Spécialisation automatique de modèles acoustiques

Georges Linares, Serigne Gueye, Loïc Lefort, Philippe Michelon, Pascal Nocera

Laboratoire d'Informatique d'Avignon
LIA, Avignon, France
Tél.: ++33 (0)4 90 84 35 20 - Fax: ++33 (0)4 90 84 35 01
Mél: georges.linares@lia.univ-avignon.fr - http://www.llia.univ-avignon.fr

ABSTRACT

In this paper, we present a method for automatic generation of acoustic models from simple generic models. This method use the internal structure of non-contextual acoustic models in order to build new specialized states which are supposed to modelize specific patterns of a phoneme. The proposed technique use temporal information for state splitting. This method is compared to a maximum likelihood based approach. Our experiments show that this last criterion leads to better performance. Nevertheless, unsupervised model splitting seems to be less efficient than model specialization based on a priori knowledge.

1. INTRODUCTION

Les modèles acoustiques utilisés dans les systèmes de reconnaissance "état de l'art" sont d'une grande complexité, de l'ordre de plusieurs millions de paramètres. Ils modélisent des phonèmes contextuels qui peuvent partager des paramètres à différents niveaux (gaussiennes, états, transitions, etc..). Ce partage permet de réduire la complexité globale des modèles et de résoudre les éventuels problèmes d'estimation liés à la taille ou à la représentativité des corpus d'apprentissage ([You92], [Hwa93]). Il est généralement réalisé au niveau des GMM, les états "logiques" des modèles contextuels étant regroupés en un nombre réduit d'états "physiques" effectivement estimés. Différentes approches ont étés évaluées pour déterminer la classification optimale des états ; la majorité des systèmes actuels utilisent des arbres de décision dont les feuilles sont des états partagés et les nœuds des questions relatives au contexte linguistique du phonème modélisé. Le jeu de questions utilisé est évidement assez déterminant pour la qualité des modèles obtenus, mais aussi pour la taille finale des modèles. De nombreuses équipes ont adopté une approche mixant des questions linguistiques formulées par des experts et des questions générées automatiquement à partir d'un critère d'entropie ou de perte de vraisemblance ([Beu98]).

Ce type de méthodes permet d'obtenir des résultats très concluants en terme de taux de reconnaissance et de réduction de la complexité des modèles, dès lors que les ressources disponibles (corpus, mémoire, processeur, etc.) sont suffisantes. Les systèmes de reconnaissance embarqués mettent à disposition des décodeurs des ressources souvent limitées, incompatibles avec la dimension des modèles utilisés dans les systèmes de laboratoire.

Dans ce papier, nous évaluons une approche non-supervisée susceptible d'améliorer la qualité des modèles non-contextuels au prix d'une très faible augmentation de leur complexité. Le principe de cette méthode est d'utiliser les probabilités de transitions entre les composantes gaussiennes des modèles pour adapter leur structure aux données. Ce mécanisme d'extension automatique des modèles est détaillé dans la première partie de cet article, puis comparé à une méthode de subdivision plus classique, basée sur un critère de maximum de vraisemblance.

2. SUBDIVISION HIÉRARCHIQUE DE MODÈLES

2.1 Principe

Les modèles standards sont entraînés sur l'ensemble des exemples du corpus d'apprentissage, indépendamment du contexte phonétique, du locuteur ou des différentes sources de variabilités (canal de transmission, conditions d'acquisition, etc.). La modélisation de ces formes très variables conduit à une certaine généralité des modèles et nécessite l'utilisation d'un grand nombre de gaussiennes, dont certaines vont représenter des modes particuliers (locuteur masculin/féminin, contexte linguistique, etc.). Le principe de la méthode que nous proposons est d'isoler ces différents modes par l'étude de la structure interne des modèles, puis de créer des modèles spécialisés à partir du modèle générique initial.

2.2 Bipartition de modèles

Notre objectif est d'extraire automatiquement des types de réalisations distincts. La méthode que nous proposons repose sur l'hypothèse que les modèles multi-gaussiens qui composent le MMC générique résultent du mélange de modèles mono-gaussiens à priori inconnus. Pour chercher ces modèles à partir du modèle générique appris, nous commençons par le décomposer en un graphe ergodique dont chaque nœud correspond à une gaussienne du modèle initial. Nous cherchons ensuite une subdivision optimale de l'ensemble de ces composantes gaussiennes. Cette subdivision nous permet enfin de recomposer un modèle classique, composé d'états multi-gaussiens partiellement connectés. L'algorithme de spécialisation comporte donc 4 étapes principales qui correspondent à ce mécanisme général de décomposition et recomposition.

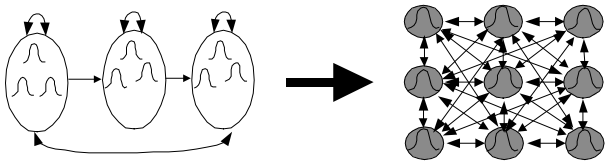

Figure 1: décomposition d'un modèle générique à 3 états et *n* gaussiennes en un modèle ergodique à *n* états mono-gaussiens

Décomposition du modèle :

Les états du modèle initial sont décomposés en un ensemble d'états mono-gaussiens interconnectés (fig. 1) ; cet éclatement des mixtures de gaussiennes aboutit à la formation de modèles ergodiques composés d'un grand nombre d'états auxquels sont associés les distributions gaussiennes extraites des mixtures du modèle générique. Toutes les transitions entre états sont possibles, avec des valeurs initiales toutes égales. On suppose donc, dans un premier temps, que les transitions sont équiprobables. Cette initialisation permet ensuite un décodage acoustico-phonétique de l'ensemble de la base, à partir duquel les probabilités de transition sont à nouveau estimées. Ce procédé nécessite donc un décodage complet du corpus d'apprentissage.

Bipartition du graphe d'états :

Dans le MMC décomposé, le décodage d'une séquence d'observations produit la séquence d'états parcourus S_t maximisant la vraisemblance des observations. On appellera *trajectoire* cette séquence optimale. On identifie des modes en détectant des trajectoires caractéristiques dans le MMC. Nous procédons de façon indirecte, en cherchant la bipartition du modèle qui minimise la probabilité qu'une trajectoire soit coupée, ce qui revient à rechercher la bipartition de coût minimal d'un graphe dont chaque nœuds représente un état et chaque lien la probabilité de transition d'un état à un autre. La seconde étape consiste donc à partitionner le modèle de façon à minimiser la somme des valeurs des transitions coupées. La recherche de cette coupe optimale est faite pour chaque cardinalité par une heuristique ([Gue99]). La solution retenue est celle correspondant à la cardinalité pour laquelle la coupe minimale est la meilleure.

On peut noter que le coût de l'algorithme de bipartition est négligeable par rapport à celui du décodage pour l'estimation des probabilités de transition, ainsi qu'à celui de la recomposition des modèles.

A l'issue de cette subdivision du modèle générique, on obtient deux modèles de structures similaires. Cependant, les transitions du modèle ergodique initial sont préservées ; la valeur des transitions sortantes d'un état ne sont plus des probabilités. L'étape suivante consiste à identifier, dans le corpus, les exemples qui sont émis par l'un ou l'autre des modèles générés.

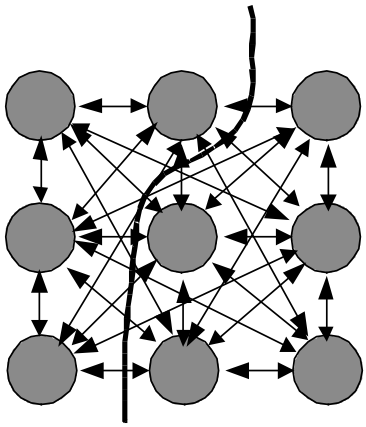

Figure 2: bipartition du graphe d'états par recherche de la coupe minimisant la somme des log-probabilités des transitions coupées.

Bipartition du corpus :

La séparation du graphe en deux sous graphes est utilisée, lors de cette étape, pour étiqueter les exemples du corpus d'apprentissage suivant les deux modes concurrents isolés par la bipartition du modèle. Chaque exemple du corpus est affecté au modèle pour lequel sa vraisemblance est maximale, ce qui nécessite un nouveau décodage du corpus par les deux sous modèles générés.

Ici, le processus de subdivision des modèles peut être vu comme une méthode de classification non-supervisée des exemples.

- réestimation des modèles : la bipartition du corpus permet une réestimation de deux sous-modèles, le sous modèle de plus faible vraisemblance étant à son tour « splitté » suivant le même processus. La subdivision des modèles se poursuit ainsi jusqu'à ce qu'une vraisemblance minimale ou un nombre de modèles maximal soient atteints. A l'issue de cette recomposition des MMC, un réalignement du corpus et réalisé et les modèles sont réestimés.

2.3 Décodage

Les différents modèles spécialisés dérivés d'un même modèle générique sont mis en parallèle (cf. figure 3). Cela signifie que les modèles générés sont mis en concurrence directe lors du décodage, contrairement à ce qui est fait lorsque les modèles spécialisés sont associés à des contextes linguistiques facilement identifiables et exploitables par le moteur de reconnaissance. De plus, il n'y a pas de liens entre les états correspondant à des modes concurrents, ce qui est cohérent avec notre choix initial de classifier au niveau des exemples plutôt qu'au niveau des trames. Un segment est donc décodé par l'un ou l'autre des modèles dérivés, mais jamais partiellement par l'un et partiellement par l'autre.

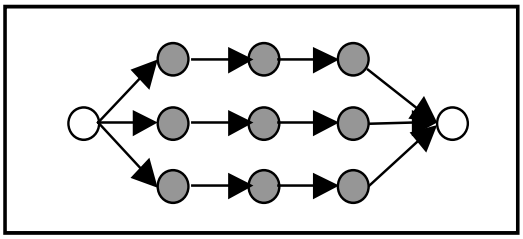

Figure 3 : les états générés par la bipartition sont mis en parallèle dans le MMC modélisant un phonème.

2.4 Expériences

Pour évaluer la méthode que nous proposons, nous avons utilisé *Speeral*, le système de reconnaissance de la parole développé au LIA, sur le corpus BREF120. L'ensemble de test est constitué des 300 phrases de test proposé pour la tâche ARC-B1 de la campagne d'évaluation AUPELF.

Les modèles initiaux que nous avons utilisés sont des modèles non-contextuels classiques ; ils sont composés de 38 MMC à 3 états émetteurs de 32 gaussiennes. Ces modèles contiennent en tout environ 280 000 paramètres. Nous avons appliqué notre algorithme de bipartition de modèle à ces 38 modèles génériques, seuls les états centraux étant subdivisés en 2 puis 4 états concurrents (fig. 4).

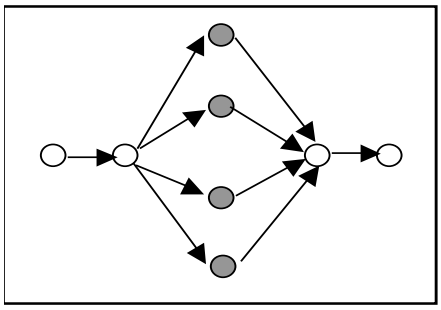

Figure 4 : topologie des modèles évalués : l'état central est subdivisé en plusieurs états concurrents, connectés au premier et au dernier état émetteur du modèle initial.

Nous évaluons donc une version simplifiée de l'algorithme, dans laquelle seul l'état central est décomposé et subdivisé. On obtient au total des modèles à 360 000 et 520 000 paramètres.

La table 1 montre les résultats obtenus en terme de taux de reconnaissance et de complexité (notée *CPX*), et permet leur comparaison avec des modèles classiques.

Table 1 : résultats des taux de reconnaissance comparés pour des modèles non-contextuels de 32 (NCTX 32) , 42 (NCTX 42), 64 (NCTX 64) gaussiennes par état , des modèles obtenus par subdivision de l'état central en 2 états (Bipart2), puis en 4 états (Bipart4), des modèles spécifiques hommes/femmes (H/F 32).

Modèles	Rec.	Sub.	Del.	Ins.	Err.	Cpx
NCTX32	76.7	17.3	6.1	2.3	25.6	280k
NCTX 42	76.7	17.5	5.8	2.1	25.3	320k
Bipart 2	77.O	17.5	5.5	2.4	25.4	320k
Bipart 4	77.1	17.3	5.5	2.3	25.2	560k
NCTX 64	78.3	16.3	5.3	2.3	24.0	560k
H/F 32	81.8	14.5	3.7	3.4	21.6	2x280k

Les résultats obtenus montrent une très légère diminution du taux d'erreur par rapport au NCTX32, de l'ordre de 0.2% pour le modèle Bipart2, de 0.4% pour le Bipart4.

Nous avons comparé ces performances avec celles de modèles classiques à 42 et 64 gaussiennes, de complexités identiques aux modèles à 2 et 4 états centraux. On voit que les résultats obtenus sont proches pour NCTX42 (+0.1%). Par contre, les modèles standards à 64 gaussiennes obtiennent des résultats sensiblement meilleurs (24.0%) ; enfin, l'utilisation d'une connaissance à priori sur le sexe du locuteur permet un gain bien plus important, de l'ordre de 4%.

Pour conclure, on peut dire que le gain de performance apporté par la bipartition de modèles reste très limité, surtout comparé avec l'apport de modèles spécifiques hommes/femmes. Deux aspects de notre algorithme peuvent expliquer ces résultats modestes ; d'une part, l'utilisation d'une stratégie non supervisée, dont le principe est d'utiliser la structure sous-jacente des données pour améliorer la qualité de la modélisation est probablement inférieure à des techniques supervisés ou l'ajout d'information supplémentaire (du type le locuteur est une femme ou un homme) contraint l'apprentissage et le décodage ; d'autre part, le critère de bipartition est fondé sur l'information portée par les probabilités de transitions des modèles « éclatés », paramètres dont la pertinence a été fréquemment discutée.

De façon à évaluer la qualité de ce critère, nous avons comparé cet algorithme de bipartition basé sur le minimum de probabilité de transitions coupées avec un autre basé sur un critère plus classique de maximum de vraisemblance.

3 CLASSIFICATION NON-SUPERVISÉE DES SEGMENTS.

3.1 Principe

L'algorithme de bipartition de modèles décrit précédemment peut être vu comme un algorithme de classification d'exemples : un modèle générique est subdivisé en deux sous-modèles ; le résultat cette division du modèle permet le bipartitionnement du corpus, qui servira à son tour à l'estimation de deux nouveaux modèles. Quelques méthodes de subdivision de modèles par classification à priori des exemples ont été publiées récemment ; [Rod97] propose un algorithme basé sur un critère de maximum de vraisemblance qui itère des bipartitions du corpus pour subdiviser des modèles sans utiliser de connaissance phonologique. L'algorithme de

subdivision s'interrompt lorsque un seuil minimal de quantité de données d'apprentissage est atteint. Evaluée en décodage acoustico-phonétique sur l'espagnol, cette technique permet de passer de 44.14% à 47.15% de taux de reconnaissance.

Nous avons mis en oeuvre une version simplifiée de cet algorithme. La bipartition est ici réalisée par l'application de l'algorithme des k-moyennes avec une distance et une fonction d'estimation des centres des classes particulières : le centre d'une classe est le modèle mono-gaussien générant les exemples de la classe, et la distance d'un segment au centre est l'opposé de la log-vraisemblance moyenne des trames du segment sachant la gaussienne centre. Le centre de la classe étant bien l'élément maximisant la vraisemblance des segments, fonction d'estimation des centres et distance sont cohérentes et garantissent la convergence de l'algorithme.

Contrairement à l'algorithme de bipartition du graphe de gaussiennes présenté dans la première partie de cet article, le modèle générique n'est pas utilisé pour construire des sous modèles. De plus , le critère de segmentation du corpus est fondamentalement différent, basé sur la topologie « naturelle » du corpus plutôt que sur l'identification de trajectoires caractéristiques.

3.2 Expériences

Nous avons comparé cette méthode dans les conditions pour lesquelles ont a obtenu un rapport complexité/performance intéressant pour la bipartition, soit 2 états centraux par modèle. On obtient des résultats légèrement supérieurs (25.1% de taux d'erreurs), équivalents à ceux obtenus par le modèle Bipart4 légèrement plus complexe.

Modèles	Rec.	Sub.	Del.	Ins.	Err.	Cpx
KMEANS 2	77.1	17.0	5.6	2.5	25.1	360k

Ces résultats montrent que l'utilisation des probabilités de transition ne permet d'aboutir à des résultats significativement meilleurs que ceux obtenus par une approche de type maximum de vraisemblance.

4. CONCLUSION

Les deux méthodes de subdivision d'états que nous avons évaluées sont basées sur des critères fondamentalement différents ; les probabilités de transitions entre gaussiennes des modèles génériques codent une information temporelle, tandis que la méthode de bipartition du corpus par l'algorithme des k-moyennes regroupe les exemples en minimisant la variance des classes. Faibles dans l'absolu, les performances obtenues par la méthode de bipartition sont légèrement inférieures à celles basées sur des approches classiques, qu'il s'agisse de l'algorithme de classification non-supervisé de segments ou bien de l'augmentation du nombre de paramètres des modèles estimés par l'algorithme EM appliqué sur l'ensemble du corpus. Globalement, l'utilisation de connaissance à priori lors de l'apprentissage et le décodage procure un gain nettement plus important. Observée ici sur des modèles de faible complexité, on peut penser que cette supériorité de l'approche supervisée sur les techniques que nous décrivons serait préservée sur des modèles plus complexes. Enfin, les performances réalisées par ces modèles de faible complexité restent très inférieures à celles obtenues par notre moteur de reconnaissance utilisant des modèles contextuels adaptés au locuteurs, pour lesquels ont obtient un taux d'erreur de 17.2% dans les mêmes conditions d'évaluation.

Dans l'avenir, nous envisageons de compléter l'évaluation de cette méthode en la mettant en oeuvre sur l'ensemble composantes gaussiennes des modèles, et non pas seulement sur l'état central. Cette décomposition plus radicale des MMC ainsi que l'augmentation de la durée des segments utilisés pourraient améliorer l'estimation des probabilités de transition et, d'une façon plus générale, les performances de l'algorithme.

BIBLIOGRAPHIE

[You92] Young S.J. & Al, "The general use of tying in phoneme based HMM speech recognizers", ICASSP 92, pp 569-572

[Tak92] Takami J., Sagayama S.,"A successive splitting algorithm for efficient allophone modelling", ICASSP 92, pp573-576

[Hwa93] Hwang M.Y. "Subphonetic Acoustic Modelling for Speaker-Independant Continuous Speech Recognition System", PhD Thesis, CMU-CS-93-230, Carnegie Mellon University, 1993

[Rod97] Rodriguez L.J., Torres M.I, "Viterbi based splitting of phoneme HMMs", EuroSpeech 97, Rhodes, Greece

[Beu98] Beulen K., Ney H., "Automatic question generation for decision tree based state tying", ICASSP 98, Vol .2, pp 805-809, 1998

[Gue99] Gueye S., Mautor T., Michelon P. (1999), « Some numerical experiments on the graph bipartionning problem », IFORS 99, Bejing, Chine