la fonction $pl_w(x^*)$ correspond à la pseudo-vraisemblance de la configuration x^* et remplace la vraisemblance qui n'est pas calculable [Cha89]. La configuration x^* correspond à une estimation au sens du maximum *a posteriori* de X et peut-être déterminée à l'aide de l'algorithme ICM [Cha89] ou par recuit simulé. La différence entre ces deux algorithmes réside dans le fait que l'algorithme ICM converge rapidement mais n'est pas optimal, tandis que le recuit simulé converge vers un optimum global mais de manière plus lente.

4. Reconnaissance de mots isolés

4.1. Protocole expérimental

Le modèle par champ Markovien proposé est étudié en reconnaissance mono-locuteur de mots isolés sur de la parole téléphonique. Le vocabulaire est constitué de 10 mots courants. Un corpus contenant 100 occurrences de chaque mot du vocabulaire, prononcées par un même locuteur, a été extrait de la base de données PolyVar pour réaliser les expériences. Les cinquante premières occurrences de chaque mot sont utilisées pour l'estimation des paramètres des modèles, les 50 dernières étant réservées pour les tests.

Le signal de parole est divisé en K sous-bandes régulièrement réparties sur une échelle MEL, une représentation cepstrale du signal étant adoptée dans chaque bande. Le nombre N d'états par bande dépend du nombre de phonèmes composant le mot.

4.2. Résultats

Nous présentons dans la table 1 les résultats obtenus pour différents algorithmes d'estimation des paramètres et de décodages. Nous considérons une décomposition de la bande passante en 1, 3, 5 et 7 bandes représentées par, respectivement, 12, 5, 3 et 2 coefficients cepstraux. L'estimation ICM correspond à une simple initialisation des paramètres, x^* étant déterminé par l'algorithme ICM. Dans l'estimation ICM-EM, l'initialisation est suivie de 10 itérations de l'algorithme EM proposé. Dans les deux cas, le décodage peut-être basé sur l'algorithme ICM ou bien sur le recuit simulé (RS) pour la calcul de x^* dans (8). Enfin, pour comparaison, la première ligne du tableau correspond à une estimation par l'algorithme de Baum-Welch des paramètres des MMC de manière indépendante dans chaque bande. Dans ce dernier cas, pour le décodage V-ICM, la meilleure configuration x^* est déterminée par l'algorithme ICM initialisé par une segmentation obtenue en appliquant Viterbi dans chaque bande [GSC98b]. Le score est ensuite calculé comme précédemment à l'aide de l'équation (8).

Table 1: Taux de reconnaissance (en %) en fonction du nombre de bandes pour différentes techniques d'estimation des paramètres et de décodage.

estimation	scoring	b1c12	b3c5	b5c3	b7c2
BW	V-ICM	99.8	99.4	97.6	95.0
ICM	ICM	87.2	84.6	78.8	78.2
ICM-EM	ICM	88.6	80.6	75.4	76.0
ICM	RS	99.6	97.8	92.6	88.2
ICM-EM	RS	99.0	97.8	95.0	94.2

Les résultats mettent en évidence le problème du décodage basé sur l'algorithme ICM, ce dernier étant trop sensible aux conditions initiales. En dépit des différences dans les procédures d'estimation des paramètres entre les trois premières lignes du tableau, l'écart de performance entre le décodage V-ICM et ICM montre clairement les défauts de l'algorithme ICM. Dans le cas mono-bande (b1c12), le modèle de champ de Markov proposé donne des résultats similaires à ceux obtenus avec les MMC, pour un décodage à base de recuit simulé. La comparaison des différentes divisions en sous-bandes du signal montre que dans tous les cas, le taux de reconnaissance est inversement proportionnel au nombre de bandes. Ce résultat pourrait être expliqué par le fait que l'on

a une moins bonne représentation de la parole, le nombre de coefficients cepstraux utilisés dans chaque bande devenant de plus en plus faible. Cependant, lorsque l'on prend 5 coefficients cepstraux par bandes dans un modèle à 7 bandes, le taux de reconnaissance n'augmente que de manière marginal, passant de 95% à 96.4% dans l'approche par MMC indépendants. Dans le décodage par recuit simulé, on peut voir que l'estimation des poids de synchronisation par l'algorithme EM permet de limiter la baisse de performance. La modélisation de la synchronisation inter-bandes a donc une influence importante sur les performances du modèle. Ce résultat souligne la nécessité de disposer d'un bon modèle *a priori* du processus caché X lorsque l'observation devient plus variable. En effet, la variabilité de l'observation dans une bande croît avec le nombre de bandes. Dans ce cas, une meilleure modélisation du processus X permet une meilleure régularisation de la segmentation, et donc un meilleur taux de reconnaissance.

Pour étudier le comportement de l'approche proposée en présence de bruit additif, nous avons artificiellement ajouté un bruit aux données de test. Le bruit ajouté est coloré par un filtre de réponse impulsionelle $H(z) = 1/1 - 0.9z^{-1}$, concentrant ainsi l'énergie du bruit dans les basses fréquences. Les trois premières lignes du tableau 2 montrent les résultats obtenus pour 3 bandes, en appliquant une reconnaissance par MMC dans chacune des bandes pour différents rapports S/B. La quatrième ligne donne le taux de reconnaissance obtenu si l'on fusionne les scores obtenus avec les MMC dans chaque bande en faisant une moyenne. Enfin, la dernière ligne montre les résultats obtenus avec une approche par champs de Markov. Ces résultats montrent que dans ce cas, la fusion des scores par moyenne donne de mauvais résultats lorsque le rapport S/B augmente, les deux premières bandes étant fortement dégradées comme le montrent les taux de reconnaissance par bande. Le modèle de champ de Markov proposé donne de meilleurs résultats que la fusion par moyenne mais les performances obtenues sur la seule bande peu bruitée (bande #3) restent meilleures. Enfin, lorsque la synchronie entre les bandes n'est pas modélisée (*i.e.* $f_{kl} = 0$), le taux de reconnaissance à 30 dB n'est plus que de 44.8 % au lieu de 49.4 %, ce qui montre l'intérêt de la synchronie en présence de bruit.

Table 2: Taux de reconnaissance (en %) en fonction du rapport S/B dans une approche 3 bandes.

S/B (dB)	∞	30	20	10
bande #1	96.4	9.0	9.8	10.0
bande #2	94.4	17.4	17.4	14.2
bande #3	92.2	80.4	59.8	28.4
moyenne	99.6	28.0	13.2	10.4
ICM-GEM/RS	93.6	49.4	37.8	24.8

5. Conclusion

Dans cet article, nous avons proposé un modèle multibande dans lequel la synchronie entre les bandes est modélisée. Ce modèle, qui repose sur la théorie des champs de Markov, est étudié pour la reconnaissance de mots isolés. La méthode proposé donne des performances comparables aux MMC dans le cas monobande. Quelque soit l'approche choisie, les performances des systèmes baissent dans le cas de la parole non bruitée lorsque le nombre de bandes augmentent. Cependant, l'introduction d'un terme de synchronisation entre les bandes permet de limiter la baisse des performances. Enfin, en présence de bruit additif, le modèle proposé permet d'améliorer les performances par rapport à une approche par MMC avec une fusion par moyenne des scores partiels de chaque sous-bande.

Cependant, le modèle de synchronisation introduit dans notre approche est stationnaire dans le sens où les poids de synchronisation inter-bandes sont constants pour un segment donné. Bien que pratique, cette hypothèse de stationnarité de la synchronisation est fausse et un modèle plus complet devrait être étudié. Soulignons pour conclure que l'intérêt d'une modélisation par champs de Markov réside dans la souplesse de cette approche, de nombreux types d'interactions et de nombreuses familles de potentiels pouvant être envisagés.

Bibliographie

[Bes74] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Royal Statistical Soc.*, B-48:192–236, 1974.

[Cha89] Bernard Chalmond. An iterative Gibbsian technique for reconstruction of m-ary images. *Pattern Recognition*, 22(6):747–761, 1989.

[Geo88] Hans-Otto Georgii. *Gibbs measures and phase transitions*, volume 9 of *Studies in Mathematics*. de Gruyter, 1988.

[GSC98a] G. Gravier, M. Sigelle, and G. Chollet. Markov random field modeling for speech recognition. *Australian J. for Intelligent Info. Proc. Systems*, 5(4), 1998.

[GSC98b] G. Gravier, M. Sigelle, and G. Chollet. Toward Markov random field modeling of speech. In *ICSLP*, December 1998.

[HPT96] H. Hermansky, M. Pavel, and S. Tibrewala. Towards ASR using partially corrupted speech. In *ICSLP*, Oct. 1996.

[Jel98] Frederick Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, 1998.

[Lan93] Kenneth Lange. A gradient algorithm locally equivalent to the EM algorithm. *J. Royal Statistical Soc.*, 57(2):425–437, 1993.

[TRM+97] M. J. Tomlinson, M. J. Russel, R. K. Moore, A. P. Buckland, and M. A. Fawley. Modelling asynchrony in speech using elementary single-signal decomposition. In *ICASSP*, volume 2, pages 1247–1250, 1997.

[ZAZ91] Yunxin Zhao, Lee A. Atlas, and Xinhua Zhuang. Application of the Gibbs ditribution to hidden Markov modeling in speaker independent isolated word recognition. *IEEE Trans. on Signal Processing*, 39(6):1291–1298, 1991.