Modélisation acoustique compacte basée sur les sous-espaces de mixture de gaussiennes

Mohamed Bouallegue, Driss Matrouf, Georges Linarès

Laboratoire d'Informatique d'Avignon, Université d'Avignon et des Pays de Vaucluse, France
mohamed.bouallegue@etd.univ-avignon.fr, driss.matrouf@univ-avignon.fr, georges.linares@univ-avignon.fr

ABSTRACT

In the context of HMM-based speech recognizers, each HMM-state distribution is modeled independently from to the other and has a large amount of parameters. In spite of using state-tying techniques, the size of the acoustic models stays large and certain redundancy remains between states. In this paper, we investigate the capacity of the Subspace Gaussian Mixture approach to reduce the acoustic models size while keeping good performances. We introduce a simplification concerning state specific Gaussians weights estimation, which is a very complex and time consuming procedure in the original approach. With this approach, we show that the acoustic model size can be reduced by 92% with almost the same performance as the standard acoustic modeling.
Keywords: Compact Acoustic Models, Subspace Gaussian Mixture, Embedded speech recognition, Gaussian Mixture Models, Hidden Markov Models

1. Introduction

La plupart des Systèmes de Reconnaissance Vocale Continue sont basés sur les Modèles de Markov Cachés (HMM) qui représentent les unités élémentaires de parole, typiquement les phonèmes ou les triphones. Habituellement, les probabilités des états sont estimées en utilisant les Modèles de Mélanges Gaussiens (GMM) qui offrent plusieurs avantages : des formalismes mathématiques bien établis, un apprentissage automatique des paramètres et une amélioration de la performance. Pour atteindre une bonne performance, le nombre d'états de HMM et par conséquent le nombre de paramètres à estimer devient de plus en plus important (des milliers de paramètres). Dans ce document, nous aborderons le problème de la réduction du nombre de paramètres en gardant un performances proche de Baseline. D'autre travaux ont traité ce problème, principalement dans le but de réduire la quantité de mémoire pris par les modèles acoustiques [1, 2]. Dans les Modèles Markov semi-continus (SCHMM : Semi Continuous Hidden Markov Models), les modèles partagent un dictionnaire commun de gaussiennes. C'est un modèle générique qui représente tout l'espace acoustique, les états étant caractérisés par un vecteur de poids généralement estimé par maximisation de la vraisemblance. Une version détaillée peut être trouvée dans [3].

Le technique du *tying* permet une réduction significative de la complexité du modèle mais avec une perte significative de performance [4]. D'autres chercheurs étendent cette modélisation à l'aide de fonctions de transformations compactes qui modèlisent les différents états de HMM depuis un modèle générique [1].

Malgré l'utilisation de techniques d'attachement : *tying*, la taille des modèles acoustiques reste importante et certaines redondances demeurent entre les états. Dans cet article, nous proposons d'utiliser l'approche du Subspace Gaussian Mixture (SGM) pour réduire la taille du modèle acoustique. Tous les GMM des états sont dérivés d'un seul GMM appelé GMM-UBM (Universal background model) avec un nombre limité de paramètres spécifiques à l'état. Cette approche est un peu similaire à Eigenvoice [5] et Cluster Adaptive training [6]. Dans l'approche SGM, les poids spécifiques aux états sont estimés en utilisant une procédure complexe et très longue. Nous remplaçons cette étape en utilisant une simple estimation EM on conserve les N meilleurs poids dans chaque HMM.

Nous effectuons un bref rappel des modélisations standards de modèles acoustiques [7] dans la section 2. Nous décrivons l'approche proposée pour des modèles acoustiques compactes en utilisant l'approche SGM dans la section 3, en mettant en évidence la différence par rapport aux autres approches. Dans le section 4, nous présentons les résultats expérimentaux. Pour finir, conclure l'article dans la section 5.

2. Structure standard de modèle acoustique pour la reconnaissance de parole

Le système de base adopté dans ce travail utilise une architecture HMM gauche-droite avec 10 002 phonèmes contextes-dépendants. Pour réduire la taille des modèles acoustiques, nous avons utilisé le principe de regroupement (*clustering*) des états, dans lequel on remplace deux états ou plus, très similaires (les paramètres GMM sont similaires), par un seul état, en attachant ces états ensemble [8]. Cette technique nous permet de modéliser 10 002 phonèmes contextes-dépendants par 3327 états au lieu de 36 006, avec 64 gaussiennes par état et 39 coefficients PLP (Perceptual Linear Predictive) par trame (13 paramètres avec dérivées première et seconde). Même la technique d'attachement d'état *tying* a été utilisée,

la taille des modèles acoustiques reste important et quelques redondances demeurent entre les états. Dans le prochain paragraphe, nous décrivons l'utilisation de l'approche SGM pour réduire cette redondance, par conséquent, réduire la taille des modèles.

3. SGM pour la modélisation acoustique compacte

Dans l'approche de modélisation SGM, tous les GMM des états sont dérivés d'un modèle générique GMM-UBM : UBM=$(\alpha_g, m_g, \Sigma_g)$, où α_g, m_g et Σ_g représentent le poids, la moyenne et la matrice de covariance du g^{th} gaussienne. soit m le super-vecteur de moyennes obtenu par concaténation de toutes les moyennes de gaussiennes. Dans la modélisation SGM, le super-vecteur de moyennes de l'état s est une variable aléatoire donnée par :

$$\mathbf{m}_{(s)} = m + \mathbf{U}\mathbf{x}_{(s)} \tag{1}$$

Où $\mathbf{m}_s$ est le super-vecteur de moyennes d'un état s (vecteur à variable aléatoire), $\mathbf{U}$ est la matrice de variabilité inter-états (matrice $MD \times R$) de faible rang R. Où M est le nombre de gaussiennes dans l'UBM et D est la taille du paramètre cepstral. les vecteurs $\mathbf{x}_s$ sont les facteurs d'état de taille R. $\mathbf{x}_s$ sont supposés normalement distribués sur $\mathcal{N}(0, I)$. Durant la phase d'entraînement, la matrice $\mathbf{U}$ est estimée sur toutes les données d'apprentissage et $x_{(s)}$ est estimé sur les données de chaque état [9].

Dans l'approche SGM [10], les poids spécifiques des gaussiennes sont dérivés du modèle du monde UBM par :

$$w_g^s = \frac{exp\ \mathbf{w}_g^T \mathbf{x}_s}{\sum_{g'=1}^{M} exp\ \mathbf{w}_{g'}^T \mathbf{x}_s} \tag{2}$$

Où $\mathbf{x}_s \in \Re^R$ est le " vecteur d'état " avec R la dimension du sous-espace. $\mathbf{w}_g$ est le vecteur poids dépendant de la gaussienne mais pas de l'état.
Cet estimation des poids spécifiques à l'état est difficile à réaliser : C'est une dérivation basée sur une combinaison de l'inégalité Jensen- type, des expansions de séries de Taylor locales de second ordre et une modification de la résultante de la fonction quadratique auxiliaire qui assure la stabilité tout en conservant le même gradient local [11]. Comme alternative, nous proposons de ré-estimer les poids des gaussiens par une simple itération EM (algorithme de maximisation d'espérance). Sa estimation est beaucoup plus simple et nous permet de gagner en temps de calcul. Soit w_g le poids de la gaussienne g dans l'UBM. Le poids w_g^s de cette gaussienne dans l'état s est calculé comme suit :

$$w_g^s = \frac{\sum_{x \in s} P(g|x)}{N_s} \tag{3}$$

où,

$$P(g|x) = \frac{w_g * f(x|g)}{\sum_{g'} w_g * f(x|g')} \tag{4}$$

$N_{(s)}$ est le nombre des trames de l'état (s), $P(g|x)$ correspond à la vraisemblance des données relatives à l'état s pour la gaussienne g.

Pour une réduction plus importante du nombre de paramètres dans un état GMM, nous sélectionnons, pour un état s, les $\mathbf{N}$ gaussiennes possédant les poids les plus importants et nous ignorons les autres. $\mathbf{N}$ est choisi telle manière que la somme des poids des gaussiennes choisies atteigne un seuil prédéfini. Dans notre expérience, 333 000 gaussiennes ont été choisis, ce qui représente une moyenne de 100 gaussiennes sélectionnées par état.

3.1. Estimation de U et variable latente x_s

Soient $\mathbf{N}_{(s)}$ et $\mathbf{X}_{(s)}$ des vecteurs contenant respectivement les statistiques d'état d'ordre zero et de premier ordre :

$$\mathbf{N}_s[g] = \sum_{t \in s} \gamma_g(t);\ \{\mathbf{X}_{(s)}\}_{[g]} = \sum_{t \in (s)} \gamma_g(t) \cdot t \tag{5}$$

où $\gamma_g(t)$ est la probabilité *a posteriori* de la gaussienne g pour l'observation de la trame t. L'équation $\sum_{t \in s}$ est la somme de toutes les trames appartiennant à l'état s. Nous présentons maintenant les équations d'estimation du vecteur état. Soit $\overline{\mathbf{X}}_{(s)}$ les statistiques d'état dépendantes définies comme suit :

$$\{\overline{\mathbf{X}}_{(s)}\}_{[g]} = \{\mathbf{X}_{(s)}\}_{[g]} - \mathbf{m}_{[g]} \cdot \sum_{h \in s} \mathbf{N}_{(s)}[g] \tag{6}$$

Soit $\mathbf{L}_{(s)}$ la matrice $R \times R$, et $\mathbf{B}_{(s)}$ un vecteur de dimension R, définis par :

$$\begin{aligned} \mathbf{L}_{(s)} &= \mathbf{I} + \sum_{g \in \text{UBM}} \mathbf{N}_{(s)}[g] \cdot \{\mathbf{U}\}_{[g]}^t \cdot \mathbf{\Sigma}_{[g]}^{-1} \cdot \{\mathbf{U}\}_{[g]} \\ \mathbf{B}_{(s)} &= \sum_{g \in \text{UBM}} \{\mathbf{U}\}_{[g]}^t \cdot \mathbf{\Sigma}_g^{-1} \cdot \{\overline{\mathbf{X}}_{(s)}\}_{[g]}, \end{aligned} \tag{7}$$

En utilisant $\mathbf{L}_{(s)}$ et $\mathbf{B}_{(s)}$, $\mathbf{x}_{(s)}$ obtenu par l'équation suivante :

$$\mathbf{x}_{(s)} = \mathbf{L}_{(s)}^{-1} \cdot \mathbf{B}_{(s)} \tag{8}$$

La matrice $\mathbf{U}$ peu être estimée ligne par ligne, avec $\{\mathbf{U}\}_{[g]}^i$ étant la ligne i^{th} de $\{\mathbf{U}\}_{[g]}$ alors :

$$\mathbf{U}_{[g]}^i = \mathbf{L}\mathbf{U}_g^{-1} \cdot \mathbf{R}\mathbf{U}_g^i, \tag{9}$$

où $\mathbf{R}\mathbf{U}_g^i$ et $\mathbf{L}\mathbf{U}_g$ sont donnés par :

$$\begin{aligned} \mathbf{L}\mathbf{U}_g &= \sum_s \sum_{h \in s} (\mathbf{L}_{(s)}^{-1} + \mathbf{x}_{(s)} \mathbf{x}_{(s)}^T) \cdot \mathbf{N}_{(s)}[g] \\ \mathbf{R}\mathbf{U}_g^i &= \sum_s \sum_{h \in s} \{\overline{\mathbf{X}}_{(s)}\}_{[g]}[i] \cdot \mathbf{x}_{(s)} \end{aligned} \tag{10}$$

L'algorithme 1 présente la stratégie adoptée pour estimer la matrice de variabilité des états avec les équations développées ci-dessus (la fonction de vraisemblance standard peut être utilisée pour estimer la convergence).

For each state s : $\mathbf{x}_{(s)} \leftarrow 0$, $\mathbf{U} \leftarrow random$;
Estimate statistics : $\mathbf{N}_{(s)}$, $\mathbf{X}_{(s)}$ (eq.5);
for $i = 1$ *to nb_iterations* **do**
  for *all s and h* **do**
    Center statistics : $\overline{\mathbf{X}}_{(s)}$ (eq.6);
    Estimate $\mathbf{L}_{(s)}$ and $\mathbf{B}_{(s)}$ (eq.7);
    Estimate $\mathbf{x}_{(s)}$ (eq.8);
  end
  Estimate matrix $\mathbf{U}$ (eq. 9 and 10) ;
end

Algorithm 1: Estimation algorithm of $\mathbf{U}$ and latent variable $\mathbf{x}$.

3.2. Modèles acoustiques compacts : Subspace GMM contre SCHMM

Dans le modèle de Markov semi-continu (SCHMM), qui peut être considéré comme une forme spéciale de mélange de HMM continu, 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. Par contre, le gain obtenu en terme de temps de calcul est moins décisif [3].

La principale différence entre la modélisation acoustique SGM et celle du SCHMM est que dans le premier cas les moyennes et les poids sont dépendants des états, alors que dans le second cas, seuls les poids gaussiens sont dépendants des états.

Pour comparer l'approche SGM proposée avec celle du SCHMM, nous avons comparé les performances en taux d'erreur mot, de deux modèles acoustiques, ayant le même nombre de paramètres, de ces deux différentes approches. Pour le SGM nous avons utilisé un UBM de 500 gaussiennes et une matrice de variabilités inter-états $\mathbf{U}$ de rang 88. Dans le système SCHMM, l'UBM comporte 600 gaussiennes. En re-entraînant seulement les 100 meilleurs gaussiennes, nous avons évalué le même nombre de paramètres pour les deux systèmes, ce qui correspond à une réduction de 89% du nombre de paramètres.

Le tableau 1 donne les résultats du système de base par rapport aux modèles compacts SCHMM et SGM. En utilisant l'approche proposée, nous pouvons observer un gain relatif d'environ 37,4% par rapport au SCHMM et presque les mêmes performances que les modèles acoustiques standards (perte relative de 3%).

Radio broadcasts	Baseline	SCHMM	modèle compacte SGM
RTM (2h)	35.51	44.31	35.81
RFI (1h30mn)	25.72	38.36	25.77
INFO (2h)	25.97	39.60	28.37
CLASSIQUE (1h)	21.7	34.5	22.4
Total (6h30)	28.06%	39.97%	29.09%
réduction de la taille de modèle	-	89%	89%

Table 1: *Résultats du modèle standard, du système compact SCHMM et SGM, en % de taux d'erreur de mots .*

4. Cadre experimental et résultats

4.1. Le système de LIA

Les expériences sont effectuées en utilisant le système du LIA (BN) qui a été utilisé lors de la campagne d'évaluation ESTER. Ce système repose sur le décodeur de base HMM, développé au Laboratoire Informatique d'Avignon (LIA) : Speeral [12]. Les modèles acoustiques sont basés sur les HMM, dépendants du contexte avec des *cross-word* triphones. Les modèles de langage sont des trigrammes classiques estimés sur environ 200M de mots du journal Le Monde et sur environ 1M de mots du corpus de radio ESTER. Le lexique contient 67 000 mots. Dans ces expériences, seul une passe de décodage est exécuté en 3X le temps réel.

Les modèles acoustiques testé sont entrainé sur le corpus de traine fourni par la campagne d'évaluation ESTER. Le corpus ESTER se compose des radios françaises du groupe Radio-France. Nous testons notre approche sur 7,5 heures de parole extraites de l'ensemble de test d'ESTER.

4.2. Resultats

Dans ce paragraphe, nous présentons les résultats obtenus par notre modèle acoustique compact. D'abord, nous établissons les équations pour calculer le nombre de paramètres pour le système de base et pour ceux de nos nouveaux modèles. Le nombre de paramètres du système de base reste important même après l'état de groupement. Ce nombre est donné par l'équation (11) :

$$nbs * [\quad \underbrace{(2 * nbp + 1)}_{\text{one Gaussian}} * nbg \quad] \tag{11}$$

où nbg est le nombre de gaussiennes (fixé à 64), nbs le nombre d'états émetteurs (3300 états), et nbp le nombre des paramètres acoustiques (39). Pour notre modèle, nous avons les paramètres du GMM-UBM, la matrice de variabilité inter-états $\mathbf{U}$, la matrice de facteurs d'état $\mathbf{X}$ et les vecteurs de nouveaux poids. Les paramètres sont donnés par l'équation suivante :

$$\underbrace{(2 * nbp+1) * nbg}_{GMM-UBM} + \underbrace{ndg * nbp * R}_{U\ matrix} + \underbrace{nbs * R}_{X\ matrix} + \underbrace{nbs * n\_best}_{new\ weights} \tag{12}$$

où n-best est le nombre des gaussiennes sélectionnées par état. Dans nos expériences, n-best est égal à 100.

Pour étudier l'impact de la variation de la taille du modèle acoustique sur la performance du modèle, nous pouvons changer la taille de l'UBM-GMM ou le rang de la Matrice U. Dans nos expériences, nous avons choisi de fixer la taille du UBM-GMM à 600 gaussiennes et de tester plusieurs rangs de la matrice $\mathbf{U}$ (40, 100, 150, 190 et 250).

Dans le tableau 2, nous donnons les résultats en terme

de taux d'erreur de mot (WER) sur notre corpus de test. Dans la première colonne, nous établissons les différentes stations radios auxquelles appartiennent les fichiers de test. La seconde colonne montre le taux d'erreur de mot du système de base et dans les cinq dernières colonnes nous montrons les taux d'erreur de mots des cinq différents modèles compacts. La dernière ligne indique le pourcentage de réduction du nombre de paramètres par rapport à celui du système de base.

Radio broadcasts	Base	40	100	150	190	250
RTM (2h)	35.5	36.0	35.7	35.7	35.6	35.5
RFI (1h30mn)	25.7	26.5	26.2	25.8	25.7	25.3
INFO (2h)	25.8	28.3	27.7	27.6	27.1	27.1
CLASSIQUE (1h)	21.7	22.6	22.1	22.2	21.8	22.1
CULTURE (1h)	34.0	36.8	36.0	35.6	35.3	35.1
Total (7h30)	28.8%	30.4 %	29.9%	29.7 %	29.5 %	29.4 %
réduction de la tail de modèle	-	92.1%	83.8%	76.9 %	70.9 %	63.0 %

Table 2: *Taux d'erreur de mot en % pour les différentes radios.*

Le tableau 2 montre que pour les fichiers tests appartienant aux radios RTM et RFI, nous gardons quasiment la performance de base, alors que nous avons une perte négligeable pour les autres (entre 1,6% et 0,6% de perte relative globale). Cette perte négligeable, malgré la réduction très significative du nombre de paramètres à estimer, montre la capacité du SGM modifié de réduire la taille des modèles tout en maintenant presque la même performance que la modélisation standard.

5. Conclusion

Dans cet article, nous avons exploré la capacité du Subspace Gaussian Mixture afin de réduire le nombre de paramètres du modèle acoustique. Nous proposons également une simplification concernant l'estimation des poids spécifiques à l'état : nous utilisons une estimation simple d'algorithme EM. Dans nos expériences, nous avons montré 92% de réduction relative de la taille des modèles, tout en maintenant une performance similaire (perte relative entre 1,6% et 0,6%). Avec ce gain en termes de taille du modèle, nous pouvons également améliorer le temps de calcul, car le calcul de probabilité pour chaque trame de test peut être limité à un petit nombre de gaussiennes sélectionnées à partir de l'UBM.

Références

[1] Christophe Lévy, Georges Linarès and Jean françois Bonastre, *"Compact acoustic models for embedded speech recognition"*, EURASIP Journal on Audio, Speech, and Music Processing,Vol9, pp. 806-186, 2009.

[2] Christophe Lévy, Georges Linarès, Pascal Nocera, and J-F Bonastre, *"Reducing computational and memory cost for cellular phone embedded speech recognition system*, In IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, Montrál, Canada, Vol. 5, pp. 309-312, 2004.

[3] J.Duchateau, K. Demuynck, D. Van Compernolle, *"A Novel Node Splitting Criterion in Decision Tree Construction for Semi-Continuous HMMs"*, Eurospeech'97, Rhodes, pp. 183-1186, 1997.

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

[5] J.-C. Junqua, P. Nguyen, and N. Niedzielski, *"Rapid Speaker Adaptation in Eigenvoice Space"*, IEEE Transactions on Speech and Audio, Speech and Audio Processingn vol. 8, no. 6,Nov.2000.

[6] M. J. F.Gales, *"Multiple-cluster adaptive training schemes,"*, in ICASSP,2001.

[7] L. R. Rabiner and B. H. Juang, *"An introduction to hidden Markov models"*. IEEE Acoustics, Speech & Signal Processing Magazine, pp. 4Ű16, 1986.

[8] S. J. Young, J. J. Odell, and P. C. Woodland, *"Tree-based state tying for high accuracy acoustic modeling"*, Proc. ARPA Human Language Technology Workshop, pp. 307-312, 1994.

[9] Driss Matrouf, Nicolas Scheffer, Benoit Fauve, Jean-Franois Bonastre, *"A Straightforward and Efficient Implementation of the Factor Analysis Model for Speaker Verification"*, INTERSPEECH Conference 2007, Antewerp, Belgium, pp. 1242-1245, 2007.

[10] D. Povey, et al., *"Subspace Gaussian mixture models for speech recognition"*, In : Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '10), March 14-19, 2010, Dallas, Texas, U.S.A, pp.1144-1160 , 2010.

[11] D. Povey, et al., *"A Tutoriel Introduction to Subspace Gaussien Mixture Models for Speech Recognition"*, Tech. Rep. MSR-TR-2009-111, Microsoft Research,2009.

[12] P. Nocera, C. Fredouille, G. Linares, D. Matrouf, S. Meignier, J.-F. Bonastre, D. Massonié, and F. Béchet, *"The LIA's french broadcast news transcription system"*, in SWIM : Lectures by Masters in Speech Processing, 2004.