Estimation rapide de modèles de Markov semi-continus discriminants

Georges Linarès, Christophe Lévy, Jean-christophe Plagniol

Laboratoire Informatique d'Avignon
339 Chemin des meinajaries, BP 1228, 84911 Avignon, France
{georges.linares, christophe.levy, jean-christophe.plaignol}@univ-avignon.fr

ABSTRACT

In this paper, we present a fast estimation rule for MMIE (Maximum Mutual Information Estimation) training of semi-continuous HMM (Hidden Markov Models).

The first experiments validate this method by comparing our fast MMI estimator (FMMIE) and the original one. We observe that, on a digit recognition task, FMMIE and full MMIE estimation obtain similar results, when our method decreases significantly the computational time.

Then, we incorporate our semi-continuous MMIE models in a real-time Large Vocabulary Continuous Speech Recognition (LVCSR) system. The evaluation corpora is extracted from the French BN Evaluation campaign Ester. The results show that the proposed MMIE models outperform significantly the system based on continuous models while remaining at the same level of complexity.

1. INTRODUCTION

Différentes voies ont été explorées dans le passé pour limiter les ressources utilisées par les modèles acoustiques, tant en terme d'espace mémoire que de temps de calcul. Des solutions basées sur une modélisation par modèles de Markov semi-continus (SCHMM, Semi Continuous Hidden Markov Models) ont étés proposées dans la littérature ([3], [9], [7]). Dans ces architectures, les modèles partagent un dictionnaire commun de gaussiennes, les états étant caractérisés par un vecteur de poids généralement estimé par maximisation de la vraisemblance. Cette mutualisation massive des paramètres permet de réduire de façon très significative l'espace mémoire requis par le stockage des modèles acoustiques et peut limiter les problèmes d'estimation liés à des tailles de corpus d'apprentissage limitées. Par contre, le gain obtenu en terme de temps de calcul est moins décisif, les méthodes de calcul rapide des vraisemblances sélection de gaussiennes permettant d'atteindre le temps réel sur des systèmes grand vocabulaire tout en préservant la précision des modèles à plusieurs millions de paramètres.

Plusieurs méthodes d'estimation du dictionnaire de gaussiennes d'un SCHMM ont été proposées : dictionnaires multiples, gaussiennes issues d'un jeu de HMM classique ([11]), *etc*. L'estimation des poids, quand à elle, est généralement effectuée par maximisation de la vraisemblance (MLE, Maximun Likelihood Estimation) alors que l'apprentissage discriminant par MMIE s'est généralisé pour l'estimation des HMM continus ([1]), malgré l'augmentation très sensible du temps de calcul requis par ce type de techniques.

Dans ce papier, nous présentons une méthode d'estimation rapide des poids d'un SCHMM par maximisation de l'information mutuelle. Nous reformulons les règles de réestimation proposées dans [8] dans le cadre spécifique des modèles semi-continus et nous proposons une heuristique qui permet une estimation très rapide des poids. Cette méthode est d'abord évaluée sur un système mot-isolé petit vocabulaire, puis sur un système de reconnaissance grand vocabulaire.

La première partie de cet article décrit les principes de l'estimation des paramètres par maximisation de l'information mutuelle. La seconde partie décrit des expériences confrontant l'approximation proposée à la règle initiale dans le cadre d'un système embarqué de reconnaissance de chiffres. Dans la troisième partie, la méthode proposée est évaluée en reconnaissance de parole continue grand vocabulaire, sur une tâche de transcription temps-réel d'émission radiophoniques.

2. MMIE RAPIDE POUR L'ESTIMATION DES SCHMM

L'estimation des modèles acoustiques par maximisation de l'information mutuelle a donné lieu à de nombreux travaux ces dernières années. Le principe général est de minimiser le risque d'erreur en maximisant l'écart de vraisemblance entre la bonne transcription et les mauvaises. La recherche des paramètres λ améliorant la capacité discriminante des modèles est réalisée par des algorithmes d'optimisation maximisant la fonction objective :

$$F_{mmie}(\lambda) = \sum_{r=1}^{R} log(\frac{P_\lambda(O_r|M_{w_r})P(w_r)}{\sum_{\tilde{w}} P_\lambda(O_r|M\tilde{w})} \qquad (1)$$

où w_r est la transcription correcte, M_w la séquence de modèles correpondant à l'hypothèse w, $P(w)$ la probabilité linguistique et O_r une séquence d'observations. Le dénominateur somme les produits des probabilités acoustiques et linguistiques sur toutes les hypothèses possibles.

Une des difficultés majeures rencontrées pour l'estimation des paramètres optimaux réside dans la complexité de la fonction objective qui intègre, dans son dénominateur, l'ensemble des chemins incorrects suceptibles d'être empruntés (et les séquences de modèles associées). Pour atteindre une complexité acceptable, il faut limiter le nombre de ces chemins, par exemple en ne gardant que

les n meilleurs issus d'un treillis de mots ou de phonèmes ([10]). L'estimation des modèles par MMIE reste néanmoins bien plus coûteuse que par MLE, qui ne nécessite pas l'évaluation d'hypothèses incorrectes.

Dans le cas particulier des modèles semi-continus, seule la ré-estimation des poids est nécessaire. Par ailleurs, le partage massif des gaussiennes permet de réduire considérablement la complexité de l'estimation des paramètres. En effet, le calcul des vraisemblances est limité au nombre réduit des gaussiennes du dictionnaire. D'autre part, la présence des mêmes composantes dans tous les états peut permettre une sélection directe des gaussiennes discriminantes. Nous mettons ce point en évidence en développant la formule de ré-estimation des poids proposée dans [8]. Dans cet article, les auteurs montrent que les poids $\tilde{c}_{jm}$ maximisant la fonction objective peuvent être obtenus par maximisation de l'expression suivante :

$$\sum_{j,m} \left[\gamma_{jm}^{num} log(\tilde{c}_{jm}) - \frac{\gamma_{jm}^{den}}{c_{jm}} \tilde{c}_{jm} \right] \quad (2)$$

où γ_{jm}^{num} et γ_{jm}^{den} sont respectivement les taux d'occupation estimés sur les exemples corrects (num) et incorrects (den) ; c_{jm}, le poids de la composante m de l'état j à l'itération précédente ; $\tilde{c}_{jm}$, le nouveau poids de la composante (j, m).

En optimisant chaque terme de cette somme pour un ensemble de poids fixé, la convergence peut être obtenue après quelques itérations. Chacun de ces termes étant convexe, la formule de mise à jour se déduit directement de l'équation précédente :

$$\tilde{c}_{jm} = \frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} c_{jm} \quad (3)$$

où γ_{jm}^{k} est la probabilité d'être dans la composante m de l'état j estimée sur l'ensemble de données Ω_k, qui regroupe les trames associées à l'état k. Ce taux d'occupation peut s'exprimer en fonction des vraisemblances $L()$:

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} \frac{L(X|S_j)}{\sum_i L(X|S_i)} \frac{c_{jm} L(X|G_{jm})}{L(X|S_j)} \quad (4)$$

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} c_{jm} \frac{L(X|G_{jm})}{\sum_i L(X|S_i)} \quad (5)$$

En isolant, dans le dénominateur, la vraisemblance de la trame X sachant l'état S_k, on obtient :

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} c_{jm} \frac{L(X|G_{jm})}{L(X|S_k) + \sum_{i \neq k} L(X|S_i)} \quad (6)$$

Dans des modèles semi-continus, les composantes gaussiennes G_{jm} sont indépendantes de l'état j ; dans ce cas, les taux d'occupations peuvent s'écrire :

$$\gamma_{jm}^{k} = \sum_{X \in \Omega^k} c_{jm} \frac{L(X|G_{km})}{L(X|S_k) + \sum_{i \neq k} L(X|S_i)} \quad (7)$$

En notant $\epsilon_k = \sum_{i \neq k} L(X|S_i)$, le rapport des taux d'occupations devient :

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} = \frac{\sum_{X \in \Omega^j} \frac{L(X|G_{jm})}{L(X|S_j) + \epsilon_j}}{\sum_l \sum_{X \in \Omega^l} \frac{L(X|G_{lm})}{L(X|S_l) + \epsilon_l}} \quad (8)$$

Nous proposons d'approximer le rapport précédent par :

$$\frac{\gamma_{jm}^{num}}{\gamma_{jm}^{den}} = \frac{c_{jm}}{\sum_l c_{lm}} \quad (9)$$

En introduisant cette approximation dans l'équation 3, on obtient la formule de ré-estimation rapide des poids :

$$\tilde{c}_{jm} = \frac{c_{jm}^2}{\sum_l c_{lm}} \quad (10)$$

Après ré-estimation, les vecteurs de poids de chaque état sont re-normalisés.

Cette fonction de mise à jour des poids à chaque itération ne nécessite pas l'estimation des vraisemblances sur le corpus d'apprentissage. En terme de temps de calcul, son coût est limité à celui de l'apprentissage du modèle MLE initial, les poids MMIE se déduisant directement des poids MLE (*cf.* éq. 10).

Dans la section suivante, nous comparons les résultats obtenus par cette méthode aux résultats obtenus par estimation MMIE standard.

3. VALIDATION EXPÉRIMENTALE EN PETIT VOCABULAIRE

De façon à valider expérimentalement la formule de ré-estimation proposée, nous avons entraîné des modèles semi-continus par MMIE avec les formules originelles, puis avec l'approximation proposée. Le système est évalué sur une tâche de reconnaissance de chiffres isolés, avec une faible quantité de données d'apprentissage, ce qui autorise un calcul exact des formules de ré-estimation.

La plateforme utilisée est celle développée dans [5], dans le cadre d'un système conçu pour la reconnaissance de petits vocabulaires sur un téléphone portable.

3.1. Système de base

Ce système a été développé pour la reconnaissance de chiffres sous la contrainte de ressources matérielles très limitées. Il repose sur une modélisation par SCHMM et un processus d'adaptation à l'environnement acoustique avec très peu de données. Un GMM (Gaussian Mixture Model) initial, dont seront dérivés les GMM d'état, a été appris avec le corpus BREF120 ([4]), sur une centaine d'heure de parole.

Le GMM initial est ensuite adapté par MAP (Maximum A Postériori) sur un corpus de taille réduite, mais néanmoins caractéristique des conditions de test. Ces ensembles d'adaptation et de test sont issus de la base BDSONS ([2]), qui a été divisée en 2 sous-ensembles afin d'obtenir un corpus d'adaptation (destiné à l'adaptation du GMM et à la ré-estimation des poids) et un corpus de test. Ces 2 ensembles contiennent respectivement 700 et 2300 occurences de chiffres. Les vecteurs acoustiques sont composés de 12 coefficents PLP, auxquels on a ajouté l'énergie du signal. Ici, les paramètres dynamiques (dérivées premières et secondes) ne sont pas utilisés.

3.2. Résultats

Le tableau 1 présente les performances du système dans une configuration de reconnaissance embarquée. A par-

tir du même dictionnaire de gaussiennes, 2 approches sont utilisés pour l'estimation des poids : MMIE avec les formules complètes de ré-estimation (*cf.* éq. 3) et la ré-estimation rapide que nous proposons (FMMIE).

TAB. 1: Taux d'erreur mot en fonction de la taille du GMM initial exprimée en nombre de gaussiennes (# GAUSS) pour les modèles MMIE et MMIE rapide (FMMIE). Mesure effectuée sur 2300 chiffres issus de la base BDSONS.

# GAUSS	MMIE	FMMIE
216	4,26%	4,09%
432	2,70%	2,39%
864	2,48%	2,57%
1728	2,30%	2,00%

Les résultats obtenus montrent que les deux algorithmes se comportent globalement de la même façon. Dans la plupart des cas, FMMIE obtient des résultats légèrement meilleurs mais toujours dans l'intervalle de confiance. Par ailleurs, alors que les performances du modèle MMIE standard s'améliorent régulièrement avec l'augmentation de la taille du GMM initial, FMMIE semble légèrement plus irrégulier. Là aussi, ces variations restent limitées et toujours incluses dans l'intervalle de confiance.

4. EVALUATION SUR UN SYSTÈME GRAND VOCABULAIRE

De façon à évaluer notre méthode d'estimation rapide des poids MMIE dans un contexte de LVCSR, nous avons estimé un modèle semi-continu compact qui a été intégré au système "Broadcast News" du LIA ([6]). Les expériences ont ensuite été menées en utilisant le cadre expérimental défini pour la tâche temps-réel de la campagne d'évaluation ESTER.

4.1. Estimation du HMM semi-continu

Construction du dictionnaire de gaussiennes Les gaussiennes du dictionnaire sont extraites d'un HMM continu de petite taille. Il s'agit de modèles indépendants du contexte avec 64 gaussiennes par état (C-CI). Les vecteurs acoustiques utilisés sont composés de 12 coefficients PLP (Perceptual Linear Prediction), de l'énergie du signal dans la fenêtre d'analyse, puis des dérivées premières et secondes de ces 13 premiers coefficients. Enfin, les paramètres sont centrés et réduits sur une fenêtre glissante de $5s$. La première étape du processus d'estimation consiste à entraîner ces modèles continus classiques, pour lesquels chaque GMM est estimé sur les données spécifiques de l'état par un algorithme de type EM (Expectation-Maximisation). Les gaussiennes issues de cet apprentissage sont ensuite regroupées dans le dictionnaire qui est constitué d'environ 7000 éléments. Le modèle initial a été entraîné sur le corpus d'apprentissage d'ESTER, qui contient 200 heures de parole transcrite, issue d'émissions radiophoniques du groupe Radio-France. L'essentiel du corpus est composé de journaux d'information, avec une diversité de locuteurs assez importante.

Estimation des modèles semi-continus par MLE L'étape suivante constiste à entraîner des modèles dépendant du contexte par ré-estimation du vecteur de poids de chaque état. Le jeu de modèles utilisé est constitué de 10000 triphones partageant un ensemble de 3600 états émetteurs. Le partage des états a été déterminé par un arbre de décision.

Pour chacun de ces 3600 états, nous estimons d'abord un vecteur de poids par MLE sur le corpus d'apprentissage aligné par un HMM continu modélisant le même jeu d'unités acoustiques. Nous obtenons alors un premier ensemble de modèles dépendants du contexte (SC0), partageant les 7000 gaussiennes du dictionnaire, et dont chaque état est caractérisé par un vecteur de 7000 poids.

Les composantes significatives des états de SC0 sont ensuites selectionnées de façon à ce que la somme de leur poids atteigne un seuil γ fixé *a priori*. Pour une valeur de γ à $0,9999$, le modèle contient 580000 poids, ce qui représente une moyenne de 161 gaussiennes indexées par état. On obtient un modèle (SC-MLE) de 1100000 paramètres, soit une complexité qui dépasse très largement celle du modèle continu dont est issu le dictionnaire de gaussiennes (C-CI).

Pour comparer cette approche avec les modèles continus, nous avons entraîné un autre modèle semi-continu par MLE. Cette fois çi, γ est choisi de façon à atteindre une complexité équivalente à celle du modèle continu C-CI. Pour une valeur de γ à $0,3$, nous obtenons un modèle compact (SC-MLE-C) contenant 610000 paramètres.

Estimation des modèles semi-continus par MMIE Avec des corpus de taille importante ou sur des tâches grand vocabulaire, l'estimation exhaustive du dénominateur de la fonction objective MMIE ne peut pas être réalisée dans un temps raisonnable. Nous utilisons directement l'estimateur rapide proposé qui sera comparé aux approches à base de HMM continus ou semi-continus estimés par MLE.

Les poids MMIE sont ré-estimés à partir du modèle SC0 décrit dans le paragraphe précédent. Les composantes les plus significatives sont ensuite sélectionnées, comme pour le modèle SC-MLE, en supprimant les gaussiennes de poids très faible. Les modèles obtenus sont composés d'environ 80000 poids, ce qui représente un moyenne de 22 gaussiennes par état.

On obtient finalement un modèle (SC-FMMIE) d'une complexité de l'ordre de 615000 paramètres pour une valeur de γ identique à celle utilisée pour le modèle SC-MLE (qui est, lui, composé de plus de 1 million de paramètres). De nombreuses composantes ne sont donc pas spécifiques à un état, ce qui est une conséquence naturelle de l'estimation par maximum de vraisemblance des GMM, mais aussi probablement lié au manque de précision des modèles initiaux (indépendants du contexte et composés de seulement 7000 gaussiennes).

4.2. Résultats

Le tableau 2 présente les résultats obtenus sur 1 heure de parole extraite du corpus de développement de la base ESTER.

On peut d'abord constater que le modèle semi-continu entraîné par MMIE (SC-FMMIE) améliore les performances du modèle continu indépendant du contexte (C-CI) de

3,1% (en absolu). Dans le même temps, le nombre de paramètres n'a augmenté que de 12,7%. Le modèle semi-continu compact entraîné par MLE (SC-MLE) obtient, lui, un taux d'erreurs nettement inférieurs au modèle continu (35,4% contre 43,3% de WER pour le C-CI).

La coupure appliquée au modèle SC-MLE ($\gamma = 0,3$) pour obtenir le modèle SC-MLE-C (dont la complexité est comparable aux modèles SC-FMMIE et C-CI) est très stricte. Elle conduit à une dégradation conséquente de la qualité de ces modèles ; le taux d'erreur atteint 50,8% (soit une augmentation absolu de WER de 7,5%). Sans cette coupure (modèle SC-MLE) le nombre de paramètre est bien plus élevé (+202% par rapport à C-CI), mais le taux d'erreur s'abaisse à 35,4% (-7.5% en absolu par rapport à C-CI). Néanmoins, ces performances restent éloignées de celles obtenues par un modèle continu de grande taille : en utilisant un modèle de plus grande complexité (1000 états émetteurs, 4,8 millions de paramètres), le taux d'erreur est de 29,1% (sur la même base de test).

TAB. 2: Résultats (en taux d'erreur-mot) et complexité (en nombre de paramètres, noté # PAR) des systèmes basés sur des HMM continus indépendants du contexte (C-CI), des SCHMM avec estimation MMIE (SC-FMMIE), MLE réduit à 610000 paramètres (SC-MLE-C) et MLE (SC-MLE). Evaluation portant sur 1 heure d'émission radiophonique issue du corpus ESTER.

	C-CI	SC-FMMIE	SC-MLE-C	SC-MLE
WER	43.3%	40.2%	50.8%	35.4%
# PAR	544k	615k	610k	1100k

5. CONCLUSION

Nous avons présenté une technique rapide de ré-estimation des poids MMIE dans le cadre de HMM semi-continus. Les évaluations comparatives ont montré que cet algorithme permet d'obtenir des résultats proches de ceux obtenus par la méthode initiale (et parfois légèrement meilleurs), malgré une relative instabilité. Cependant, son principal intérêt tient à la consommation très faible de temps CPU qu'elle requiert : partant de SCHMM estimés par MLE, le coût additionel d'estimation des poids MMIE est quasiment nul.

Les premières évaluations que nous avons menées en LVCSR sont encourageantes. Elles montrent le potentiel des SCHMM discriminants dans des contextes de ressources mémoires et CPU limitées. Nous envisageons de poursuivre cette étude dans deux directions. D'une part, des évaluations complémentaires doivent être menées de façon à mesurer les performances de cette technique en fonction de la complexité des modèles. D'autre part, la stratégie de constitution du dictionnaire de gaussiennes peut être affinée.

Enfin, ce type d'architecture pourrait permettre des adaptations rapides à l'environnement et/ou au locuteur, difficiles à intégrer dans des systèmes contraints par la vitesse de décodage ou par les ressources mémoire disponibles.

RÉFÉRENCES

[1] L.R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1986)*, pages 49–52, Tokyo, Japan, April 1986.

[2] R. Carré, R. Descout, M. Eskénazi, J. Mariani, and M. Rossi. The French language database : defining, planning and recording a large database. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1984)*, pages 324–327, San Diego, California, USA, March 1984.

[3] Xuedong Huang, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, and Ronald Rosenfeld. The SPHINX-II speech recognition system : an overview. *Computer Speech and Language*, 7(2) :137–148, 1993.

[4] L.F. Lamel, J.L. Gauvain, and M. Eskénazi. BREF, a large vocabulary spoken corpus for French. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'1991)*, pages 505–508, Gênes, Italie, September 1991.

[5] C. Lévy, G. Linarès, P. Nocera, and J.F. Bonastre. *Embedded mobile phone digit-recognition*, chapter 7 in Digital Signal Processing for In-Vehicle and Mobile Systems 2. Springer Science, H. Abut, J.H.L. Hansen and K. Takeda edition, 2006, *à paraître*.

[6] G. Linarès, P. Nocéra, D. Matrouf, F. Béchet D. Massonié, and C. Fredouille. Le système de transcription du lia pour ester-2005, 2005.

[7] Brian Kan-Wing Mak. Towards A compact speech recognizer : Subspace distribution clusteringhidden markov model. Technical Report CSE-TH-98-001, 20, 1998.

[8] D. Povey and P. Woodland. Frame discrimination training of hmms for large vocabulary speech recognition, 1999.

[9] T. Vaich and A. Cohen. Comparison of continuous-density and semi-continuous hmm in isolated words recognition systems. In *EUROSPEECH'99*, pages 1515–1518, 1999.

[10] V. Valtchev, J. Odell, P. Woodland, and S. Young. Mmie training of large vocabulary recognition systems, 1997.

[11] K.F. Lee X. D. Huang and H. Hon. On semi-continuous hidden markov modeling. In *ICASSP'90*, pages 689–692, 1990.