Réseaux Bayésiens Dynamiques pour la Reconnaissance Multi-Bandes de la Parole

Khalid Daoudi, Dominique Fohr et Christophe Antoine

LORIA/INRIA, Speech Group.
615 rue du jardin botanique 54602 Villers-lès-Nancy FRANCE
Tél.: ++33 (0)3 83 59 30 64 - Fax: ++33 (0)3 83 27 83 19
Mél: daoudi,fohr,antoinec@loria.fr - http://www.loria.fr/equipes/parole/

ABSTRACT

This paper presents a new approach to multi-band automatic speech recognition which has the advantage to overcome many limitations of classical muti-band systems. The principle of this new approach is to build a speech model in the time-frequency domain using the formalism of Bayesian networks. Contrarily to classical multi-band modeling, this formalism leads to a probabilistic speech model which allows communications between the different sub-bands and, consequently, no recombination step is required in recognition. We develop efficient learning and decoding algorithms and present illustrative experiments on a connected digit recognition task. The experiments show that the Bayesian network's approach is very promising in the field of noisy speech recognition.

1. INTRODUCTION

Les systèmes de reconnaissance automatique de la parole actuels sont fondés sur des modèles probabilistes utilisant les modèles de Markov cachés (en anglais Hidden Markov Models ou HMM). Ces modèles obtiennent de bonnes performances en condition de laboratoire (locuteur coopératif, sans accent marqué, tâche simple...). Cependant, en conditions réelles (bruits de fond, parole spontanée, locuteur non natif...) les performances des systèmes HMM se dégradent fortement. Une des raison de cette dégradation tient au fait que la modélisation dans un système HMM ne parvient pas à capturer des phénomènes acoustiques qui sont spécifiques à la parole. Alors que la dynamique temporelle est bien capturée par les HMMs, la dynamique fréquentielle est mal modélisée par ces derniers.

Récemment, une nouvelle approche pour la reconnaissance automatique de la parole, nommée "multi-bandes" a été proposée dans [5]. Cette approche s'inspire d'une étude de Harvey Fletcher [11] sur la perception de la parole. En résumé, cette étude, (reprise par Jont B. Allen in [1]), suggère que le système auditif traite la parole *localement* dans le domaine temps-fréquence avant de faire la reconnaissance proprement dite. L'approche multi-bandes peut se schématiser en trois étapes: dans un premier temps on divise le signal de parole en plusieurs bandes de fréquences appelées sous-bandes. Dans une deuxième étape on effectue de façon indépendante une reconnaissance dans chacune de ces sous-bandes par un HMM. Finalement les différents résultats obtenus dans chacune des sous-bandes sont combinés pour obtenir le résultat final. Une autre motivation de cette méthode est d'être plus résistant aux bruits limités à une partie du spectre de fréquence. En effet, dans un système classique qui extrait les paramètres acoustiques sur la totalité du spectre du fréquence, tous les coefficients seront perturbés même si le bruit n'occupe qu'une faible partie du spectre. Dans un système multi-bandes, seules les informations issues de la sous-bande bruitée seront dégradés et celles issus des autres sous-bandes peuvent être exploitées pour une meilleure reconnaissance.

L'approche multi-bandes est très séduisante car elle s'inspire du fonctionnement du système auditif humain et peut conduire à des systèmes plus robustes au bruit ambiant. Cependant, l'approche décrite ci dessus, est loin d'être optimale. En effet, les sous-bandes de fréquence sont supposées indépendantes, ce qui semble bien peu réaliste. De plus, l'étape de recombinaison est une tâche particulièrement difficile dans la cas de la reconnaissance de parole continue.

Le but de cet article est de proposer une nouvelle approche pour la reconnaissance multi-bandes qui présente l'avantage de palier les inconvénients que nous venons de mentionner. Dans notre nouvelle approche, nous modélisons les dépendances entre les sous-bandes en créant des "interactions" entre les différents HMM correspondants aux sous-bandes. Pour cela, nous utilisons le formalisme des réseaux Bayésiens (RB) qui constitue un cadre idéal pour deux raisons majeurs. D'une part, grâce à leur structure graphique, les RB offrent un outil naturel pour représenter les dépendances entre les différentes variables d'un système donné. D'autre part, en exploitant les indépendances conditionnelles entre les variables, ils introduisent une certaine "modularité" dans les systèmes complexes. Ainsi, non seulement les RB fournissent un outil séduisant pour modéliser des systèmes complexes, mais aussi conduisent à des algorithmes rapides d'inférence et d'apprentissage.

À la suite des travaux pionniers de Judea Pearl [14], les RB ont émergé comme un formalisme puissant qui unifie différents concepts de modélisation probabilistes utilisés en statistique, intelligence artificielle, traitement du signal... Par exemple, les HMMs et les filtres de Kalman sont des cas particuliers du formalisme des RB. Ainsi, les RB sont devenus courament employés pour le raisonnement sous incertitude et sont très utilisés dans la conception de systèmes experts et d'aide à la décision. Cependant, l'utilisation de RB pour la reconnaissance automatique de la parole a attiré l'attention seulement très récemment [2, 3, 4, 7, 8, 10, 15, 16].

Ce papier présente un nouveau système multi-bandes

où nous construisons un RB dans le domaine temps-fréquence en couplant les HMMs associés aux différentes sous-bandes. Nous développons des algorithmes d'apprentissage et de décodage pour ce nouveau modèle et nous conduisons des expériences pour illustrer le potentiel de ce nouveau modèle.

2. RÉSEAUX BAYÉSIENS

Durant les dix dernières années, les réseaux Bayésiens (et les modèles graphiques en général) sont devenus très populaires en intelligence artificielle grâce à de nombreuses avancées dans différents aspects de l'apprentissage et de l'inférence. La littérature est maintenant riches en livres et articles traitant de la théorie et de l'application des réseaux Bayésiens, le lecteur intéressé pourra trouver une très bonne introduction dans [6]. Formellement, un RB (statique) est défini par la connaissance de deux éléments: un graphe acyclique orienté S et une paramétrisation numérique Θ. Étant donné un ensemble de variables aléatoires $X = \{X_1, ..., X_N\}$ et $P(X)$ sa distribution jointe de probabilité (DJP), le graphe S code les indépendances conditionnelles qui existent dans le DJP. La paramétrisation Θ est donnée en terme de probabilités conditionnelles des variables connaissant leurs parents. Une fois S et Θ donnés, le DJP peut être exprimée sous la forme :[1]

$$P(x) = \prod_{i=1}^{N} P(x_i|pa(x_i)) \quad (1)$$

où $pa(x_i)$ désigne une réalisation des parents de X_i. La sémantique des indépendances conditionnelles (ou propriétés de Markov) d'un RB implique que, sachant ses parents, une variable est indépendante de toutes les autres variables du réseau à l'exception des ses descendants.

Les réseaux Bayésiens dynamiques (RBD) sont une extension des réseaux Bayésiens qui permet de représenter l'évolution temporelle des variables. Si on considère un ensemble $X[t] = \{X_1[t], \ldots, X_N[t]\}$ de variables évoluant dans le temps, un RBD représente la distribution de probabilité jointe de ces variables pour un intervalle borné $[0, T]$. En général, cette DJP peut être codée par un gros réseau Bayésien statique avec $T \times N$ variables avec la possibilité d'avoir une structure et/ou des paramètres différents pour chaque temps. Si le processus est stationnaire, les hypothèses d'indépendance et les probabilités conditionnelles associées sont identiques pour tous les temps t. Dans ce cas, le RBD peut être représenté par un RB dont la structure est dupliquée pour chaque pas de temps t. De ce point de vue, il est évident qu'un HMM est un cas particulier de RBD comme le montre la figure 2. contrairement à la représentation habituelle des HMM, un noeud H_t (resp. O_t) est une variable aléatoire dont la valeur indique l'état occupé (resp. l'observation) au temps t. Le temps apparaît donc de façon explicite et les flèches qui lient les H_t indiquent les dépendances (et non pas les transitions entre états). C'est cette représentation que nous utiliseront dans la suite de cet article.

[1] dans cet article, les lettres majuscules sont utilisées pour les variable aléatoires et les minuscules pour leurs réalisations.

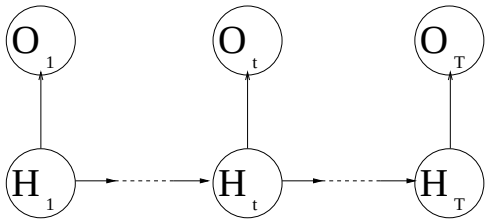

Figure 1: un HMM représenté comme un réseau Bayésien

3. RECONNAISSANCE MULTI-BANDES PAR RBD

Étant donné un vocabulaire V composé de $|V|$ mots. L'idée principale de notre approche est la suivante: pour chaque mot v de V, au lieu de considérer un HMM indépendant dans chaque sous-bande, nous construisons un RBD plus complexe (mais uniforme) dans le domaine temps-fréquence en "couplant" les différents HMMs associées aux différentes sous-bandes. Ce couplage est réalisé en ajoutant des liens (orientés) entre les variables pour capturer les dépendances entre les sous-bandes. Une question naturelle est: quels sont les liens à ajouter?. Probablement, la meilleure solution serait d'apprendre la structure graphique (les dépendances) à partir des données. Cependant, cette stratégie, appelée *apprentissage structurelle*, qui est extrêmement intéressante et que nous sommes en train d'étudier [10] n'est pas l'objectif de ce papier. Notre but ici est (d'abord) de fixer une structure graphique "raisonnable" puis évaluer si notre nouvelle approche multi-bandes est prometteuse.

3.1. *Définition du modèle (ou de la structure du graphe)*

Nous fixons cette structure "raisonnable" en nous basant sur les critères suivants:

- le modèle doit avoir un nombre relativement faible de paramètres pour que la complexité des calculs reste abordable,
- aucune variable continue ne doit avoir de fils discrets afin de pouvoir appliquer un algorithme d'inférence exacte,
- des liens doivent exister entre les variables cachées le long de l'axe des fréquences afin de capturer l'asynchronisme potentiel entre elles.

La figure 2 présente un modèle qui satisfait ces critères. Dans ce RB les variable cachées de la sous-bande n sont liées à celles de la sous-bande $n+1$ de telle façon que l'état d'une variable cachée dans la sous-bande $n+1$ au temps t dépend de l'état de deux autres variables cachées: au temps $t-1$ dans la même sous-bande et au temps t dans la sous-bande n. $H_t^{(n)}\left(= H_t^{(n)}(v)\right)$ est une variable discrète prenant ses valeurs dans un ensemble ordonnés d'étiquettes $I_v = \{1_v, ..., m_v\}$, $|I_v|$ étant le nombre d'états cachés. $O_t^{(n)}\left(= O_t^{(n)}(v)\right)$ est une variable continue avec un mélange de Gaussiennes comme distribution (sachant la valeur de la variable cachée correspondante $H_t^{(n)}$) représentant l'observation au temps t dans la sous-bande n $(n = 1, ..., B)$, B est le nombre de sous-bandes. Nous imposons une topologie gauche-droite dans chaque

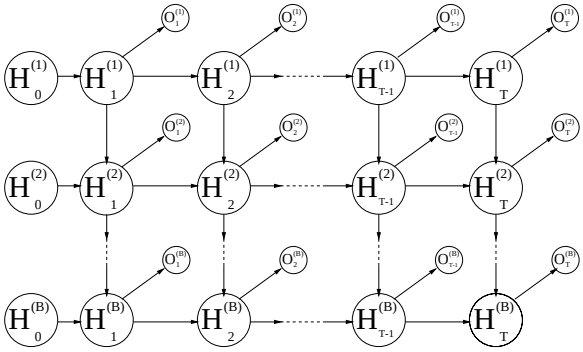

Figure 2: Réseau Bayésien dynamique B-bandes

sous-bande et nous faisons l'hypothèse que les paramètres sont stationnaires. Ainsi, pour un mot $v \in V$ (et pour tout $(i,j,k) \in I_v^3$) , la paramétrisation numérique Θ_v de son RBD B-bandes est:

$$\begin{cases} a_{ij}(v) \triangleq P(H_t^{(1)} = j | H_{t-1}^{(1)} = i) \\ u_{ijk}^{(n)}(v) \triangleq P(H_t^{(n)} = k | H_t^{(n-1)} = i, H_{t-1}^{(n)} = j) \\ b_i^{(n)}(v,\cdot) \triangleq P(O_t^{(n)} = \cdot | H_t^{(n)} = i) \end{cases} \quad (2)$$

où $b_i^{(n)}(v,\cdot)$ est un mélange de lois Gaussiennes. L'asynchronie entre les sous-bandes est prise en compte en autorisant les $u_{ijk}^{(n)}(v)$ à être différents de 0 (sauf quand $k < j$ ou $k > j+1$ pour respecter la topologie gauche-droite).

Contrairement à un HMM classique, notre modèle multi-bandes fournit une modélisation de la dynamique fréquentielle de la parole. A la différence d'un modèle multi-bandes classique, notre RBD permet une interaction entre les sous-bandes et une asynchronie entre elles. De plus, notre modèle utilise les informations contenue dans toutes les sous-bandes et ainsi aucun module de recombinaison n'est nécessaire. Dans le même domaine, un système multi-bandes fondé sur les champs de Markov a été proposé dans [12]. Mais cette approche, contrairement à la notre, ne permet ni une inférence exacte ni rapide et impose un modèle linéaire pour l'asynchronie entre les sous-bandes. Dans notre approche, l'asynchronie est apprise à partir des données.

3.2. Estimation des paramètres du modèle

Dans les expériences que nous avons effectuées, l'apprentissage des modèles est effectué de façon indépendante. En utilisant l'algorithme EM (Expectation-Maximisation), estimer les paramètres du modèle revient à une simple généralisation de l'algorithme de Baum-Welch. Cela est rendu possible grâce au fait que nous avons imposé que les variables continues ne soient conditionnées que par des variable discrètes. L'étape d'"Expectation" est réalisée en utilisant l'algorithme JLO [13]. Les formules de re-estimation des paramètres peuvent être trouvées dans [8].

4. APPLICATION À LA RECONNAISSANCE DE PAROLE CONTINUE

Dans une application de reconnaissance de parole continue, étant donné un un RBD B-bandes pour chaque mot du vocabulaire et une phrase prononcée par un locuteur, le problème est de trouver la séquence de mots la plus probable. Une solution naïve serait d'utiliser un algorithme de Viterbi B-dimensionnel qui est très coûteux en temps de calcul. Dans cette section, nous présentons un algorithme de décodage efficace qui repose essentiellement sur la construction d'un RBD B-bandes "augmenté". Ensuite, nous décrivons des résultats sur une tache de reconnaissance de chiffres enchaînés.

4.1. Algorithme de décodage

L'idée de base est de construire un nouveau RBD B-bandes qui représente tous les mots du vocabulaire, le décodage est ensuite fait en inférant ce nouveau RBD. Précisément, la structure graphique de ce nouveau RBD B-bandes est la même que celle de la figure 2, la seule différence est que maintenant les variables ne dépendent plus du mot sous considération, mais plutôt chaque variable $H_t^{(n)}$ prend ses valeurs maintenant dans l'ensemble $I = \bigcup_{v \in V} I_v$. Pour compléter la définition de ce nouveau RBD, nous devons spécifier sa paramétrisation numérique. Soit $(i,j,k) \in I^3$ tel que $(i,j,k) \in {I_v}^3$ pour un certain $v \in V$. Les probabilités conditionnelles des observations sont simplement données par celles correspondant à chaque mot :

$$P(O_t^{(n)} = \cdot | H_t^{(n)} = i) \triangleq b_i^{(n)}(v,\cdot).$$

Pour spécifier la paramétrisation du processus caché, nous devons introduire le modèle de langage, nous faisant aussi des hypothèses de (a)synchronie: nous continuons à autoriser une complète asynchronie à *l'intérieur* d'un mot, mais nous imposons une complète synchronie lors de la *transition* entre deux mot. Précisément, puisqu'on a une topologie gauche-droite, les seules probabilités conditionnelles non nulles sont les suivantes:

• La transition synchrone entre deux mots v et v' (pas nécessairement différents) :

$$P(H_t^{(1)} = 1_v | H_{t-1}^{(1)} = m_{v'}) \triangleq P(v|v')$$

$$P(H_t^{(n)} = 1_v | H_t^{(n-1)} = 1_v, H_{t-1}^{(n)} = m_{v'}) \triangleq P(v|v')$$

où $P(v|v')$ est donnée par le modèle de langage.

• Les probabilités conditionnelles à l'intérieur d'un mot:

$$P(H_t^{(1)} = j | H_{t-1}^{(1)} = i) \triangleq a_{ij}(v)$$

$$P(H_t^{(n)} = k | H_t^{(n-1)} = i, H_{t-1}^{(n)} = j) \triangleq u_{ijk}^{(n)}(v).$$

Nous avons ainsi un RBD B-bandes complètement défini sur lequel nous pouvons opérer le décodage. Pour ce faire, nous utilisons l'algorithme de Dawid [9] qui permet d'identifier (avec la même complexité que l'algorithme JLO [13]) la séquence la plus probable des états cachés sachant les observations.

Étant donné les hypothèses de (a)synchronie et la topologie gauche-droite, la complexité totale de notre algorithme de décodage est $O(Bm^B T + |V|^2 T)$. Nous signalons aussi que dans le cas 1-bande (c.à.d. un HMM), cet algorithme est exactement équivalent à une décodage de Viterbi.

4.2. Résultats expérimentaux

Les expérimentations sont effectuées sur le corpus de chiffres connectés TIDIGITS dans laquelle 112 locuteurs

Noise 3-6 KHz	HMM1	HMM2	HMM4	RBD
SNR 26 db	89.95%	92.69%	97.20%	**96.16**%
SNR 20 db	82.17%	85.17%	94.19%	**94.89**%
SNR 14 db	73.27%	75.33%	87.44%	**90.81**%
SNR 8 db	62.57%	59.57%	73.85%	**82.27**%
SNR 2 db	58.86%	40.82%	54.60%	**75.51**%

Table 1: Taux de reconnaissance par mot (HMMn signifie un HMM avec un mélange de n-Gaussiennes par état)

Noise 3-6 KHz	HMM1	HMM2	HMM4	RBD
SNR 26 db	71.47%	79.05%	92.00%	**89.42**%
SNR 20 db	52.49%	59.38%	84.09%	**85.90**%
SNR 14 db	35.69%	40.13%	69.22%	**74.67**%
SNR 8 db	20.90%	20.55%	46.00%	**53.86**%
SNR 2 db	10.97%	9.696%	23.82%	**39.87**%

Table 2: Taux de reconnaissance par phrase (HMMn signifie un HMM avec un mélange de n-Gaussiennes par état)

sont utilisés pour l'apprentissage et 113 pour le test. Chaque locuteur a énoncé 77 séquences, qui chacune contient entre 1 et 7 chiffres. Nous comparons les performances d'un RBD 2-bandes avec une Gaussienne par état avec un modèle HMM classique à une ou plusieurs Gaussiennes. Les onze chiffres (de 0 à 9 plus le "oh") et le silence ont le même nombre d'états ($m = 7$) et toutes le matrice de covariances sont diagonales. le modèle de langage est uniforme, c'est à dire $P(v|v') = \frac{1}{12}$.

La paramétrisation du modèle HMM classique (une seule bande) est la suivante: le signal est passé à travers 24 filtres MEL et 12 paramètres MFCC sont extraits. Le vecteur de paramètres contient 35 composantes : 11 coefficients statiques (C0 à été retiré), 12 Δ et 12 $\Delta\Delta$. Pour le modèle RBD 2-bandes, les paramètres sont extraits à partir des 16 premiers filtres pour la première bande et des 8 derniers filtres pour la deuxième bande ce qui correspond à une bande passante de $[0..2152Hz]$ pour la bande 1 et de $[1777Hz..10000Hz]$ pour la bande 2. Chaque vecteur de paramètres contient 17 coefficients: 5 MFCC statiques, 6 Δ et 6 $\Delta\Delta$. L'apprentissage des modèles est effectué uniquement à partir des fichiers originaux sans bruit. Pour les fichiers de test, un bruit blanc filtré entre 3000 et 6000Hz a été ajouté aux signaux de parole de TIDIGITS.

Les Tableaux 1 et 2 montre les résultats de reconnaissance au niveau du mot et de la phrase complète, respectivement. Les performances de notre modèle RBD 2-bandes à une Gaussienne sont nettement supérieures à celles obtenues avec un modèle HMM à une gaussienne. On pourrait penser que c'est dû au fait que notre modèle possède (légèrement) plus de paramètres que le HMM. Mais cet argument ne tient pas si on regarde les résultats obtenus lorsqu'on augmente le nombre de Gaussiennes (et donc le nombre de paramètres) du HMM. En effet, tous les modèles HMM avec plus de 2 Gaussiennes par état ont plus de paramètres que le RBD 2-bandes. En particulier, plus le SNR est faible plus la performance de notre modèle est supérieure à celle des HMMs. Cela illustre le potentiel de notre approche à exploiter l'information contenue dans la bande de fréquence qui n'est pas perturbée par le bruit et montre que l'approche RBD à la reconnaissance multi-bandes de la parole est très prometteuse.

BIBLIOGRAPHIE

[1] J. Allen. How do humans process and recognize speech. *IEEE Trans. Speech and Audio Processing*, 2(4):567–576, 1994.

[2] Jeff A. Bilmes. Data-driven extensions to hmm statistical dependencies. In *International Conference on Spoken Language Processing (ICSLP)*, 1998.

[3] Jeff A. Bilmes. Buried markov models for speech recognition. In *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 1999.

[4] Jeff A. Bilmes. *Natural Statistical Models for Automatic Speech Recognition*. PhD thesis, International Compute Science Institute, Berkeley, California, 1999.

[5] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. International Conference on Spoken Language Processing (ICSLP), 1996.

[6] R. Cowell, A. Dawid, S. Lauritzen, and D. Spiegelhalter. Probabilistic Networks and Expert Systems. *Springer-Verlag*, 1999.

[7] K. Daoudi, D. Fohr, and C. Antoine. A new approach for multi-band speech recognition based on probabilistic graphical models. In *International Conference on Spoken Language Processing (ICSLP), Beijing, China*, October 2000.

[8] K. Daoudi, D. Fohr, and C. Antoine. Continuous Multi-Band Speech Recognition using Bayesian Networks. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Terento, Italy*, December 2001.

[9] A. Dawid. Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, (2):25–36, 1992.

[10] M. Deviren and K. Daoudi. Structural Learning of Dynamic Bayesian Networks in Speech Recognition. In *EUROSPEECH, Alborg, Denemark*, September 2001.

[11] H. Fletcher. *Speech and hearing in communication*. Krieger, New-York, 1953.

[12] G. Gravier. *Analyse statistique à deux dimensions pour la modélisation segmentale du signal de parole: Application à la reconnaissance*. PhD thesis, ENST Paris, 2000.

[13] F. Jensen, S. Lauritzen, and K. Olsen. Bayesian updating in recursive graphical models by local computations. *Computational Statistics and Data Analysis*, (4):269–282, 1990.

[14] J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[15] G. G. Zweig and S. Russell. Speech recognition with dynamic bayesian networks. In *Proceedings Fifteenth National Conference on Artificial Intelligence*, Madison, Wisconsin, 1998.

[16] G.G. Zweig. *Speech Recognition with Dynamic Bayesian Networks*. PhD thesis, University of California, Berkeley, 1998.