Transformation linéaire discriminante pour l'apprentissage des HMM à analyse factorielle

*Fabrice Lefèvre*et Jean-Luc Gauvain*

ABSTRACT

Factor analysis has been recently used to model the covariance of the feature vector in speech recognition systems. Maximum likelihood estimation of the parameters of factor analyzed HMMs (FAHMMs) is usually done via the EM algorithm. The initial estimates of the model parameters is then a key issue for their correct training. In this paper we report on experiments showing some evidence that the use of a discriminative criterion to initialize the FAHMM maximum likelihood parameter estimation can be effective.

The proposed approach relies on the estimation of a discriminant linear transformation to provide initial values for the factor loading matrices. Solutions for the appropriate initializations of the other model parameters are presented as well. Speech recognition experiments were carried out on the *Wall Street Journal* LVCSR task with a 65k vocabulary. Contrastive results are reported with various model sizes using discriminant and non discriminant initialization.

1. INTRODUCTION

Ces dernières années, un renouveau d'intérêt a été observé pour la modélisation de la covariance dans les systèmes de reconnaissance de la parole à base de HMM [1, 3, 10, 8]. L'utilisation de matrices de covariance pleines, bien que souhaitable en principe, augmente considérablement en pratique le nombre de paramètres des modèles et complique l'estimation des paramètres. C'est pourquoi des covariances à matrices diagonales sont généralement utilisées dans les systèmes. L'analyse factorielle fournit une stratégie de modélisation intermédiaire qui permet d'obtenir des matrices de covariances pleines avec un moins grand nombre de paramètres. Un modèle générateur de la parole, basé sur un schéma de filtrage statistique du signal, est utilisé. Dans ce modèle, l'hypothèse est faîte que les observations de parole sont le résultat d'une transformation linéaire bruitée à partir d'un espace d'états de plus faible dimension. Les modèles à analyse factorielle ont été récemment généralisés dans le contexte des modèles markoviens [8]. Le HMM à analyse factorielle (ou FAHMM) est un modèle linéaire gaussien basé sur un processus d'évolution d'états constant par morceaux. Les vecteurs d'états sont générés par un HMM utilisant des mixtures de gaussiennes à covariance diagonale. Comme le montre [8], les FAHMM fournissent une cadre général intégrant beaucoup d'autres modèles standards de covariances comme l'analyse factorielle partagée (SFA) [10], l'analyse factorielle indépendante (IFA) [1] ou encore les modèles à covariances semi-liées (STC) [3]. Avec les FAHMM, différents niveaux de partage de paramètres peuvent être envisagés conduisant à différents degrés de complexité des composants statistiques.

L'utilisation optimale des FAHMM repose sur la configuration d'un nombre important de paramètres. Tout comme les HMM traditionnels, la taille des vecteurs d'observation ainsi que le nombre de gaussiennes par état doivent être fixés. De plus, dans le cas des FAHMM, la taille de l'espace d'état doit être fixé ainsi que le nombre de gaussiennes qui lui sont associées. Les autres paramètres du modèle sont appris en utilisant une procédure de type EM. La qualité des valeurs initiales est alors essentielle pour permettre une bonne convergence. L'approche évaluée dans notre travail tente d'améliorer ce point en introduisant un critère discriminant lors de la sélection des dimensions de l'espace d'état. Pour cela nous proposons d'obtenir les dimensions de l'espace d'état à partir d'une transformation linéaire discriminante héteroscedastique (*Heteroscedastic Linear Discriminant Analysis*, HLDA) [5].

Les expériences ont été menées sur une tâche de reconnaissance de la parole continue à grand vocabulaire, *Wall Street Journal* [7]. Une comparaison est faîte entre les HMM à matrices de covariance diagonales et pleines et les FAHMM appris avec les initialisations standard ou discriminante.

2. HMM À ANALYSE FACTORIELLE

Les FAHMM sont une généralisation à espace d'état dynamique des systèmes à analyse factorielle à composants multiples. Le modèle générateur utilisé dans les FAHMM pour chaque indice de temps t et chaque état j est décrit par les équations suivantes :

$$o_t = C_j x_t + v_t \tag{1}$$

$$x_t \sim \sum_k c_{jk}^{(x)} \mathcal{N}(x_t; \mu_{jk}^{(x)}, \Sigma_{jk}^{(x)})) \tag{2}$$

$$v_t \sim \sum_l c_{jl}^{(o)} \mathcal{N}(v_t; \mu_{jl}^{(o)}, \Sigma_{jl}^{(o)})) \tag{3}$$

avec o_t un vecteur d'observation de dimension n, x_t un vecteur d'état de dimension p et v_t un vecteur d'erreur d'observation de dimension n. Toute les matrices de covariance sont diagonales. La structure interne de la covariance est capturée par la matrice C_j, appelée la matrice

*F. Lefèvre est aussi avec l'équipe Dialogue Homme-Machine du LIA-Université d'Avignon

de chargement des facteurs. La distribution d'un vecteur d'observation o_t pour un état donné j, une composant de l'espace d'état k et une composante du bruit d'observation l, est obtenue par intégration sur le vecteur d'état x_t. La distribution qui en résulte est une gaussienne, dont le vecteur moyenne et la matrice de covariance sont donnés par :

$$\mu_{jkl} = C_j \mu_{jk}^{(x)} + \mu_{jl}^{(o)} \quad (4)$$

$$\Sigma_{jkl} = C_j \Sigma_{jk}^{(x)} C_j' + \Sigma_{jk}^{(o)}. \quad (5)$$

La densité de probabilité conditionnelle des observations sur les états des FAHMM peut être vue comme une mixture de densités gaussiennes avec $M^{(o)}M^{(x)}$ matrices de covariance pleines, dont les moyennes et les matrices de covariances sont données par les équations (4) et (5). La vraisemblance marginale d'une observation considérant l'état j est alors obtenue en sommant sur les deux ensembles de gaussiennes. Ce calcul nécessite d'inverser $M^{(o)}M^{(x)}$ matrices de covariance pleines de taille $n \times n$.

Une présentation détaillée des formules des reestimation EM ainsi que les conditions générales de l'apprentissage des FAHMM peut être trouvée dans [8]. Dans notre système, ces formules de réestimation ont été intégrées avec la nuance qu'une segmentation constante est utilisée pour l'apprentissage EM (*apprentissage de Viterbi*). Afin de réduire l'effort de calcul durant l'apprentissage des modèles, l'algorithme à deux niveau, a été adopté : une boucle d'itération interne plus rapide permet d'accélérer la convergence.

L'initialisation des paramètres des modèles est un problème important dans le cadre de l'algorithme EM dans la mesure où elle conditionne la possibilité d'atteindre une bonne solution après convergence. Pour les FAHMM, un point initial raisonnable consiste à convertir un HMM standard (avec des composants mono-gaussiens) en un FAHMM équivalent en utilisant les cepstres statiques comme dimensions de l'espace d'état. Cette proposition fait l'hypothèse que les vecteurs de l'espace d'état sont fortement corrélés avec les dimensions statiques des vecteurs d'observation cepstraux, ce qui n'est pas forcement très réaliste. Pour cette raison, nous proposons d'utiliser une projection discriminante de type HLDA pour l'initialisation des modèles.

3. INITIALISATION DISCRIMINANTE

Des travaux récents sur la modélisation acoustique [5, 3, 9] ont conduit à une adoption généralisée de la technique de transformation linéaire discriminante HLDA dans les systèmes de reconnaissance de l'état-de-l'art. L'objectif de la transformation discriminante des observations consiste à trouver un espace de projection de faible dimension mais conservant l'information discriminante. HLDA est une méthode basée sur le maximum de vraisemblance pour estimer une projection linéaire des vecteurs d'observation de dimension n vers un sous-espace de dimension p. Comme pour LDA, le sous-espace obtenu doit permettre une meilleure séparation des classes, chaque classe étant modélisée par une gaussienne. HLDA généralise LDA en éliminant le recours à une matrice unique comme matrice de covariance intra-classe, ce qui conduit à un meilleur espace de projection lorsque les classes sont héteroscedastiques.

En pratique, la transformation linéaire A est divisée en 2 sous-matrices : A_p transforme l'espace original vers un sous-espace de dimension p et A_{n-p} vers le sous-espace des dimensions rejetées. L'hypothèse faîte est que les moyennes et les variances des n-p dimensions rejetées sont représentées par les dimensions correspondantes des moyennes et variances transformées globales. L'optimisation de la fonction objective associée à HLDA peut être obtenue par des méthodes numériques (comme le gradient conjugué) ou en utilisant une procédure d'optimisation par maximum de vraisemblance réalisée ligne par ligne [3]. Dans cette dernière approche (utilisée dans notre expérience), chaque ligne de la matrice de transformation est mise à jour de façon séquentielle en utilisant le vecteur de cofacteurs de la ligne et la projection des paramètres actuels du modèle.

Lorsque HLDA est appliquée aux HMM pour la reconnaissance de la parole, de bons résultats sont généralement obtenus en utilisant les états liés des HMM comme classes pour HLDA, chacune étant représentée par une gaussienne à matrice de covariance pleine [9]. Plusieurs initialisations sont envisageables pour la matrice de projection. Même si une matrice identité est la solution la plus simple, nous avons observé des résultats légèrement meilleurs en utilisant le ratio de Fisher ou une solution LDA (cette dernière est utilisée par défaut dans nos expériences).

L'usage classique de la technique HLDA consiste, après avoir réduit de manière optimale un espace de grande dimension en un espace plus petit possédant de bonnes propriétés discriminantes, à de construire les nouveaux modèles acoustiques dans ce nouvel espace. Dans notre proposition, la transformation HLDA est utilisée pour définir un espace d'état discriminant pour les FAHMM. Pour ce faire, l'espace d'état est défini par les dimensions utiles de la projection HLDA.

En combinant la définition du sous-espace HLDA $x_t = A_p o_t$ avec le modèle FAHMM donné par l'équation (1), nous observons que la matrice pseudo-inverse de A_p (une matrice rectangulaire $p \times n$) peut être une bonne solution pour initialiser les matrices de chargement de facteurs. Pour nos expériences, la matrice inverse de Moore-Penrose a été utilisée [2], $\hat{C}_j = A_p^+$.

Une fois la matrice de chargement définie, les mixtures de gaussiennes des espaces d'état et de bruit d'observation doivent être initialisées. Elles sont dérivées directement de l'espace HLDA. Les paramètres peuvent alors être obtenus par un apprentissage des modèles dans l'espace HLDA ou directement en transformant les paramètres de l'espace des observations.

Avec la distribution des vecteurs d'états donnée par (2), les paramètres de la distribution par la premier méthode (*apprentissage*) sont obtenus, pour chaque état j, par un apprentissage EM operé sur le sous-espace d'observation résultant de la projection HLDA

$$\mu_{jk}^{(x)} = A_p \mu_{jk} \quad (6)$$

$$\Sigma_{jk}^{(x)} = \mathrm{diag}(A_p W_{jk} A_p^T) \quad (7)$$

avec μ_{jk} et W_{jk} le vecteur moyenne et la matrice de covariance totale des données originales associées à l'état j pour la gaussienne k. Dans la seconde méthode (*projection*), les vecteurs moyens suivent toujours l'équation (6)

mais les covariances deviennent

$$\Sigma_{jk}^{(x)} = \mathrm{diag}(A_p \Sigma_{jk} A_p^T) \quad (8)$$

avec μ_{jk} et Σ_{jk} obtenus par un apprentissage dans l'espace des observations.

A partir des statistiques de l'espace des observations complet, les valeurs initiales pour les paramètres du bruit d'observation sont obtenues en retirant la projection des paramètres de l'espace d'état des paramètres a priori de l'espace des observations. Une covariance diagonale est suffisante a priori dans la mesure où aucune compensation des éléments non-diagonaux n'est réalisée. Lors de l'initialisation, les mixtures d'observation et de bruit doivent avoir le même nombre de gaussiennes ($M^{(o)}$) et la distribution de l'espace d'état est mono-gaussienne. Sinon, l'association entre les composants des observations et de l'espace d'état serait indéfinie et complexe à établir.

Avec la distribution des vecteurs de bruit d'observation définie par l'équation (3), l'initialisation par *apprentissage* conduit à :

$$\mu_{jl}^{(o)} = \mu_{jl} - \hat{C}\mu_{j1}^{(x)} \quad (9)$$

$$\Sigma_{jl}^{(o)} = \Sigma_{jl} - \mathrm{diag}(\hat{C}\Sigma_{j1}^{(x)}\hat{C}^T) \quad (10)$$

et la méthode par *projection* modifie les covariances selon :

$$\Sigma_{jl}^{(o)} = \Sigma_{jl} - \mathrm{diag}(\hat{C}\mathrm{diag}(A_p\Sigma_{j1}A_p^T)\hat{C}^T) \quad (11)$$

Dans ce contexte, un grand soin doit être apporté à appliquer un seuil aux variances du bruit d'observation de sorte à pouvoir calculer les gaussiennes résultantes.

4. DESCRIPTION DU CORPUS ET DU SYSTÈME

Les expériences on été menées dans le cadre de la tâche de dictée vocale de textes, représentée par le corpus *Wall Street Journal* [7], et les conditions de test correspondent à l'évaluation ARPA Hub3 de 1995. Les données acoustiques d'apprentissage ont été prononcées par 355 locuteurs pour un total de 100 heures de parole. Le test Hub3 consiste en parole lue en studio par 20 locuteurs pour un total de 45 minutes. L'analyse acoustique produit des paramètres cepstraux à partir de l'échelle de fréquence Mel estimée sur la bande 0-8kHz toutes les 10ms. La moyenne cesptrale est retirée aux paramètres. Le vecteur acoustique d'observation de 39 dimensions est composé de 12 coefficients cepstraux avec la log-énergie, complétés par leur dérivée des premier et deuxième ordres.

Le système de reconnaissance de dictée vocale du LIMSI utilise des modèles de Markov cachés à densités continues avec des mélanges de gaussiennes pour la modélisation acoustique et des modèles linguistiques de type n-grammes estimés sur de grands corpus de textes. Chaque modèle phonétique dépendant du contexte est un HMM gauche-droit à états liés obtenus grâce à un arbre de décision.

La reconnaissance est opérée en trois étapes : 1) génération d'une hypothèse initiale, 2) adaptation des modèles et génération d'un graphe de mots, 3) adaptation des modèles et génération de l'hypothèse finale. L'hypothèse initiale est utilisée pour l'adaptation des modèles acoustiques à l'aide de la technique du MLLR [6] préalablement à la génération du graphe de mots. Un modèle de langage de type 3-grammes avec back-off est utilisé lors des deux premières étapes. L'hypothèse finale est engendrée à partir d'un modèle 4-grammes et des modèles acoustiques adaptés lors de la seconde étape.

TAB. 1: Taux d'erreur en mots (WER %) et nombre de paramètres par état (η) pour les systèmes HMM à matrices de covariance diagonales et pleines. $M^{(o)}$ est le nombre de gaussiennes par état.

	$M^{(o)}$	*1*	*8*	*16*	*32*
Diagonales	η	78	624	1248	2496
	WER	13.8	8.8	8.5	8.0
	$M^{(o)}$	*1*	*2*	*4*	*8*
Pleines	η	819	1638	3276	6552
	WER	10.9	9.5	9.2	10.3

L'apprentissage EM est réalisé avec une segmentation fixe (*apprentissage de Viterbi*). Les modèles acoustiques dépendants du genre sont obtenus par une adaptation MAP des modèles indépendants du genre [4]. Le jeu de modèles acoustiques comprend 4k phones en contexte, inter-mots et dépendants de la position dans le mot, comptant pour 9k états liés. Le vocabulaire de reconnaissance contient 65k mots avec 77k prononciations phonétiques. Les modèles de langage 3 et 4-grammes résultent de l'interpolation linéaire de modèles appris sur différents ensemble de données (transcriptions des données acoustiques d'apprentissage, journaux...). Un graphe de prononciation est associé à chaque mot permettant des prononciations alternatives.

5. RÉSULTATS EXPÉRIMENTAUX

Afin de définir une référence, une configuration de modèles HMM à matrices de covariance diagonales est évaluée. Tous les résultats rapportés ont été obtenus avec des modèles dépendant du genre du locuteur et adaptés de façon non supervisée au locuteur. Le tableau 1 donne les taux d'erreur en mots et le nombre de paramètres libres par états (η) pour quatre valeurs de $M^{(o)}$ (le nombre de composantes gaussiennes par état). On constate que le taux d'erreur décroît de manière constante avec le nombre de paramètres, jusqu'à atteindre 8% avec 32 gaussiennes par état.

Une expérience contrastive a été menée avec des HMM à base de covariances pleines. Les résultats obtenus avec de 1 à 8 gaussiennes par état sont donnés dans la seconde partie du tableau 1. Le taux d'erreur le plus bas est 9.2% avec 4 gaussiennes par état, c'est à dire environ 1% au dessus du meilleur résultat obtenu avec des matrices de covariance diagonales. Ce résultat est cohérent avec d'autres résultats rapportés pour de la parole conversationnelle [11], et il justifie le développement d'une modélisation intermédiaire.

Le tableau 2 présente les taux d'erreur pour plusieurs configurations de FAHMM avec le nombre de paramètres par état. La taille de l'espace d'état a été fixée à 13 avec un espace d'observation de 39 dimensions. Dans le cas de l'initialisation standard des paramètres, un ensemble de HMM à mixtures de gaussiennes avec matrices de covariance diagonales est utilisé pour initialiser les FAHMM comme proposé dans [8]. Une seule matrice de charge-

TAB. 2: Taux d'erreur en mots ($WER\%$) et nombre de paramètres par état (η) pour les 3 configurations des FAHMM (toutes avec $n = 39$ et $p = 13$) : standard, initialisations HLDA apprise et par projection. $M^{(x)}$ et $M^{(o)}$ sont respectivement le nombre de composants dans les espaces d'etat et d'observation.

$M^{(x)}$	$M^{(o)}$	*1*	*2*	*3*	*4*	*5*	*6*
		Initialisation standard					
6	η	715	793	871	949	1027	1105
	WER	9.2	8.9	8.8	8.9	9.1	8.8
		Init. HLDA (projection)					
	WER	9.0	8.5	8.5	8.4	8.6	8.3
		Init. HLDA (apprentissage)					
	WER	9.2	8.9	8.6	8.8	8.5	8.3
3	η	637	715	793	871	949	1027
	WER	9.5	9.4	9.0	8.7	8.7	8.3
1	η	585	663	741	819	897	975
	WER	11.0	10.0	9.9	9.2	9.4	9.2

ment de facteurs est utilisée par état, elle est partagée entre toutes les gaussiennes. Les deux premières lignes du tableau correspondent à l'initialisation standard avec 6 gaussiennes pour l'espace d'état et de 1 à 6 gaussiennes pour l'espace d'observation. Le meilleur taux d'erreur 8.8% est obtenu avec $M^{(o)} = 6$. Bien que le nombre de parametres indépendants par état reste peu élevé (1105) en comparaison des modèles à covariances diagonales, le coût de décodage est significativement plus important. Ceci explique qu'il est difficile d'aller au delà de la configuration 6×6.

Les résultats de l'initialisation discriminante des FAHMM sont donnés dans les deux lignes suivantes du tableau 2, pour l'initialisation HLDA par apprentissage d'une part et par projection d'autre part. Lorsque que le nombre de composants dédiés au bruit est augmenté, les performances s'améliorent plus rapidement pour la méthode *par projection*, bien que au final les 2 approches obtiennent des résultats comparables pour $M^{(o)} = 6$ (8.3%). Avec l'initialisation discriminante, le taux d'erreur est diminué de 0.5% pour la meilleure configuration ($M^{(x)}$=6, $M^{(o)}$=6) comparée à l'initialisation standard. Avec environ un millier de paramètres indépendants, le taux d'erreur est aussi inférieur à celui obtenu avec des modèles à matrices de covariance diagonales.

La dernière partie du tableau donne des résultats supplémentaires pour l'initialisation de type *apprentissage* en utilisant moins de composants pour l'espace d'état (3 et 1). Avec $M^{(x)} = 3$, les résultats sont meilleurs que ceux observés pour $M^{(x)} = 6$ à nombre de paramètres fixe. Finalement, les performances baissent avec $M^{(x)} = 1$. Ces résultats tendent à montrer que l'équilibre entre les valeurs de $M^{(x)}$ and $M^{(o)}$ est un point délicat pour obtenir de bonnes performances avec les FAHMM.

CONCLUSION

Les HMM à analyse factorielle ont été appliqués sur une tâche de reconnaissance de parole continue grand vocabulaire, utilisant 100 heures de données d'apprentissage. Le système intègre un modèle de langage 4-grammes appris sur un vocabulaire de 65k mots et des adaptations supervisée (MAP) et non-supervisée (MLLR) des modèles acoustiques FAHMM. Une méthode a été proposée pour améliorer l'initialisation des dimensions de l'espace d'état pour l'apprentissage des FAHMM, basée sur un critère discriminant (HLDA).

A la suite d'une série d'expériences, nous avons observé que les FAHMM appris selon l'approche proposée présentent des performances légèrement inférieures à celles des HMM à matrices de covariance diagonales (8.3% vs 8.0%) mais supérieures à celles des HMM à covariances pleines (8.3% vs 9.2%) et des FAHMM standards (8.3% vs 8.8%). De plus, si nous fixons les nombre de paramètres indépendants du système vers 1k par état, les FAHMM avec une initialisation discriminante donnent alors de meilleurs résultats que tous les autres modèles, y compris les modèles à covariance diagonale.

RÉFÉRENCES

[1] H. Attias. Independent factor analysis. *Neural Computation*, 11(4) :803–851, 1999.

[2] S. Campbell and C. Meyer. *Generalized Inverses of Linear Transformations*. Dover Publications, New-York, 1991.

[3] M. Gales. Semi-tied covariance matrices for hidden markov models. *IEEE Transactions on Speech and Audio Processing*, 7(3) :272–281, 1999.

[4] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2) :291–298, 1994.

[5] N Kumar and A Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech Communication*, 26(4) :283–297, 1998.

[6] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9 :171–185, 1995.

[7] D. Paul and J. Baker. The design for the wall street journal-based csr corpus. In *Proceedings of the ICSLP*, pages 899–902, Banff, 1992.

[8] A-V.I. Rosti and M.J.F. Gales. Factor analysed hidden markov models for speech recognition. *Computer Speech and Language*, 18(2) :181–200, 2004.

[9] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen. Maximum likelihood discriminant feature spaces. In *Proceedings of the IEEE ICASSP*, Istanbul, 2000.

[10] L. Saul and M. Rahim. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 8(2) :115–125, 2000.

[11] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig. The ibm 2004 conversational telephony system for rich transcription. In *Proceedings of the IEEE ICASSP*, volume I, pages 205–208, Philadelphia, 2005.