Représentation acoustique compacte pour un système de reconnaissance de la parole embarquée

Christophe Lévy, Georges Linarès, Jean-François Bonastre

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

ABSTRACT

Speech recognition applications are known to require a significant amount of resources (training data, memory, computing power). However, the targeted context of this work -mobile phone embedded speech recognition system- only authorizes few resources. In order to fit the resource constraints, an approach based on a HMM system using a GMM-based state-independent acoustic modeling is proposed in this paper. A transformation is computed and applied to the global GMM in order to obtain each of the HMM state-dependent probability density functions.

The proposed approach is evaluated on a French digit recognition task. Our method leads a Digit Error Rate (DER) of 2%, when the system respects the resource constraints. Compared to an HMM with comparable resource, our approach achieved a DER relative decrease more than 50%.

1. INTRODUCTION

Le marché de la téléphonie mobile n'a cessé de croître depuis le milieu des années 90. D'après l'ARCEP[1] au troisième trimestre 2005 ce marché atteignait les 46 millions d'abonnements (plus de 73% de la population française - enfants compris - possèderait un téléphone).

Les téléphones de dernière génération offrent de nouvelles fonctionnalités : gestion des rendez-vous, outils de bureautique, jeux,... Certains modèles résultant de la fusion entre un téléphone et un organiseur numérique (PDA) offrent même l'ensemble des fonctionnalités de ces deux appareils. Ajouté à cela, une course à la miniaturisation a ammnené à des outils (téléphone, PDA, PDAphone, ...) de taille très réduite. L'intégration de la Reconnaissance Automatique de la Parole (RAP) dans ces systèmes devient un enjeu majeur en terme d'ergonomie.

Les performances obtenues par les systèmes de RAP actuels permettent d'envisager des applications réelles. Ces systèmes sont généralement basés sur une modélisation stochastique d'unités acoustiques (mots, phonèmes, diphones, tri-phones, ...). Cette approche probabiliste nécessite beaucoup de données pour l'apprentissage des modèles acoustiques, un espace de stockage conséquent pour les modèles et une puissance de calcul non négligeable pour la reconnaissance. Ces points deviennent cruciaux dès lors que le moteur de RAP est intégré dans un système embarqué, ces derniers ne possèdent que très peu de ressources (mémoire et/ou puissance de calcul).

[1] Autorité de Régulation des Communications Electroniques et des Postes - `http://www.art-telecom.fr/`

Dans ce papier, nous essayons d'apporter une réponse à ces problèmes. Pour cela, nous proposons une appoche basée sur un HMM (Hidden Markov Model). Notre proposition consiste à représenter l'espace acoustique dans un seul modèle générique (un GMM, Gaussian Mixture Model) puis à dériver les modèles des différents phonèmes depuis ce modèle générique, en mutualisant une large partie des modèles (*cf.* figure 1). Une première ébauche de cette approche a été présentée dans [6], une version détaillée peut être trouvée dans [7].

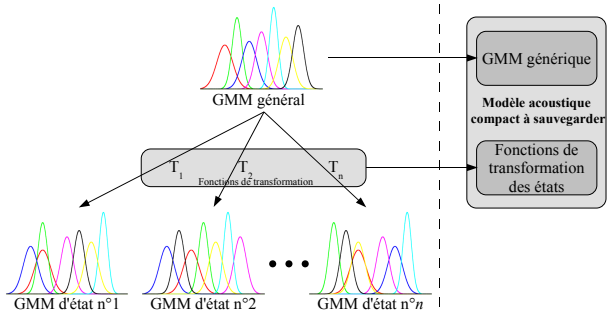

FIG. 1: principe général de l'approche proposée.

Dans la littérature différentes méthodes pour intégrer un moteur de RAP dans des systèmes embarqués sont proposées [1], [10], [9]. Les deux dernières travaux cités précédemment réduisent *a posteriori* la taille des modèles en mutualisant les paramètres par une approche ascendante, qui part d'un ensemble de modèles spécifiques à chaque unité acoustique et les regroupe hiérarchiquement en partageant les paramètres les plus proches. Notre approche vise les mêmes objectifs, mais propose une démarche descendante qui modélise l'environnement acoustique par un unique modèle générique, puis dérive celui-ci pour obtenir les modèles correspondant à chaque unité acoustique. L'intérêt d'une approche descendante réside dans une meilleure représentation de l'espace acoustique global étant donné que le modèle générique est appris sur l'ensemble des données d'apprentissage.

Après une brève présentation des deux corpus utilisés (section 2), nous détaillons l'approche proposée dans la section 3. Les principaux résultats sont présentés dans la section 4. Pour finir, quelques conclusions et perspectives sont présentées dans la section 5.

2. CORPUS

Afin d'évaluer l'approche proposée dans ce papier, deux corpus ont été utilisés. Le premier, BREF120 ([5]), a servi uniquement à l'apprentissage du modèle générique. Le se-

cond corpus, BDSONS ([3]), permet de construire les modèles de chaque unité acoustique (ici les phonèmes). Il est également utilisé pour les expériences de reconnaissance.

2.1. BREF120

Le corpus BREF120 est composé de parole lue issue du journal français *Le Monde*. Il est constitué d'une centaine d'heures de parole enregistrées par 120 personnes (55 hommes et 65 femmes).

Ce corpus a servi à l'apprentissage du modèle acoustique générique. Ce modèle a ensuite été adapté au corpus de test, en utilisantle sous-ensemble ADAPT_SET du corpus BDSONS (*cf.* la section 2.2).

2.2. BDSONS

Le corpus BDSONS est composé de phrases phonétiquement équilibrées, de suites CVC, de logatomes, de chiffres, ... Ces séquences ont été prononcées par 32 locuteurs (16 hommes et 16 femmes).

Seul le sous-corpus contenant les chiffres isolés a été retenu (composé de 15 hommes et 15 femmes). Ce sous-corpus a été divisé en deux sous-ensembles :
– un pour l'adapation (ADAPT_SET). Il contient 700 occurences de chiffre prononcées par 7 locuteurs (4 hommes et 3 femmes). Ce sous-ensemble a été utilisé pour adapter le modèle acoustique générique. Il a également été utilisé pour construire les modèles correspondant aux phonèmes.
– un pour l'évaluation (TEST_SET). Il est composé de 2300 occurences de chiffre prononcées par 23 locuteurs (11 hommes et 12 femmes).

Les locuteurs du corpus BDSONS sont différents de ceux du corpus BREF. Les locuteurs des corpus ADAPT_SET et TEST_SET sont aussi différents. Les tests sont donc réalisés en mode indépendant du locuteur, les locuteurs de test n'apparaissant jamais durant les phases d'apprentissages.

L'objectif de notre travail est de proposer un système de RAP embarqué généraliste, capable de reconnaitre des commandes vocales, des noms propres ... Nous utilisons donc une modélisation en unités acoustiques plutôt qu'une représentation par mots.

Cependant, le corpus de test étant limité à des chiffres isolés, les résultats sont exprimés en DER (Digit Error Rate - taux d'erreur de reconnaissance de chiffre).

3. APPROCHE BASÉE SUR UN MODÈLE GÉNÉRIQUE

L'approche proposée dans ce papier consiste à apprendre un modèle générique avec un ensemble suffisant de données puis à dériver ce modèle pour chaque unité acoustique, à l'aide d'une transformation basique dont seuls les paramètres seront stockés pour caractériser le modèle d'un phonème donné. Les modèles utilisés (modèle générique et modèles des états des HMMs) sont des GMM. La transformation utilisée consiste en une adaptation MAP (Maximum *A Posteriori* - [4]) sur les poids suivie d'une sélection des *NBest* gaussiennes (seulement les *NBest* gaussiennes sont sauvegardées). Cette transformation est appelée WRE (Weight Re-Estimation) dans ce papier. Elle peut être précédée d'une adaptation linéaire globale du GMM (avant l'adaptation des poids). Durant cette phase, la transformation est la même pour toutes les gaussiennes. Cette phase d'adaptation est nommée ULT (Unique Linear Transformation) dans la suite de ce papier.

3.1. Ré-estimation des poids : WRE

Cette étape consiste à adapter les poids du GMM générique afin d'obtenir le GMM dépendant de l'état. Les états sont donc différenciés entre eux uniquement par le vecteur de poids du GMM.

Pour la ré-estimation des poids, deux méthodes sont proposées :
– la première basée sur le critère de maximum de vraisemblance (MLE),
– la seconde basée sur une approche discriminante (MMIE).

Cette étape ne nécessite que très peu de ressource. En effet, pour stocker le modèle acoustique il suffit de stocker le GMM générique et un vecteur de poids. De plus, Tous les poids n'ont pas été conservés, seuls les poids correspondant aux $NBest$ gaussiennes ont été conservés, de manière à réduire l'espace nécessaire au stockage des modèles acoustiques. Pour le calcul de la vraisemblance, il suffit de calculer la vraisemblance de chaque gaussienne du GMM générique puis d'effectuer une somme pondérée en fonction du vecteur de poids considéré.

• **MLE** :

Le poids de la $i^{ème}$ gaussienne (w_i) est ré-estimé suivant le critère de maximum de vraisemblance (le critère MAP pourrait être utilisé mais devant la faible quantité de données disponibles, nous avons préféré utiliser un critère MLE). La fonction de ré-estimation utilisée est donc :

$$w'_i = \frac{w_i * Vrais(tr|g_i)}{\sum_{g_j=1}^{nb_g} w_j * Vrais(tr|g_j)} \quad (1)$$

où $Vrais(tr|g_x)$ correspond à la vraisemblance de données relatives à l'état (tr) pour la gaussienne g_x.

• **MMIE** :

L'apprentissage des HMM en utilisant un critère discriminant maximisant l'information mutuelle (Maximum Mutual Information Estimation - MMIE) a déjà été largement étudié, notamment dans [2].

Dans [7], une technique rapide pour ré-estimer les poids suivant un critère MMIE a été présentée. La règle de mise à jours des poids proposée est :

$$w'_{jm} = w_{jm} * \frac{w_{jm}}{\sum_k w_{km}}$$

où w_{jm} représente le poids de la $j^{ème}$ gaussienne de l'état m.

3.2. *Transformation linéaire : ULT*

La méthode LIAMAP présentée dans [8] permet d'adapter globalement un GMM en utilisant uniquement une transformation simple. Cette transformation nous permet d'adapter le GMM générique avec les données propres à chaque état. La transformation porte sur les paramètres de moyenne et de variance. La forme générale de cette transformation est donnée ci-dessous :

$$\mu_{GMMetat} = \alpha * \mu_{GMMgnral} + \beta$$
$$\Sigma_{GMMetat} = \alpha^2 * \Sigma_{GMMgnral}$$

avec α (commun pour $\mu_{GMMetat}$ et $\Sigma_{GMMetat}$) et β définis ci-après.

L'idée principale de cette adaptation (*cf.* figure 2) est d'estimer une transformation entre les gaussiennes (μ, Σ) et $(\tilde{\mu}, \tilde{\Sigma})$ obtenues :

1. en fusionnant toutes les gaussiennes du GMM générique afin d'obtenir (μ, Σ).
2. en adaptant (avec MAP) le GMM générique avec les données spécifiques de l'état et en fusionnant ensuite toutes les gaussiennes de ce nouveau GMM afin d'obtenir $\widetilde{\mu}$ and $\widetilde{\Sigma}$.

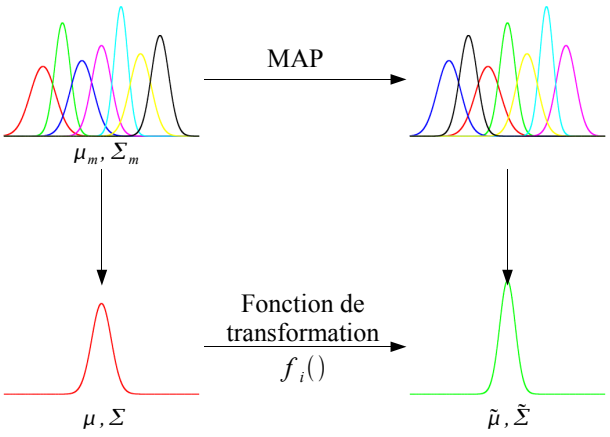

FIG. 2: LIAMAP : principe général de la transformation LIAMAP.

Chaque gaussienne finale (définie par sa moyenne μ'_m et sa matrice de covariance Σ'_m) est calculée de la manière suivante :

$$\mu'_m = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2}(\mu_m - \mu) + \widetilde{\mu} \quad (2)$$
$$\Sigma'_m = \widetilde{\Sigma}\Sigma^{-1}\Sigma_m \quad (3)$$

l'équation 2 peut-être développée en :

$$\mu'_m = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu_m - \widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu + \widetilde{\mu} \quad (4)$$

posons

$$\alpha = \widetilde{\Sigma}^{1/2}\Sigma^{-1/2} \quad (5)$$

et

$$\beta = -\widetilde{\Sigma}^{1/2}\Sigma^{-1/2}\mu + \widetilde{\mu} \quad (6)$$

les équations 2 et 3 deviennent :

$$\mu'_m = \alpha\mu_m + \beta \quad (7)$$

et

$$\Sigma'_m = \alpha^2\Sigma_m \quad (8)$$

Les équations 7 et 8 montrent d'une simple d'une transformation linéaire définie par les vecteurs α et β (la transformation est partagée par l'ensemble des gaussiennes du GMM).

ULT est ici présentée comme une première étape (optionelle) avant la ré-estimation des poids. L'étape suivante (WRE, *cf.* 3.1) toujours appliquée. La figure 3 illustre le processus complet.

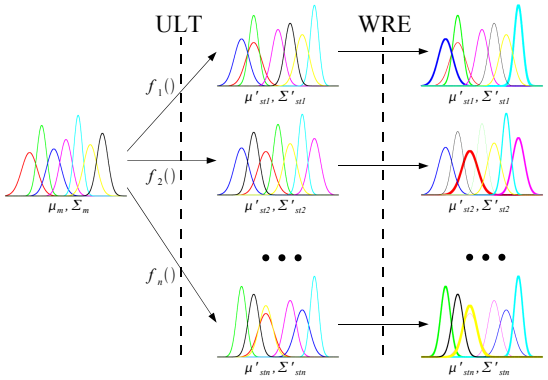

FIG. 3: Transformation permettant de passer du GMM général, indépendant des états, au GMM dépendant des états.

Durant le test, cette étape nécessite le calcul du GMM dépendant de l'état avant de pouvoir calculer la vraisemblance d'un état. Elle impose aussi de calculer la vraisemblance de chaque gaussienne de chaque état contrairement à une approche WRE simple (sans ULT préalable).

3.3. *Evaluation de l'occupation mémoire*

Nous avons fixé deux limites d'occupation mémoire proches des capacités réelles disponibles dans un téléphone portable.

Pour chaque approche - WRE et ULT+WRE - nous avons estimé la taille des modèles acoustiques en terme de nombre de paramètres. Ceci permet de fixer le nombre de gaussiennes pour chacune des trois approches afin que chaque approche nécessite la même quantité de mémoire.

4. RÉSULTATS

Les expériences ont été effectuées sur une tâche de reconnaissance de chiffre isolé avec le corpus BDSONS. A titre de comparaison, nous donnons aussi les résultats par un système HMM classique respectant les contraintes mémoire que nous nous sommes fixées.

Le tableau 1 résume l'ensemble des expériences effectuées. Il présente les performances en terme de taux d'erreur de reconnaissance (Digit Error Rate - DER) de chaque méthode.

Au regard de ce tableau, l'approche présentée dans ce papier (ULT+WRE ou WRE uniquement) permet une diminution importante du DER. La réduction, relative, du DER varie entre 9% (WRE/MMIE limite mémoire inférieure) et 53% (WRE/MMIE limite mémoire supérieure). Le DER atteint 2,09% avec notre meilleure approche alors qu'il se situe aux alentours de 4,43% pour l'approche HMM classique.

TAB. 1: DER pour les approches WRE et ULT+WRE (2700 tests).

Model size	HMM	WRE/MLE	WRE/MMIE	ULT+WRE/MLE	ULT+WRE/MMIE
6k	4,96%	4,17%	4,52%	3,39%	3,22%
11k	4,43%	3,09%	2,09%	3,00%	2,70%

Le gain, en terme de DER, apporté par l'étape ULT est à nuancer au regard du surcoût de calcul qu'impose cette étape. En effet, l'approche WRE est très économe en puissance de calcul, elle permet de calculer l'ensemble des vraisemblances pour une observation donnée (une vraisemblance par état des machines phonétiques) en ne calculant qu'une seule fois la vraisemblance des gaussiennes du modèle générique. Les vraisemblances pour chaque état correspondent simplement à une somme pondérée des vraisemblances calculées avec le modèle générique. A contrario, l'approche ULT+WRE nécessite de calculer le GMM de l'état puis les vraisemblances de chacune des composantes avant d'obtenir le résultat par une somme pondrée de ces dernières vraisemblances. Il convient donc de choisir entre ULT+WRE et WRE en fonction des ressources (d'un point de vue puissance de calcul) disponibles.

5. CONCLUSION ET PERSPECTIVES

Dans ce papier, nous avons présenté une approche permettant d'intégrer un système de reconnaissance automatique de la parole dans un système embarqué (téléphone portable, PDA, ...).

La technique présentée dans ce papier est basée sur le formalisme des HMM (classiquement utilisé en RAP). Nous proposons une optimisation des HMM en terme de ressources mémoire et de calcul qui permet d'atteindre un niveau de performance intéressant dans le contexte applicatif des systèmes embarqués.

Les résultas présentés dans ce papier montrent que notre approche permet une réduction relative des taux d'erreur variant entre 9% (WRE/MMIE, petit modèle) et 53% (WRE/MLE, grand modèle) comparée à un modèle HMM classique. Par exemple, notre méthode permet un taux d'erreur à 2,09% pour les plus gros modèles, à comparer au résultat du HMM classique, 4,43% (pour comparaison, le système HMM classique obtient un taux d'erreur avoisinant les 1% avec un modèle acoustique composé de plus d'un million de paramètres).

Dans l'étude présentée aucune adaptation, au locuteur ou à l'environnement n'a été réalisée. Une suite logique de ce travail serait d'envisager différentes formes d'adaptation. Une approche prometteuse consiste à réaliser cette adaptation au niveau du GMM générique. En effet, en se basant sur les méthodes proposées en reconnaissance du locuteur, quelques trames peuvent suffir à adapter ce modèle au locuteur (ou à l'environnement), indépendamment des mots prononcés. Dans le cadre de l'adaptation ULT, un second niveau d'adaptation pourrait être envisagé, agissant au niveau de chaque état.

RÉFÉRENCES

[1] S. Astrov, J.G. Bauer, and S. Stan. High performance speaker and vocabulary independent asr technology for mobile phones. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'2003)*, pages 281–284, Hong Kong, April 2003.

[2] 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.

[3] 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.

[4] J.L. Gauvain and C.H. Lee. Maximum A Posteriori estimation for multivariate gaussian mixture observations of markov chains. In *IEEE Transactions on Speech and Audio Processing*, volume 2-2, pages 291–298, April 1994.

[5] 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.

[6] C. Lévy, G. Linarès, and J.F. Bonastre. Mobile phone embedded digit-recognition. In *Workshop on DSP in Mobile and Vehicular Systems*, Sesimbra, Portugal, September 2005.

[7] 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*.

[8] D. Matrouf, O. Bellot, P. Nocera, Linarès, and J. F. Bonastre. Structural linear model-space transformations for speaker adaptation. In *Proceedings of European Conference on Speech Communication and Technology (Eurospeech'2003)*, pages 1625–1628, Geneva, Switzerland, September 2003.

[9] J. Park and H. Ko. Compact acoustic model for embedded implementation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP'2004)*, pages 693–696, Jeju Island, Korea, October 2004.

[10] S.J. Young. The general use of tying in phoneme-based HMM speech recognisers. In *Proceedings of International Conference on Acoustics Speech and Signal Processing (ICASSP'1992)*, pages 569–572, San Francisco, California, USA, March 1992.