Modélisation de trajectoires et de classes de locuteurs pour la reconnaissance de voix d'enfants et d'adultes

Arseniy Gorin, Denis Jouvet

Modélisation de trajectoires et de classes de locuteurs pour la reconnaissance de voix d'enfants et d'adultes

Arseniy Gorin[1,2,3] Denis Jouvet[1,2,3]
Equipe Parole, LORIA
(1) Inria, 615 rue du Jardin Botanique, F-54600, Villers-lès-Nancy, France
(2) Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
(3) CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
{arseniy.gorin, denis.jouvet}@inria.fr

RÉSUMÉ

Lorsque l'on considère de la parole produite par des enfants et des adultes, la variabilité acoustique de chaque unité phonétique devient grande, ce qui dégrade les performances de reconnaissance. Un moyen d'aller au-delà des modèles de Markov traditionnels, est de prendre en considération des classes de locuteurs. Les classes de locuteurs peuvent être obtenues automatiquement. Elles servent à fabriquer des modèles acoustiques spécifiques de chaque classe. Ce papier propose une structuration des composantes des densités multigaussiennes (GMMs) en relation avec des classes de locuteurs. Dans une première approche, cette structuration des densités est complétée par des pondérations des composantes gaussiennes dépendantes des classes de locuteurs, et dans une deuxième approche, par des matrices de transition entre les composantes gaussiennes des densités (comme dans les *stranded GMMs*). Ces deux approches apportent des gains substantiels pour la reconnaissance de voix d'enfants et d'adultes. La structuration des composantes gaussiennes complétée par des matrices de transition entre composantes réduit de plus d'un tiers le taux d'erreur mot sur le corpus TIDIGIT.

ABSTRACT

Explicit trajectories and speaker class modeling for child and adult speech recognition

When the speech data is produced by speakers of different age and gender, the acoustic variability of any given phonetic unit becomes large, which degrades speech recognition performance. One way to go beyond conventional Hidden Markov Model is to explicitly include speaker class information in the modeling. Speaker classes can be obtained automatically, and they are used for building speaker class-specific acoustic models. This paper introduces a structuring of the Gaussian components of the GMM densities with respect to the speaker classes. In a first approach, this structuring of the Gaussian components is completed with speaker class-dependent mixture weights, and in a second approach, with transition matrices, which add dependencies between Gaussian components of mixture densities (as in stranded GMMs). The two approaches bring substantial performance improvements when recognizing adult and child speech. Using class-structured components plus mixture transition matrices reduces by more than one third the word error rate on the TIDIGIT corpus.

MOTS-CLÉS : Reconnaissance de la parole ; classification non supervisée ; modèles de classes de locuteurs ; modèles stochastiques de trajectoire.

KEYWORDS: Speech recognition ; unsupervised clustering ; speaker class modeling ; stochastic trajectory modeling.

1 Introduction

Le traitement de la variabilité interlocuteur reste un problème pour les systèmes de reconnaissance automatique de la parole (RAP) ; elle est due entre autres au sexe du locuteur, à son âge et à son accent (Benzeghiba *et al.*, 2007; Stern et Morgan, 2012). La variabilité résultante des paramètres acoustiques doit être prise en compte par les modèles acoustiques indépendants du locuteur ; or les modèles de Markov cachés avec densités multigaussiennes (HMM-GMM : *Hidden Markov Model with Gaussian Mixture Model observation densities*) ne sont pas capables de représenter précisément des distributions très hétérogènes de paramètres acoustiques. Les techniques d'adaptation permettent d'améliorer la précision, en modifiant les paramètres acoustiques (par exemple, VTLN (Zhan et Waibel, 1997), fMLLR (Gales, 1998)) ou les paramètres des modèles (par exemple, MLLR, MAP (Gauvain et Lee, 1994)).

Ce papier traite de la situation générale où les données proviennent de locuteurs d'âge et de sexe différents, et dont la classe d'appartenance est inconnue pour le système de reconnaissance. Une classification automatique peut alors être appliquée sur chaque phrase, en supposant que le locuteur ne change pas au cours de la phrase (Beaufays *et al.*, 2010). La quantité de données disponibles pour l'apprentissage du modèle de chaque classe diminue lorsque le nombre de classes augmente, et cela peut rendre la modélisation moins fiable. Le manque de données peut être partiellement compensé par l'utilisation d'approches comme les voix propres, où les paramètres pour un locuteur sont déterminés comme une combinaison de modèles de classes (Kuhn *et al.*, 1998), ou en élargissant les classes de données pour l'apprentissage en autorisant une donnée d'apprentissage à appartenir à plusieurs classes (Jouvet *et al.*, 2012; Gorin et Jouvet, 2012).

Alors que dans les approches HMM (*Hidden Markov Model*) conventionnelles, les composantes des GMM (*Gaussian Mixture Models*) sont apprises indépendamment pour chaque densité, la structuration des composantes des densités en fonction de classes de locuteurs conduit à des GMM pour lesquels la $k^{ème}$ composante de chaque densité est associée à une même classe de données. La structuration en classes des HMM-GMM, résulte de l'initialisation des GMM à partir de GMM de plus faible dimension (i.e. ayant moins de composantes), et a été initialement proposée et étudiée pour la transcription d'émissions radio (Gorin et Jouvet, 2013). Pour cette modélisation HMM-SWGMM (*HMM with Speaker class-dependent Weights*), les composantes gaussiennes des mélanges GMM sont structurées en fonction des classes et partagées entre toutes les classes, tandis que les pondérations des composantes des densités sont dépendantes de la classe.

Dans ce papier, nous proposons de combiner la structuration des composantes des densités GMM en fonction des classes avec l'utilisation de matrices de transition entre composantes (MTM : *Mixture Transition Matrices*) comme dans les SGMM (*Stranded GMM*). Les SGMM sont similaires aux modèles Gaussiens conditionnels (Wellekens, 1987) qui ont été récemment étendus, reformulés et étudiés pour la reconnaissance robuste (Zhao et Juang, 2012). Dans les SGMM, les matrices de transition MTM définissent les dépendances entre les composantes gaussiennes des densités GMM adjacentes. Alors que les SGMM étaient originellement initialisés à partir de HMM-GMM conventionnels, la modélisation SSGMM (*class-Structured SGMM*) proposée dans ce papier combine SGMM et structuration des composantes en classes. Les matrices MTM modélisent alors la probabilité de rester sur une composante de la même classe au cours du temps, ou de passer vers une composante d'une autre classe. Cette modélisation explicite de trajectoires améliore la précision des modèles acoustiques. De plus, cette approche ne requiert pas d'étape additionnelle

de classification pour estimer la classe à laquelle appartient la phrase à décoder.

Le papier est organisé de la manière suivante. La section 2 présente le problème de la reconnaissance de voix d'enfants et d'adultes. La section 3 commente l'utilisation de modèles de classes. La section 4 introduit la structuration en classes pour les HMM-SWGMM. La section 5 rappelle le principe des SGMM et détaille la modélisation SSGMM proposée qui combine la structuration en classes des composantes et l'utilisation de matrices de transition entre composantes.

2 Système de référence et formulation du problème

Les expériences présentées dans ce papier ont été menées sur le corpus TIDIGIT (Leonard et Doddington, 1993) de chiffres connectés en anglais ; l'un des rares corpus disponibles à contenir à la fois des voix d'adultes et des voix d'enfants. L'ensemble d'apprentissage contient 41224 occurrences de chiffres (28329 prononcés par des adultes et 12895 par des enfants). L'ensemble de test contient 41087 occurrences de chiffres (28554 par des adultes et 12533 par des enfants). Les outils Sphinx3 (CMU, 2014) ont été utilisés, et enrichis pour traiter les SGMM (matrices de transition entre composantes - cf. Section 5). Les chiffres sont modélisés par des séquences de phones dépendants du mot. Chaque phonème est modélisé par un HMM à 3 états, sans saut. Chaque densité d'émission est une densité multigaussienne à 32 composantes. L'analyse acoutique calcule 13 coefficients (12 MFCC et le logarithme de l'énergie), auxquels on adjoint les dérivées premières et secondes. Le signal est sous-échantillonné à 8kHz, comme indiqué dans d'autres publications traitant ce corpus TIDIGITS (ex. (Burnett et Fanty, 1996)). Les taux d'erreur mot (WER : *Word Error Rate*) du système de référence sont indiqués dans la Table 1. Deux modèles indépendants du locuteur (SI : *Speaker Independent*) ont été utilisés, l'un appris sur les données provenant des adultes, et l'autre appris sur l'ensemble d'apprentissage complet (adultes et enfants). Les deux dernières lignes indiquent les performances obtenues avec des modèles adaptés par MLLR+MAP, en fonction de l'âge, ou en fonction du sexe et de l'âge.

	Adulte	Hom.	Fem.	**Enfant**	Garç.	Fille
Apprentissage sur données adultes	**0,64**	0,79	0,49	**9,92**	6,51	13,33
Apprentissage sur données adultes+enfants	**1,66**	1,86	1,46	**1,88**	1,69	2,08
+ adapt. âge (classe connue pour décodage)	**1,42**	1,56	1,28	**1,56**	1,52	1,54
+ adapt. sexe+âge (classe connue pour décodage)	**1,31**	1,57	1,04	**1,31**	1,14	1,49

TABLE 1: Taux d'erreur mot avec des HMM conventionnels pour les modèles indépendants du locuteur, et pour les modèles adaptés avec évaluation oracle (i.e. classe connue pour décodage)

L'apprentissage sur les données adulte seules fournit les meilleures performances pour les locuteurs adultes, mais de piètres performances pour les données des enfants. Quand des données d'enfants sont incluses dans le corpus d'apprentissage, les performances s'améliorent pour les enfants, mais se dégradent pour les adultes. L'utilisation de modèles adaptés, soit à l'âge du locuteur (i.e. adultes v.s. enfants) soit au sexe et à l'âge du locuteur (hommes / femmes / garçons / filles) améliore les performances. Dans la suite, l'apprentissage des modèles sera effectué en utilisant l'ensemble d'apprentissage complet (adultes + enfants), et l'information *a priori* (âge & sexe) ne sera pas utilisée (i.e. les classes de données seront estimées automatiquement).

3 Classification non supervisée et modélisation multiple

L'objectif d'une classification non supervisée est de regrouper automatiquement les données d'apprentissage en classes correspondant à des données acoustiquement similaires. Une approche de

classification automatique reposant sur des modèles multigaussiens a été présentée dans (Jouvet *et al.*, 2012). Un GMM générique est d'abord appris sur l'ensemble des données d'apprentissage. Puis, ce GMM est dupliqué, et les moyennes des gaussiennes légèrement modifiées. Ensuite les données d'apprentissage sont classifiées avec ces deux GMM, déterminant ainsi deux classes de données à partir desquelles les paramètres des deux GMM sont réestimés. Ces phases de classification et d'apprentissage sont itérées jusqu'à convergence. Ensuite, on itère si nécessaire à partir de l'étape de duplication des GMM, jusqu'à obtenir le nombre désiré de classes. Les modèles acoustiques indépendants du locuteur sont alors adaptés (MLLR+MAP) pour chaque classe en utilisant les données correspondantes.

Analyse de la classification automatique sur les données adultes et enfants avec des GMM ayant 256 composantes. La figure 1 présente la répartition des données d'apprentissage de chaque classe par rapport à l'âge et au sexe des locuteurs. Les deux premières classes séparent les données entre hommes d'un coté, et femmes et enfants de l'autre. Avec 4 classes, on voit apparaître une séparation entre les voix de femmes et les voix d'enfants. Par contre, il semble impossible de séparer les voix de garçons des voix de filles, même en augmentant le nombre de classes.

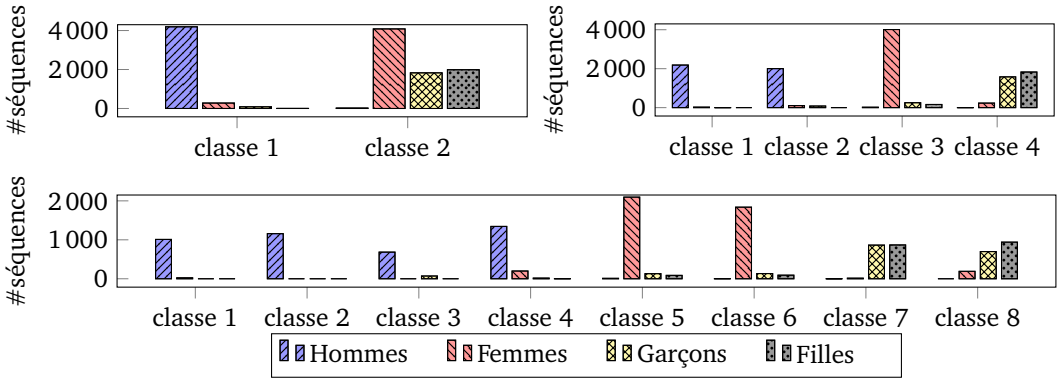

FIGURE 1: Nombre de séquences par catégorie de locuteurs en fonction du nombre de classes

Lors du décodage, chaque phrase est automatiquement classifiée et le décodage est effectué avec les modèles acoustiques correspondant à la classe identifiée. Les taux d'erreur mot, avec les intervalles de confiance à 95% associés, sont représentés par les barres "*CA-GMM*" de la figure 4. Les meilleures performances sont obtenues avec 4 classes (cf. ligne "*4 cla. CA-GMM*" de la table 2) et elles sont similaires à celles obtenues avec les modèles adaptés en fonction de l'âge et du sexe des locuteurs (cf. table 1). Au delà de 4 classes, les performances se détériorent en raison de la diminution de la quantité de données disponibles pour l'adaptation des modèles des classes.

4 Structuration des composantes pour les HMM-SWGMM

Au lieu d'adapter tous les paramètres des HMM pour chaque classe de données, l'approche HMM-SWGMM (*HMM-GMM with Speaker class-dependent Weights*) récemment proposée (Gorin et Jouvet, 2013) repose sur la structuration en classes des composantes gaussiennes, le partage des composantes gaussiennes entre les différentes classes, et la spécialisation des pondérations des gaussiennes à chaque classe. Ainsi, dans un HMM-SWGMM, la densité b_j pour une classe de locuteurs c est définie par $b_j^{(c)}(o_t) = \sum_{k=1}^{M} w_{jk}^{(c)} \mathcal{N}(o_t, \mu_{jk}, U_{jk})$, où M est le nombre de composantes par densité, o_t est le vecteur d'observation au temps t et $\mathcal{N}(o_t, \mu_{jk}, U_{jk})$ est la

composante gaussienne de moyenne μ_{jk} et de covariance U_{jk}. Lors du décodage, chaque phrase à reconnaître est d'abord automatiquement classifiée, et assignée à une classe c. Ensuite, le décodage est effectué avec le jeu de pondérations des gaussiennes qui correspond à cette classe.

La structuration des composantes gaussiennes est obtenue en initialisant les GMM par concaténation des composantes gaussiennes de GMM de plus faible dimensionalité appris sur les différentes classes. Par exemple, pour fabriquer un modèle avec M composantes gaussiennes associées à Z classes, Z modèles ayant chacun $L = M/Z$ composantes par densité sont appris. Puis, ces composantes sont regroupées dans un mélange global (cf. figure 2-a pour les moyennes).

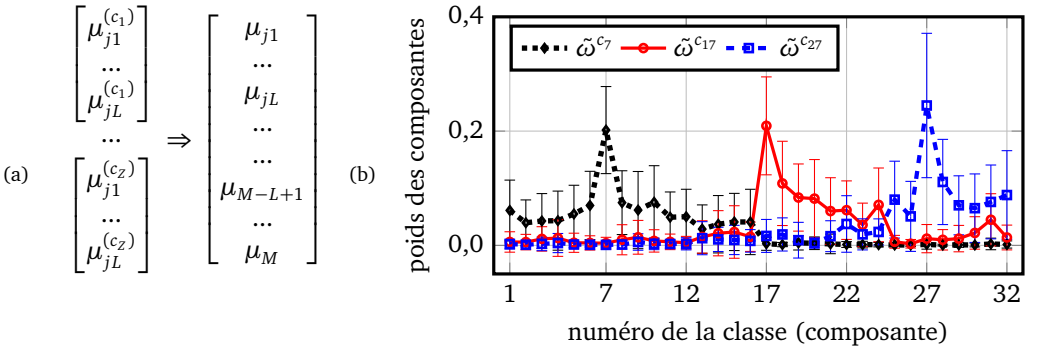

FIGURE 2: (a) Initialisation du SW-GMM à partir de plusieurs modèles associés aux classes et (b) statistiques des pondérations des composantes pour les classes c_7, c_{17} et c_{27} après ré-estimation (moyennes et écarts-types calculés sur toutes les densités ; Z=32, M=32)

Lors de l'initialisation du modèle structuré SW-GMM, les pondérations des modèles des classes sont également concaténées, puis renormalisées. Ensuite, les moyennes, variances et pondérations sont ré-estimées. Les pondérations, spécifiques à chaque classe, sont apprises à partir des données de la classe correspondante, tandis que toutes les données sont utilisées pour ré-estimer les moyennes et les variances qui sont partagées entre les classes :

$$\omega_{jk}^{(c_i)} = \frac{\sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t)}{\sum_{l=1}^{M} \sum_{t=1}^{T} \gamma_{jl}^{(c_i)}(t)} \qquad \mu_{jk} = \frac{\sum_{i=1}^{Z} \sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t) o_t}{\sum_{i=1}^{Z} \sum_{t=1}^{T} \gamma_{jk}^{(c_i)}(t)} \tag{1}$$

où $\gamma_{jk}^{(c_i)}(t)$ est le compteur Baum-Welch de la $k^{ème}$ composante de la densité b_j, générant l'observation o_t de la classe c_i. Les sommes sur t portent sur toutes les trames des phrases d'apprentissage de la classe concernée. Les variances sont ré-estimées de manière similaire aux moyennes.

Après ré-estimation, les pondérations dépendantes de la classe sont plus grandes pour les composantes associées à la même classe de données ; la figure 2-b montre l'exemple des pondérations des classes c_7, c_{17} et c_{27} , moyennés sur les toutes les densités.

Expériences avec les modèles HMM-SWGMM structurés en fonction des classes. Les modèles HMM-SWGMM à 32 gaussiennes par densité sont initialisés à partir de modèles de classes ayant moins de composantes, par exemple 2 modèles de classes à 16 gaussiennes, ou 4 modèles à 8 gaussiennes, et ainsi de suite jusqu'à 32 modèles de classes monogaussiens. Les paramètres sont ré-estimés au maximum de vraisemblance (MLE) pour les pondérations des gaussiennes dépendantes de chaque classe, et par maximum *a posteriori* (MAP) pour les moyennes et les variances partagées. Les performances sont indiquées par les barres "*SW-GMM*" dans la figure 4. Cette paramétrisation permet une estimation robuste des moyennes et des variances partagées, tout en gardant une dépendance vis-à-vis des classes avec les pondérations des composantes.

Avec un nombre limité de paramètres, le taux d'erreur est significativement réduit ; 0,80% sur les données des adultes et 1,05% sur les données enfants (cf. ligne *32 cla. SW-GMM* de la table 2).

5 Structuration des composantes pour les SGMM

Les modèles SGMM (*Stranded GMM*), proposés pour la reconnaissance robuste (Zhao et Juang, 2012), reposent sur un enrichissement des densités d'émission des HMM-GMM par l'introduction de dépendances explicites entre les composantes des densités d'états adjacents. Tandis que dans la version proposée par Zhao, les SGMM sont initialisés à partir de HMM-GMM conventionnels, cette section propose d'exploiter la structuration des composantes en fonction de classes, pour obtenir des SSGMM (*class-Structured SGMM*).

5.1 SGMM Conventionnels

Les modèles SGMM conventionnels font intervenir la suite des états $\mathscr{Q} = \{q_1, ..., q_T\}$, la suite des vecteurs d'observation $\mathscr{O} = \{o_1, ..., o_T\}$, et la suite des composantes des densités $\mathscr{M} = \{m_1, ..., m_T\}$, où $m_t \in \{1, ..., M\}$ désigne la composante gaussienne utilisée à l'instant t, et M précise le nombre de composantes par densité.

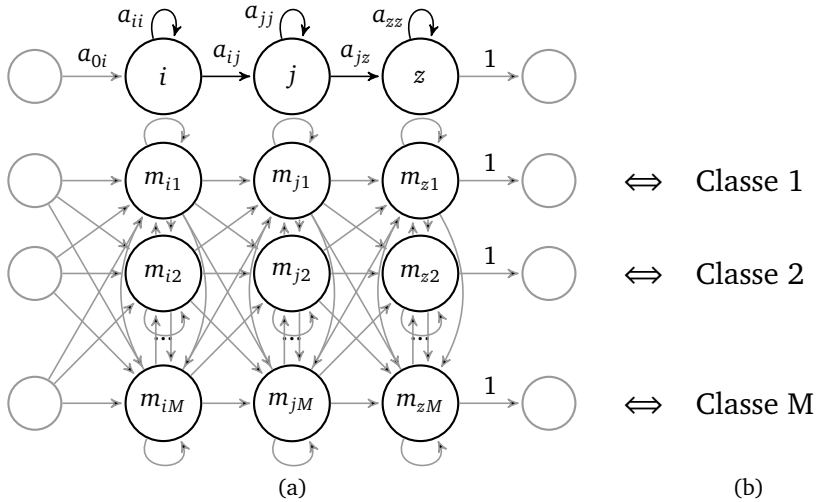

FIGURE 3: (a) *Stranded* GMM avec représentation des dépendances entre composantes gaussiennes ; (b) associations entre classes et composantes gaussiennes pour les SSGMM.

En comparaison des HMM-GMM conventionnels, les SGMM modélisent les dépendances entre la composante gaussienne m_t utilisée à l'instant t et la composante m_{t-1} utilisée à la trame précédente (cf. figure 3-a). La vraisemblance conjointe de la séquence d'observations, de la séquence d'états, et de la séquence de composantes est donnée par :

$$P(\mathscr{O}, \mathscr{Q}, \mathscr{M}|\lambda) = \prod_{t=1}^{T} P(o_t|m_t, q_t) P(m_t|m_{t-1}, q_t, q_{t-1}) P(q_t|q_{t-1}) \tag{2}$$

où $P(q_t = j|q_{t-1} = i) = a_{ij}$ est la probabilité de transition entre états, $P(o_t|m_t = l, q_t = j) = b_{jl}(o_t)$ est la probabilité de l'observation o_t pour la composante gaussienne $m_t = l$ de la densité associée à l'état $q_t = j$ et $P(m_t = l|m_{t-1} = k, q_t = j, q_{t-1} = i) = c_{kl}^{(ij)}$ est la probabilité de transition entre les composantes gaussiennes. L'ensemble des probabilités de transition

entre les composantes détermine les matrices de transition MTM (*Mixture Transition Matrices*) $C^{(ij)} = \{c_{kl}^{(ij)}\}$, avec les contraintes $\sum_{l=1}^{M} c_{kl}^{(ij)} = 1, \forall i, j, k$.

Expériences avec les SGMM conventionnels. Afin de réduire le nombre de paramètres, 2 matrices MTM seulement sont utilisées pour chaque état, l'une correspondant au bouclage sur l'état, et l'autre au passage à l'état suivant (les matrices MTM associées au dernier état d'un phonème et correspondant aux transitions vers un autre phonème sont partagées). Les taux d'erreur mot obtenus sont indiqués par les barres "*SGMM*" dans la figure 4 et dans la ligne correspondante de la table 2. Par rapport aux HMM-GMM conventionnels, les SGMM améliorent les performances, de 1,66 % à 1,11 % sur les données adultes et de 1,88 % à 1,27 % sur les données enfants. Ces améliorations sont significatives, compte tenu de l'intervalle de confiance à 95 %. L'approche SGMM, qui fonctionne en une seule passe, aboutit même à des performances meilleures que la modélisation de référence adaptée en fonction de l'âge et du sexe des locuteurs, sans toutefois dépasser les performances de l'approche HMM-SWGMM proposée dans la section précédente.

5.2 SGMM avec structuration en classes des composantes

La modélisation SSGMM (*class-Structured SGMM*) proposée repose sur la structuration en classes des composantes des densités, de telle manière que initialement, la $k^{ème}$ composante de chaque densité corresponde à une même classe de données (cf. figure 3-b). Pour obtenir cette structuration, le SSGMM est initialisé à partir du HMM-SWGMM décrit dans la section 4. Les moyennes et variances des gaussiennes sont obtenues directement à partir du HMM-SWGMM et les matrices MTMs sont initialisées avec des distributions uniformes. Les pondérations des gaussiennes du HMM-SWGMM, qui sont dépendantes des classes, ne sont pas utilisées.

Quand les HMM-SWGMM, qui servent à fabriquer les SSGMM, sont initialisés à partir ce modèles de classes monogaussiens, chaque composante correspond à une classe. Après ré-estimation, les éléments diagonaux des matrices MTM dominent, ce qui conduit à favoriser la consistance de la classe lors du décodage d'une phrase. Cependant, les éléments non-diagonaux non nuls rendent possibles les contributions d'autres composantes gaussiennes dans le calcul des scores acoustiques. L'avantage des SSGMM est qu'ils modélisent explicitement les trajectoires, tout en autorisant des changements de composantes (ou de classes). De plus, le décodage d'une phrase fonctionne en une seule passe ; il n'y a pas de classification préalable à faire.

Modèle	Decodage	Paramètres par état	**Adult.**	Hom.	Fem.	**Enf.**	Garç.	Fil.
SI GMM	1 passe	78*32+32=2528	1,66	1,86	1,46	1,88	1,69	2,08
4 cla. CA-GMM	2 passes	4*(78*32+32)=10112	**1,32**	1,47	1,17	**1,57**	1,61	1,52
8 cla. HMM-SWGMM	2 passes	78*32+8*32=2752	**0,75**	0,81	0,70	**1,21**	1,25	1,17
32 cla. HMM-SWGMM	2 passes	78*32+32*32=3520	**0,80**	0,96	0,64	**1,05**	1,03	1,08
SGMM	1 passe	78*32+2*32*32=4544	**1,11**	1,26	0,96	**1,27**	1,19	1,36
SSGMM	1 passe	78*32+2*32*32=4544	**0,52**	0,64	0,40	**0,86**	0,74	0,98

TABLE 2: Taux d'erreur mot et indication du nombre de paramètres par état. La classe du locuteur est inconnue lors du décodage, et estimée par GMM pour les versions "2 passes"

Expériences avec les SGMM structurés en classes. Pour cette expérience, les SSGMM sont fabriqués à partir du HMM-SWGMM à 32 classes (i.e. modèle *32 cla. HMM-SWGMM* dans la table 2). De même que précédemment, deux matrices MTM sont alors définies pour chaque état, et

initialisées avec des distributions uniformes, puis les paramètres des SSGMM sont ré-estimés par maximum de vraisemblance. Les performances sont indiquées avec les barres "*SSGMM* " dans la figure 4 et dans la ligne correspondante de la table 2. L'approche SSGMM proposée, qui repose sur la structuration en classes des composantes, améliore encore plus que les SGMM conventionnels, et permet d'obtenir un taux d'erreur mot de 0,52% sur les données adultes, et de 0,86% sur les données enfants. A noter que la fabrication de SSGMM à partir de HMM-SWGMM construits sur la base de différents nombres de classes (2, 4, 8 et 16) conduit aussi à des améliorations de performances par rapport au SGMM ; seul le résultat correspondant à 32 classes est indiqué.

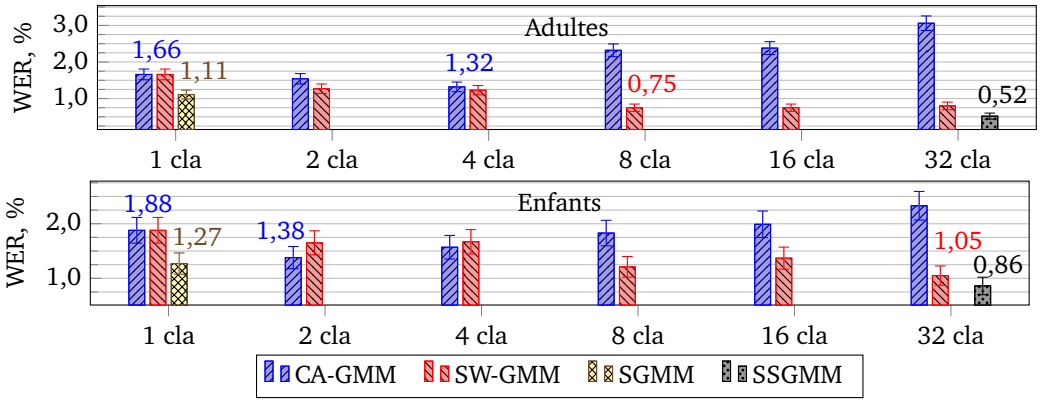

FIGURE 4: Taux d'erreur mot (WER) sur les données adultes et enfants, obtenus avec les CA-GMM (*Class-Adapted models*), les HMM-SWGMM (*HMM with Speaker class-dependent Weights*), les SGMM (*Stranded GMM*) et les SSGMM (*class-Structured SGMM*) contruits à partir de 32 classes

6 Conclusion et perspectives

Plusieurs approches pour améliorer la reconnaissance de données hétérogènes ont été étudiées. La taille limitée des corpus d'apprentissage ne permet pas de fabriquer un grand nombre de modèles de classes. Cependant, ces modèles de classes permettent de structurer les composantes gaussiennes des densités GMM en associant la $k^{ème}$ composante de chacune des densités à une même classe de locuteurs.

Deux types de modélisations exploitant cette structuration des composantes ont été proposées et approfondies, et ont montré un gain significatif en performances. Le premier modèle, HMM-SWGMM (*Speaker class-dependent Weights GMM*) repose sur un partage des composantes gaussiennes entre les différentes classes, et des pondérations des composants gaussiennes spécifiques à chaque classe. Ce partage des paramètres rend la modélisation robuste. La structuration des composantes a ensuite été appliquée aux *Stranded* GMM qui modélisent explicitement les trajectoires grâce à l'adjonction de matrices de transition entre les composantes des densités GMM. La modélisation résultante, SSGMM (*class-Structured SGMM*), conduit à un gain significatif en performances par rapport aux autres approches, pour la reconnaissance de données hétérogènes (voix d'adultes et voix d'enfants). De plus cette approche ne requiert pas de phase préalable de classification des données à reconnaître.

Une amélioration possible consistera à combiner les modélisations proposées avec une adaptation des paramètres reposant sur l'ajustement de l'axe fréquentiel, comme par exemple la normalisation VTLN (*Vocat Tract Length Normalization*).

Références

Beaufays, F., Vanhoucke, V. et Strope, B. (2010). Unsupervised Discovery and Training of Maximally Dissimilar Cluster Models. *In Proc. INTERSPEECH*, pages 66–69, Makuhari, Japan.

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Tyagi, V. et Wellekens, C. (2007). Automatic speech recognition and speech variability : A review. *Speech Communication*, 49(10):763–786.

Burnett, D. C. et Fanty, M. (1996). Rapid unsupervised adaptation to children's speech on a connected-digit task. *In Proc. ICSLP*, volume 2, pages 1145–1148. IEEE.

CMU (2014). Sphinx toolkit http ://cmusphinx.sourceforge.net.

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer speech & language*, 12(2):75–98.

Gauvain, J.-L. et Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *Speech and audio processing, IEEE transactions on*, 2(2):291–298.

Gorin, A. et Jouvet, D. (2012). Class-based speech recognition using a maximum dissimilarity criterion and a tolerance classification margin. *In Proc. Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 91–96. IEEE.

Gorin, A. et Jouvet, D. (2013). Efficient constrained parametrization of GMM with class-based mixture weights for Automatic Speech Recognition. *In Proc. LTC-6th Language & Technologies Conference*, pages 550–554.

Jouvet, D., Gorin, A. et Vinuesa, N. (2012). Exploitation d'une marge de tolérance de classification pour améliorer l'apprentissage de modèles acoustiques de classes en reconnaissance de la parole. *In JEP-TALN-RECITAL*, pages 763–770.

Kuhn, R., Nguyen, P., Junqua, J.-C., Goldwasser, L., Niedzielski, N., Fincke, S., Field, K. et Contolini, M. (1998). Eigenvoices for speaker adaptation. *In Proc. ICSLP*, volume 98, pages 1774–1777.

Leonard, R. G. et Doddington, G. (1993). Tidigits speech corpus. *Texas Instruments, Inc*.

Stern, R. M. et Morgan, N. (2012). Hearing is believing : Biologically inspired methods for robust automatic speech recognition. *Signal Processing Magazine, IEEE*, 29(6):34–43.

Wellekens, C. J. (1987). Explicit time correlation in hidden Markov models for speech recognition. *In Proc. ICASSP*, pages 384–386.

Zhan, P. et Waibel, A. (1997). Vocal tract length normalization for large vocabulary continuous speech recognition. *In Technical report*, DTIC Document.

Zhao, Y. et Juang, B.-H. (2012). Stranded Gaussian mixture hidden Markov models for robust speech recognition. *In Proc. ICASSP*, page 4301–4304.