UNIVERSITÉ PARIS SUD

Comprendre le monde, construire l'avenir®

Université Paris-Sud

Ecole Doctorale 427 : Informatique Paris Sud

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

Thèse de doctorat

Informatique

par

Thiago FRAGA DA SILVA

Réduction des coûts de développement de systèmes de reconnaissance de la parole à grand vocabulaire

Date de soutenance : 29/09/2014

Composition du jury :

Directeur de thèse :	M. Jean-Luc Gauvain	Directeur de Recherche (CNRS-LIMSI)
Rapporteurs :	M. Denis Jouvet	Directeur de Recherche (INRIA-LORIA)
	Mme. Tanja Schultz	Professeur (Karlsruhe Institute of Technology)
Examinateurs :	Mme. Anne Vilnat	Professeur (Université Paris-Sud)
	M. Ralf Schlüter	Chercheur Associé (RWTH Aachen University)
	M. Driss Matrouf	Maître de Conférences - HDR (LIA-CERI)

1 Introduction

Le but principal de la reconnaissance automatique de la parole (RAP) est d'effectuer la conversion entre un énoncé oral, représenté par un flux audio continu, vers une séquence discrète de mots écrits. La RAP peut être utile pour des différentes applications, telles que la commande vocale, l'indexation de médias audio, la recherche d'information, la traduction automatique de la parole, la dictée ou la communication assistée par ordinateur pour des personnes souffrant de déficience auditive.

La RAP est un thème de recherche étudié depuis des dizaines d'années. Actuellement, les systèmes de RAP sont capables de traiter de la parole continue à grand vocabulaire ayant une large variété de locuteurs avec une précision raisonnable. Néanmoins et pour des raisons historiques, les plus hauts niveaux de performances ne sont obtenus qu'avec les quelques langues les plus parlées dans le monde.

Depuis les dernières années, l'intérêt en construire de systèmes de RAP à grand vocabulaire pour des différentes langues, accents et domaines a fortement crû tant au milieu académique quant au milieu industriel. Des nombreuses entreprises, tels que Apple, IBM, Google, Microsoft, comme beaucoup d'autres, ont déjà montré un intérêt dans le développement des technologies de la parole pour l'utilisation dans leurs produits.

Les systèmes RAP à l'état de l'art reposent sur les principes de la reconnaissance statistique de formes. Dans cette approche, par le biais de la décision bayésienne, la tâche de la reconnaissance consiste à rechercher la séquence de mots la plus probable étant donné le flux audio et deux modèles stochastiques : le modèle acoustique (MA) et le modèle de langue (ML).

Les paramètres de ces modèles sont obtenus par l'intermédiaire d'un algorithme d'apprentissage qui vise à ajuster un ensemble d'échantillons à une fonction objectif appropriée. Étant donnée que la reconnaissance de la parole est un problème de classification, les étiquettes correctes des données d'apprentissage sont généralement nécessaires. Bien que les choix d'algorithmes et critères d'optimisation comptent encore sur des compétences humaines, l'approche statistique considère que les données sont les vrais guides du processus modélisé.

En ce qui concerne ces données, deux hypothèses fondamentales sont considérées :

1. Le corpus utilisé pour l'apprentissage doit être **suffisamment grand** afin de produire des estimations fiables et mener une bonne généralisation à des nouvelles données.
2. Les données doivent être **représentatives** pour la tâche cible. En d'autres termes, on suppose que les données d'apprentissage et de test sont des échantillons indépendants et identiquement distribués d'une même densité de probabilité.

Malgré les énormes progrès qui ont été réalisés dans le domaine de la reconnaissance de la parole au cours des dernières décennies, le développement d'un système de reconnaissance pour une nouvelle langue ou domaine reste toujours une tâche coûteuse et assez longue. Réduire les coûts de développement de systèmes est une étape nécessaire afin d'assurer l'expansion rapide des technologies de reconnaissance de la parole.

Plusieurs facteurs peuvent influencer sur les coûts et délais de développement. Une importante partie des coûts est liée à l'effort humain nécessaire sur les décisions architecturales du système (types d'attributs acoustiques, types de modèles, critères d'apprentissage, algorithmes de lissage, stratégies de décodage, etc), ainsi que sur la production de

corpora d'apprentissage annotés. Il est bien connu que l'un des coûts les plus importants impliqués dans le développement des systèmes de RAP repose sur l'annotation manuelle de données audio. Réduire la nécessité d'étiqueter des données audio conduit certainement à une réduction des coûts de développement et, par conséquent, favorise l'expansion des technologies de reconnaissance de la parole. Le première axe de recherche de cette thèse couvre l'utilisation de données audio annotées automatiquement pour l'estimation des modèles acoustiques et de langue.

La disponibilité d'une grande quantité de données n'est pas la seule condition nécessaire pour la construction des modèles performants : la similarité entre les données d'apprentissage et les données de test est également importante. Cependant, des grandes quantités de données ne sont pas toujours disponibles pour certaines langues, dialectes ou domaines, dits "peu dotés". Ces cas demandent des efforts supplémentaires pour la production des corpora d'apprentissage ainsi que pour l'affinement des paramètres du système. Le deuxième axe de recherche de cette thèse couvre l'utilisation de techniques visant à réduire le besoin de données d'apprentissage spécifiques au domaine cible, réduisant ainsi l'effort humain nécessaire pour développer des systèmes pour des tâches peu dotés.

2 Apprentissage non-supervisé de modèles acoustiques

L'apprentissage non-supervisé des modèles acoustiques est un problème d'estimation avec des données incomplètes. Seul la séquence des vecteurs acoustiques ($\mathcal{X}$) est observable, pendant que la séquence d'états des modèles de Markov (MM) ($\mathcal{S}$), la séquence d'étiquettes de Gaussiennes ($\mathcal{L}$) et la séquence des mots ($\mathcal{W}$) sont non-observables (ou cachées). Ce problème peut être résolu à l'aide de l'algorithme espérance-maximisation (EM) à travers l'optimisation de la fonction auxiliaire suivante :

$$Q(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) = E\left[\log f(\mathcal{S}|\mathcal{W}, \boldsymbol{\lambda})\middle|\mathcal{X}, \hat{\boldsymbol{\lambda}}\right] + E\left[\log f(\mathcal{X}, \mathcal{L}|\mathcal{S}, \boldsymbol{\lambda})\middle|\mathcal{X}, \hat{\boldsymbol{\lambda}}\right] \tag{2.1}$$

où $\mathcal{Y} = (\mathcal{X}, \mathcal{W}, \mathcal{S}, \mathcal{L})$ représente les données complètes.

La fonction auxiliaire (2.1) peut être optimisée en deux étapes, comme suit :

Décodage Étant donné un modèle acoustique $\hat{\boldsymbol{\lambda}}$ (ainsi que le modèles de langue), les données audio $\mathcal{X}$ sont décodées. Une probabilité d'alignement entre états et trames $P(\mathcal{S}|\mathcal{X}, \hat{\boldsymbol{\lambda}})$ ainsi que une probabilité *a posteriori* des séquences des mots $P(\mathcal{W}|\mathcal{S}, \mathcal{X}, \hat{\boldsymbol{\lambda}})$ sont attribués. Le décodage est guidé par la maximisation suivante :

$$(\mathcal{W}^*, \mathcal{S}^*) = \arg\max_{\mathcal{W}} \max_{S} P(\mathcal{W}) \cdot P(\mathcal{S}|\mathcal{W}, \hat{\boldsymbol{\lambda}}) \cdot f(\mathcal{X}|\mathcal{S}, \hat{\boldsymbol{\lambda}}) \tag{2.2}$$

Mise à jour du modèle Étant donné $P(\mathcal{W}, \mathcal{S}|\mathcal{X}, \hat{\boldsymbol{\lambda}}) = P(\mathcal{S}|\mathcal{X}, \hat{\boldsymbol{\lambda}}) \cdot P(\mathcal{W}|\mathcal{S}, \mathcal{X}, \hat{\boldsymbol{\lambda}})$, un nouveau modèle $\boldsymbol{\lambda}$ est estimé utilisant :

$$\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda}} \sum_{\mathcal{W}} \sum_{\mathcal{S}} \sum_{\mathcal{L}} P(\mathcal{L}, \mathcal{S}, \mathcal{W}|\mathcal{X}, \hat{\boldsymbol{\lambda}}) \cdot \log f\left(\mathcal{X}, \mathcal{L}, \mathcal{S}|\mathcal{W}, \boldsymbol{\lambda}\right) \tag{2.3}$$

La somme en (2.3) est faite parmi toutes les séquences de mots, état et étiquettes de Gaussiennes possibles, ce qui est évidemment un calcul assez coûteux. La méthode

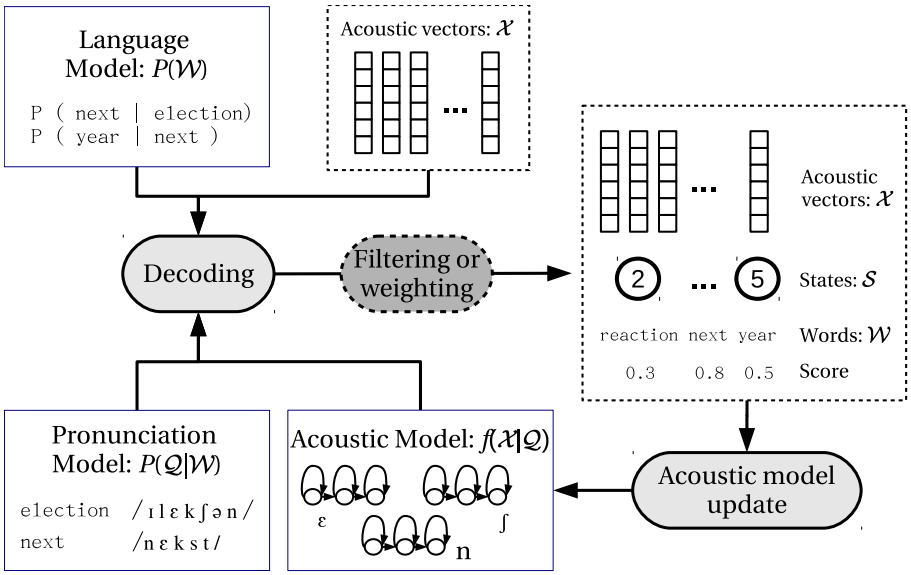

FIGURE 2.1 – Schéma de l'apprentissage non-supervisé des modèles acoustiques.

standard utilisée pour l'apprentissage non-supervisé se sert d'une approximation visant à réduire la complexité de calcul : seule la meilleure hypothèse de transcription (séquence de mots) est considérée. Cette approche est montré dans la Figure 2.1. Un système existant est utilisé pour générer des transcriptions automatiques d'un jeu de données d'entraînement qui sont ensuite utilisées pour estimer le modèle acoustique.

Le principal problème de l'approche standard (1-meilleur) est que les transcriptions automatiques contiennent des erreurs qui peuvent induire l'algorithme à des erreurs d'estimation. Ce problème est bien connu et souvent traité par des méthodes de filtrage ou pondération, tous les deux basées sur des mesures de confiance fournies par le système de reconnaissance. Ces mesures de confiance peuvent être calculées à plusieurs niveaux de granularité, tels que le mots, le phonèmes ou le états des MMs. Dans cette thèse, j'ai montré que les méthodes de filtrage et pondérations pour ces différentes granularités sont des approximations mathématiques d'un même l'algorithme d'entraînement, tel que montré précédemment en utilisant :

$$\boldsymbol{\lambda}^* = \arg \max_{\boldsymbol{\lambda}} \max_{\mathcal{W},\mathcal{S}} \sum_{\mathcal{L}} P(\mathcal{W},\mathcal{S},\mathcal{L}|\mathcal{X},\hat{\boldsymbol{\lambda}}) \cdot \log f\left(\mathcal{X},\mathcal{L},\mathcal{S}|\boldsymbol{\lambda}\right) \tag{2.4}$$

au lieu de (2.3), où $P(\mathcal{W},\mathcal{S}|\mathcal{X},\hat{\boldsymbol{\lambda}})$ est calculé à partir des mesures de confiance spécifiques.

Une autre approximation, plus souple, a été proposé dans le cadre de ce travail. Au lieu de considérer seul la meilleure, plusieurs hypothèses extraites des treillis de décodage on été utilisées, conduisant à :

$$\boldsymbol{\lambda}^* = \arg \max_{\boldsymbol{\lambda}} \sum_{\mathcal{W},\mathcal{S}\in\boldsymbol{L}} \sum_{\mathcal{L}} P(\mathcal{W},\mathcal{S},\mathcal{L}|\mathcal{X},\hat{\boldsymbol{\lambda}}) \cdot \log f\left(\mathcal{X},\mathcal{L},\mathcal{S}|\boldsymbol{\lambda}\right) \tag{2.5}$$

où $\boldsymbol{L}$ représente le treillis.

Méthode	devQ10				testQ10	testQ11	Global
	1ère (18h)	2ème (36h)	3ème (72h)	4ème (72h)	4ème (72h)	4ème (72h)	4ème (72h)
Initial	53.7				45.9	54.8	51.5
1-meilleur	41.5	36.7	33.5	33.0	26.8	33.0	30.9
1-meilleur pondéré par phone	41.1	35.9	32.5	31.9	26.3	32.2	30.1
1-meilleur filtré par état	40.6	35.7	32.6	31.8	26.3	32.2	30.1
Treillis	40.5	35.3	32.4	31.9	26.2	32.3	30.1

Table 2.1 – Comparaison des méthodes d'apprentissage non-supervisé pour la modélisation acoustique. Le taux d'erreur de mots (%) est mesuré sur les données de développement `devQ10` pour toutes les itérations, et sur les données de test `testQ10` et `testQ11` pour la dernière itération. Les chiffres entre parenthèses représentent la quantité de données d'apprentissage (en heures) utilisés à chaque itération.

Après avoir fait un étude empirique de l'influence des paramètres de décodage sur la performance des modèles acoustiques obtenus à partir de la méthode basée sur des treillis, une comparaison entre plusieurs méthodes non-supervisées a été faite sur un système de reconnaissance à grand vocabulaire d'émissions diffusées du Portugais Européen. Un extrait des résultats des expériences est montré dans le Tableau 2.1. Ces résultats montrent tout d'abord que l'application de mesures de confiance est une étape très importante lorsque l'estimation des modèles acoustiques est faite avec des données transcrites automatiquement. Les modèles obtenus à partir de treillis obtiennent les meilleurs résultats surtout dans les premières itérations, c'est-à-dire lorsque la quantité d'erreurs présente dans les transcriptions automatiques est plus importante. Les différences de performance obtenus dans les dernières itérations sont négligeables.

Hormis l'utilisation des mesures de confiance pour réduire l'effet des erreurs sur les méthodes non-supervisées, nous avons étudier l'effet de la stratégie d'entraînement sur la propagation d'erreurs d'une itération à l'autre. L'idée est d'éviter le surapprentissage à des données mal étiquetées. Les stratégies évaluées diffèrent sur la manière dont des sous-ensembles de données sont utilisées d'une itération à l'autre. Pour faire la différence entre ces stratégies, nous considérons que les données d'apprentissage soient réparties en quatre sous-ensembles, disons a, b, c et d ayant la même durée. Les approches étudiées dans ce travail peuvent être représentées comme suit :

Toutes les données	$(a,b,c,d) \Rightarrow (a,b,c,d) \Rightarrow (a,b,c,d) \ldots$
Incrémental	$(a) \Rightarrow (a,b) \Rightarrow (a,b,c,d) \Rightarrow (a,b,c,d) \ldots$
Différentiel 1	$(a) \Rightarrow (b) \Rightarrow (c,d) \Rightarrow (a,b,c,d) \Rightarrow (a,b,c,d) \ldots$
Différentiel 2	$(a) \Rightarrow (b) \Rightarrow (c) \Rightarrow (d) \Rightarrow (a,b) \Rightarrow (c,d) \Rightarrow (a,b,c,d) \ldots$

Table 2.2 – Stratégies d'entraînement non-supervisé évaluées dans ce travail.

Ces stratégies ont été évalués sur environ 10 heures de données. Le Tableau 2.3 résume les résultats obtenus. L'influence de la stratégie d'entraînement est plus claire pour

Stratégie d'entraînement	# of Itér.	Quantité de données décodées (en heures)	TEM(%)			
			1-meilleur	1-meilleur pondéré	1-meilleur filtré	Treillis
Toutes les données	3	216	31.5	30.4	30.7	30.3
Incrémental	4	198	30.9	30.1	30.1	30.1
Différentiel 1	4	144	30.6	30.2	30.1	30.1
Différentiel 1	5	216	30.3	29.9	29.8	29.7
Différentiel 2	7	216	**30.2**	-	**29.7**	**29.6**

Table 2.3 – Comparaison des stratégies d'entraînement non-supervisé pour la modélisation acoustique. Le taux d'erreur de mots (TEM) a été évalué sur 10 heures de données contenant `devQ10`, `testQ10` and `testQ11`. Le nombre d'itérations et la quantité totale de données d'entraînement décodées sont montrés. Tous les modèles ont été estimés sur les 72 heures de données de `trainQ10`.

la méthode non-supervisée "1-meilleur". Cela se justifie du fait que les autres méthodes utilisent également de mesures de confiance afin de réduire l'impact des erreurs de reconnaissance. La stratégie 'Différentiel 2' conduit aux meilleures performances, soit une amélioration absolue de 1,3% sur le taux erreur de mots (TEM) par rapport à la stratégie 'Toutes les données' lorsque les modèles acoustiques sont obtenus avec la méthode 1-meilleur. Les meilleures performances générales sont obtenues en utilisant la stratégie 'Différentiel 2' avec la méthode non-supervisée basée sur des treillis proposée dans le cadre de cette thèse, soit une amélioration absolue de 1.9% sur la référence (29.6% vs. 31.5%).

3 Apprentissage non-supervisé de Perceptron multicouche

Comme montré dans la dernière section, les méthodes non-supervisées sont bien utiles pour améliorer la performances de modèles acoustiques lorsque une grande quantité de données audio non-transcrites sont disponibles. La plupart des travaux portant sur des méthodes d'apprentissage non-supervisé font l'usage d'attributs acoustiques extraits directement du flux audio à partir d'analyses spectrales bien connues, telles que la prédiction linéaire perceptuelle ou des coefficients cepstraux sur l'échelle de fréquence Mel.

Dans les dernières années, il y a eu un grand intérêt en augmenter la capacité discriminante des attributs acoustiques en vue d'améliorer leur apport en performance. Un type d'attribut qui s'inclut dans cette catégorie sont les vecteurs acoustiques extraits à partir d'un modèle Perceptron multicouche (PCM). En particulier, l'architecture *bottleneck* (goulot) (Fousek et al., 2008) a déjà montré être capable de conduire à des gains de performance substantiels en comparaison aux attributs cepstraux pour plusieurs applications différentes. Le modèle PCM *bottleneck* est montré dans la Figure 3.1. Dans ce cas, le vecteur acoustique est extraits de la 3ème couche (le goulot). La couche de sortie n'est utilisé que pendant l'estimation des paramètres du modèle.

Tel que les modèles de Markov, les PCM ont aussi besoin des paires entrée (vecteurs acoustiques) et sorties (étiquettes) à fin d'être estimés. Dans le cas des PCM, les étiquettes

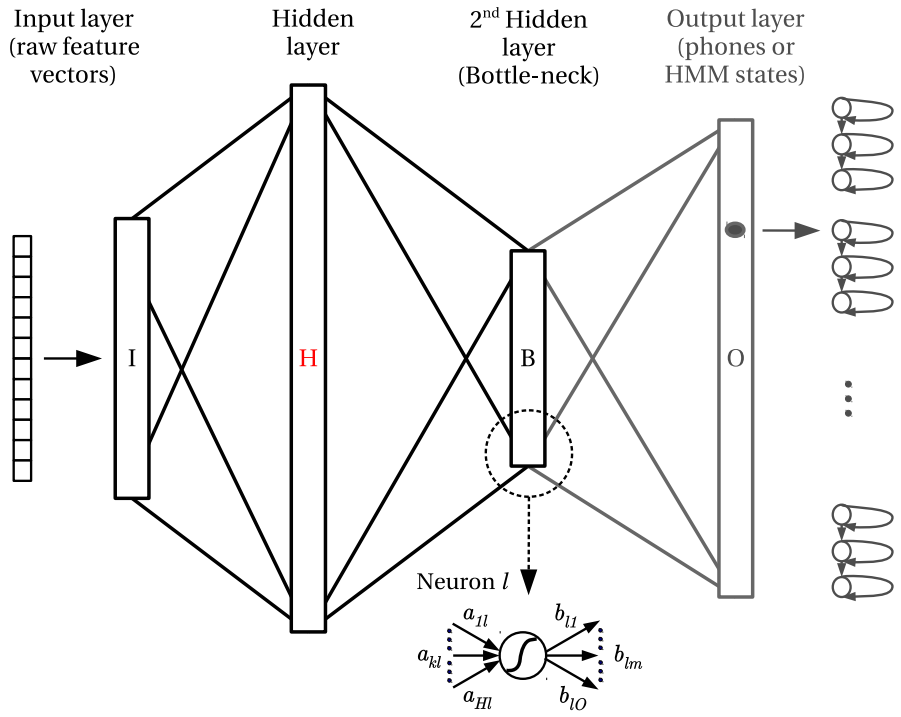

FIGURE 3.1 – Un Perceptron multicouche *bottleneck* avec 4 couches. Les attributs acoustiques sont extraits de la 3ème couche (le goulot). La couche de sortie n'est utilisé que pendant l'entraînement.

MODÈLE ACOUSTIQUE	QUANTITÉ DE DONNÉES (EN HEURES)	PLP	PCMPLPF0
`trainRVE` / Non-supervisé (référence)	173	33.0	-
`trainQ10` / Non-supervisé	72	31.1	29.9
`trainQ10` / Supervisé	72	29.1	27.3

TABLE 3.1 – Comparaison entre PCM obtenus à partir d'entraînement supervisé et non-supervisé. Le TEM (%) est mesuré dans le jeu de développement `devQ10`.

de sortie peuvent être phonèmes ou bien des états des MMs. En présence des transcriptions manuelles correspondantes à l'audio, ces étiquettes peuvent être obtenues à partir d'un alignement forcé. Comme ces transcriptions ne sont toujours pas disponibles, nous proposons dans cette thèse d'utiliser des transcriptions automatiques obtenues à partir d'un système de reconnaissance existant.

L'approche d'apprentissage non-supervisé a été comparée avec l'approche supervisée pour un même jeu de données d'entraînement. Le Tableau 3.1 résume les résultats obtenus. Tout d'abord, il est possible de voir que l'ajout des attributs PCM induit à des gains de performance importants, soit une réduction absolue du TEM de 1.2% pour les modèles PCM et MM obtenus de façon non-supervisée et 1.8% pour les modèles PCM et MM obtenus de façon supervisée.

Les dernières résultats montrent bien que les PCM peuvent aussi être entraînés avec des données audio non-transcrites. Vue que les données non-transcrites ne sont pas chers à obtenir, nous avons évaluer l'effet de l'ajout des données sur l'estimation des paramètres

PCM	Modèles acoustiques		
	trainQ10 (72h)	trainQ10 + trainQ11 (143h)	trainRVE + trainQ10 + trainQ11 (316h)
trainQ10 (72h)	29.9	-	-
trainQ10 + trainQ11 (143h)	29.5	28.2	-
trainRVE + trainQ10 + trainQ11 (316h)	28.6	**27.8**	27.9

Table 3.2 – Effet de l'ajout des données non-transcrites sur l'estimation des paramètres des PCM et MMs. Le TEM (%) a été mesuré sur le jeu de développement devQ10. Les durées des jeux d'apprentissage sont montrées en parenthèses.

PCM	Modèles Acoustiques	
	trainQ10 (72h)	trainQ10 + trainQ11 (143h)
trainRVE + trainQ10 + trainQ11 (316h)	28.6	27.8
Anglais (645h)	28.7	-
Français (600h)	28.3	27.3
trainRVE + trainQ10 + trainQ11 (2ème itération)	28.4	**27.4**
Français (600h) (2ème itération)	28.1	**27.2**

Table 3.3 – Comparaison entre PCM obtenus à partir de l'apprentissage non-supervisé et PCM cross-lingues. Les modèles acoustiques ont été obtenus à partir d'apprentissage non-supervisé sur trainQ10 ou trainQ10+trainQ11. Le TEM (%) a été mesuré sur le jeu de développement devQ10.

des PCM et des MMs. Les résultats obtenus sont montrés dans le Tableau 3.2. Pour les tests rapportés, l'ajout des données est avantageux pour l'estimation des PCM, mais ce n'est pas toujours le cas pour l'estimation des MMs. La meilleure performance a été obtenu en utilisant 316 heures de données pour estimer le PCM et 143 heures de données pour estimer le MM. Cette combinaison induit à un taux d'erreur de 27.8%, soit une perte de seulement 0.5% par rapport au système supervisé construit à partir de 72 heures de données (voir le Tableau 3.1).

Les modèles PCM non-supervisés on été comparés aux modèles PCM cross-lingues. Un PCM entraîné sur 650 heures des données audio en Anglais et un PCM entraîné sur 600 heures des données audio en Français ont été utilisés pour comparaison. Ces deux modèles ont été obtenus à partir d'une apprentissage supervisé. Ils ont été comparés au PCM entraîné sur 316 heures des données audio en Portugais Européen, obtenu via apprentissage supervisé. Les trois modèles ont la même architecture. Chacun des ces PCM a été utilisé pour extraire des attributs acoustiques pour des données audio en Portugais, desquelles des modèles acoustiques ont été estimés de façon non-supervisé. Les résultats de reconnaissance sur le jeu de développement sont montrées dans le Tableau 3.3. Même si entraîné sur presque la moitié de la quantité des données, le PCM non-supervisé Portugais présente des performances comparables au PCM supervisé Français et obtient des meilleurs résultats que le PCM supervisé Anglais. Le TEM obtenu (27.3%) est équivalent à ce obtenu avec un modèle supervisé obtenu à partir de 72 heures de données (27.2%).

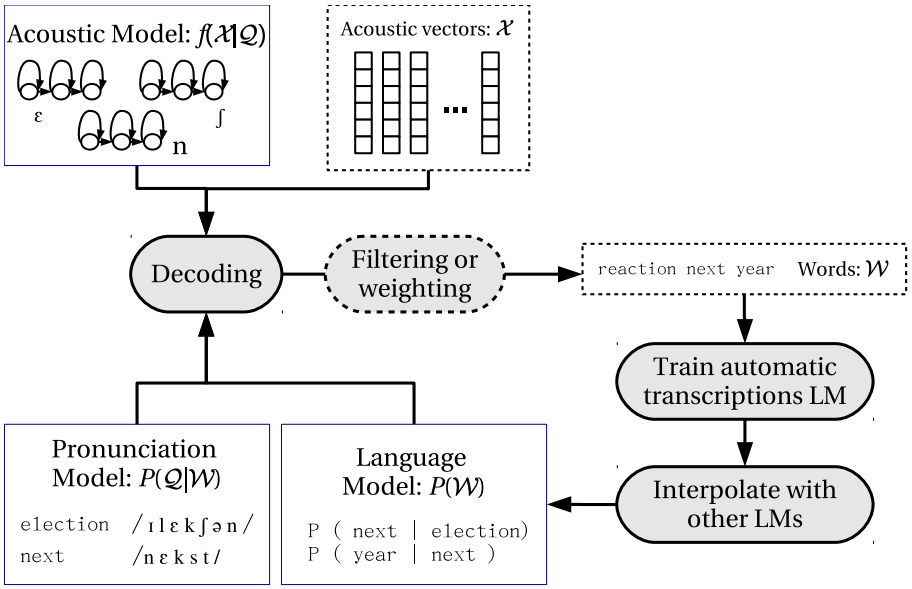

FIGURE 4.1 – Schéma de l'apprentissage non-supervisé des modèles de langue.

4 Apprentissage non-supervisé de modèles de langue

Les transcriptions audio ne sont pas obligatoire pour l'estimation des modèles de langue, mais il est bien connu qu'elles y jouent un rôle crucial et sont capables d'améliorer considérablement la performance de ces modèles. Inspiré par les travaux réalisés sur la modélisation acoustique, l'apprentissage non-supervisé des modèles de langue a été également étudiée dans cette thèse. Le principe est illustré dans la Figure 4.1. Un système de reconnaissance existant est utilisé pour générer des transcriptions automatiques, qui sont ensuite utilisées pour l'estimation ou l'adaptation des modèles de langue.

À nouveau, le principal problème de cette approche est l'effet bruit induit par des erreurs de reconnaissance. Des différentes méthodes basées sur les mesures de confiances ont été établies afin de réduire l'effet de ces erreurs :

1-meilleur Approche de référence. Pas de traitement des transcriptions automatiques.

1-meilleur filtré Les mots ayant un score de confiance au-dessous d'un seuil sont substitués par un symbole de mot inconnu.

1-meilleur pondéré Les comptes n-grammes sont obtenus à partir du produit des scores de confiance des mots.

Treillis Les comptes n-grammes sont obtenus à partir des treillis de décodage via l'algorithme *forward-backward*.

Réseau de confusion Les comptes n-grammes sont obtenus à partir des réseaux de confusion en multipliant les scores des liens.

Réseau de confusion filtré Comme avant, mais avec un filtre de plus.

Afin de se bénéficier des comptes fractionnaires obtenus par les méthodes '1-meilleur pondéré', 'Treillis' et 'Réseau de confusion' (toutes les deux), une extension de la méthode de lissage de Kneser-Ney a été proposé dans le cadre de cette thèse. Les modèles de langue

Modèle de langue	devQ10	testQ10	testQ11	testQ10 + testQ11	Global
LM_10src (référence)	31.36	25.77	31.29	28.57	29.47
+ 1-meilleur	31.06	25.85	31.25	28.59	29.38
+ 1-meilleur pondéré	**30.86**	**25.68**	**31.12**	**28.44***	**29.22***
+ 1-meilleur filtré	**30.87**	**25.69**	31.26	28.51*	29.27*
+ treillis	30.98	25.90	31.23	28.60	29.37
+ réseau de confusion	30.97	25.89	31.22	28.59	29.36
+ réseau de confusion filtré	31.17	25.71	31.19	28.49*	29.35
+ manuelle (trainQ10)	30.68	25.48	30.92	28.24	29.02

Table 4.1 – Résultats de reconnaissance obtenus avec des composantes de modèles des langues estimés à partir des transcriptions automatiques et interpolées au modèle de base LM_10src.

'1-meilleur' et '1-meilleur filtré' ont été obtenus à partir de la méthode de lissage Kneser-Ney standard. Pour chaque méthode, les transcriptions automatiques ont été utilisées pour estimer une composante de modèle de langue. À chaque fois, cette composante a été interpolée au ML de base, estimé sur des textes provenant de sources écrites (journaux, Web, blogs). Les sources écrites contiennent environ 640 millions de mots, tandis que les transcriptions automatiques ne font que 1 million. Ces modèles de langue ont été utilisés pour décoder le jeu de développement en utilisant un modèle acoustique entraîné de façon non-supervisée avec les mêmes données audio. Les résultats obtenus avec ces tests sont montrés dans le Tableau 4.1.

Les gains obtenus avec la modélisation non-supervisée de la langue sont bien plus faibles que ceux obtenus pour la modélisation acoustiques. Au mieux, un gain absolue de 0.25% a été obtenu en utilisant la méthode '1-meilleur pondéré' proposée en comparaison avec la référence (29.22% vs. 29.47%). Bien que faible, les gains apportés sont complémentaires à ceux obtenus avec la modélisation acoustique non-supervisée. Cela dit, si un certain jeu de données a été automatiquement transcrit pour l'apprentissage des MAs, le coût supplémentaire pour utiliser ces transcriptions pour l'apprentissage des MLs est négligeable.

L'approche d'apprentissage non-supervisé a été aussi utilisée avec des modèles de langue neuronaux (MLN). Différemment des modèles standards à repli, les MLNs possèdent une structure bien claire et représentent les mots sur un espace continu. Ces deux caractéristiques permettent aux MLNs de généraliser mieux, conduisant à des gains de performance considérables par rapport aux modèles à repli. Les expériences en l'apprentissage non-supervisé des modèle neuronaux ont été faites comme suit. D'abord, un MLN de référence a été estimé sur des sources écrites. Ce modèle a été ensuite adapté aux transcriptions automatiques (ou manuelles pour référence). Chaque modèle neuronal est alors interpolé à un modèle à repli de base. Le modèle final est alors utilisé pour le décodage. Le Tableau 4.2 montre les résultats obtenus avec cette approche. Au mieux, l'apprentissage non-supervisé des modèles neuronaux conduit à des gains absolues de 0.17% par rapport au modèle de référence (28.52% vs. 28.69%). Ici, le même remarque fait auparavant est valable : même si les gains sont relativement petits, l'usage de telle approche peut se faire justifier lorsque des données audio ont déjà été transcrites automatiquement.

Modèle de langue neuronal	devQ10	testQ10	testQ11	testQ10+ testQ11	Global
LM_10src (baseline)	30.68	25.21	30.22	27.75	28.69
→ *Non-upervisé* :					
trainQ10+trainQ11	30.35	**25.05**	**30.19**	**27.66**	**28.52**
trainRVE+trainQ10+trainQ11	**30.29**	25.14	30.25	27.73	28.55
→ *Supervisé* : trainQ10	29.89	24.82	29.88	27.39	28.19

Table 4.2 – Résultats de reconnaissance obtenus avec des modèles de langues neuronaux interpolés à des modèles à repli 4-grammes LM_10src. Le modèle de référence a été construit avec de sources écrites. Les modèles restants ont été obtenus après l'adaptation du modèle de référence aux transcriptions et jeux de données rapportés, en utilisant des méthodes d'apprentissage supervisé ou non-supervisé.

5 Adaptation et combinaison de modèles acoustiques

Les méthodes d'apprentissage non-supervisé discutées dans les dernières sections sont utiles lorsque une grande quantité de données audio est disponible pour la tâche cible. Pour certaines tâches, dites peu dotées, il se peut que la collection de données ne soit pas évidente. Dans ces cas, il est nécessaire d'utiliser des données provenant d'une tâche similaires. Par example, si la tâche cible est (disons) la reconnaissance de l'Anglais parlé au Moyen Orient, on peut s'en servir (disons) des données d'Anglais parlé en Grand Bretagne.

Dans ces situations, les modèles acoustiques peuvent être obtenus en regroupant tous les échantillons des données tel que s'ils provenaient d'une même source homogène. Même si ce n'est pas le cas, des bonnes performances peuvent être obtenus avec cette approche si les données d'entraînement et de tests ne sont pas éloignées. Des meilleurs niveaux de performance peuvent être atteints lorsque le modèle global (à échantillons regroupés) est adapté aux données de la tâche cible. Certaines méthodes d'adaptation, telle que l'adaptation maximum *a posteriori* (MAP) permettent d'ajuster les niveaux de pertinence entre le modèle global et les données d'adaptation. Intuitivement, on peut considérer que des niveaux de performance encore meilleurs pourraient être obtenus avec une méthode capable de permettre d'ajuster les niveaux de pertinence entre les données d'adaptation et chacune des sources des données. Dans cette thèse, nous avons proposé d'utiliser l'interpolation des modèles pour atteindre cet objectif.

La construction de modèles par biais de l'interpolation est une pratique courante dans la modélisation de la langue. Dans ce cas, un modèle est estimé indépendamment sur chacune des sources. Ensuite, ces composantes sont interpolés pour générer un seul modèle. Les coefficients d'interpolation sont estimés en sorte que le modèle final s'ajuste au mieux aux données d'adaptation. Une approche similaire a été proposé dans cette thèse pour les modèles acoustiques. Dans ce cas, les mélanges de Gaussiennes qui modélisent la sortie des étaits des modèles de Markov sont interpolés tel que illustré dans la Figure 5.1. L'interpolation est faite en fusionnant toutes les Gaussiennes et en ajustant leurs gains par rapport aux coefficients d'interpolations. Cette procédure conduit à une augmentation significative des paramètres des modèles. Pour éviter une augmentation de complexité de calcul en utilisant des modèles interpolés, un algorithme de réduction de mélange est utilisé.

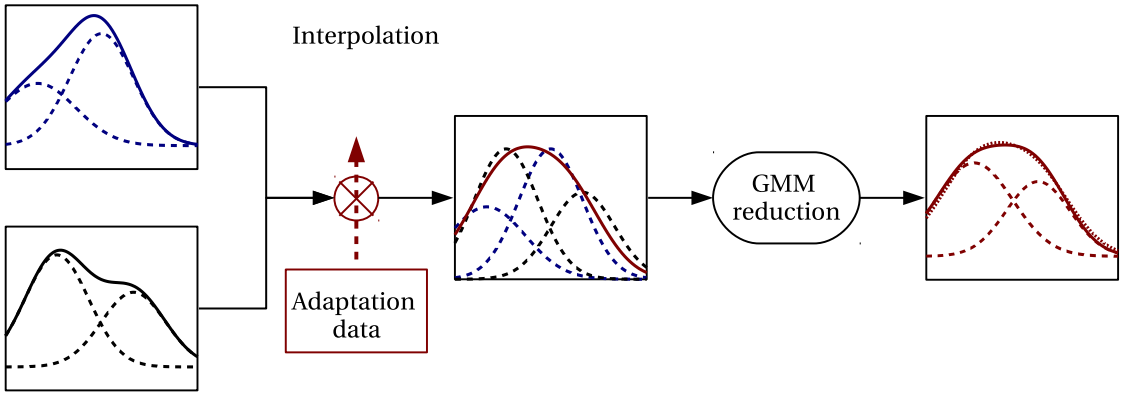

FIGURE 5.1 – Schéma de l'interpolation des mélanges de Gaussiennes. Les lignes discontinues correspondent aux Gaussiennes, tandis que les lignes solides aux mélanges. Après réduction, le mélange réduit est représenté par une ligne solide et le mélange original par une ligne pointillée.

Un algorithme de réduction des mélanges a été proposé dans cette thèse. Il est dérivé directement de l'estimation à maximum de vraisemblance en faisant l'hypothèse que chaque Gaussienne du modèle original représente un paquet des vecteurs acoustiques. Le résultat est un algorithme de regroupement mou. En comparaison avec un algorithme glouton et un algorithme basé sur du regroupement dur, l'approche proposée a conduit à des meilleurs résultats.

L'interpolation des modèles acoustiques a été évaluée sur un système de reconnaissance du Portugais et sur un système de reconnaissance de l'Anglais multi-accentué. Ici, nous commentons sur les résultats obtenus sur cette dernière tâche. Le système Anglais est destiné à la reconnaissance de six accents différents : américain (US), australien (AU), britannique (GB), mid-oriental (ME), nord-africain (NA) et indien (IN). La distributions des données n'est pas homogène dans le corpus d'entraînement. En particulier, les données US correspondent à plus que deux tiers du total (317h / 450h), tandis que le NA et IN ne contiennent que 8 et 9 heures de données.

Avec ce corpus, deux systèmes de base ont été développés par (Vergyri et al., 2010). Le premier est un système indépendant de l'accent ou le modèle acoustique a été obtenu en utilisant l'approche à échantillons regroupés. Le deuxième est un système qui est capable à identifier l'accent de l'émission de test et faire le décodage avec le modèle acoustique spécifique à cet accent. Dans ce cas, le modèles dépendants de l'accent ont été obtenus via adaptation MAP. Les résultats avec ces systèmes sont montrés dans les deux premières lignes du Tableau 5.1. Ensuite, nous avons proposé de créer les modèles spécifiques à chaque accent via interpolation. Dans ce cas, les coefficients ont été estimés pour en utilisant un jeu de données *held-out*. Pour effet de comparaison, un modèle à été généré en faisant l'interpolation avec des coefficients égales à chaque composante. Les résultats de ces approches sont montrés dans les lignes 3 à 5 du Tableau 5.1. Finalement, une autre approche a été évaluée. Plutôt que choisir un accent pour chaque émission, nous avons proposé d'interpoler les modèles acoustiques au-vol pour chaque émission ou pour chaque locuteur. Dans ce cas, les coefficients d'interpolations sont estimés sur les données de test elles-mêmes. Les résultats obtenus avec ces approches sont montrés aux deux dernières

Système	US	AU	GB	ME	NA	IN	Global	Moy
Indépendant de l'accent	14.34	11.92	12.84	15.90	26.47	39.28	16.07	20.12
ID-Accent (émission)	13.95	11.91	11.98	16.46	25.19	34.28	15.39	18.96
Interpolé (égal)	14.45	11.64	12.30	16.34	25.49	35.81	15.84	19.34
Interpolé (auto)	13.83	11.55	11.08	15.79	24.29	33.52	**15.05**	**18.34**
Interpolé réduit	13.75	11.87	11.45	15.79	24.89	33.95	15.11	18.62
Au-vol (locuteur)	14.11	11.37	11.65	15.63	24.24	33.18	15.27	18.36
Au-vol (émission)	14.06	11.18	11.30	15.86	24.24	33.69	15.22	18.39

Table 5.1 – Résultats de reconnaissance avec des différents systèmes pour chaque des six accents régionaux de l'Anglais. 'Global' correspond au TEM (%) sur tout les jeu de test, pendant que 'Moy' au TEM moyen quand chaque accent est pondéré également.

Système	US	AU	GB	ME	NA	IN	Global	Moy
Indépendant de l'accent	21.31	13.99	15.29	18.77	29.33	44.69	21.69	23.90
Adapté au `held-out`	20.69	13.02	14.39	19.15	28.58	41.11	20.90	22.82
Interpolé (auto)	20.42	13.35	13.30	17.25	27.33	40.84	**20.41**	**22.08**
Interpolé réduit	20.66	13.48	13.79	17.45	28.08	41.00	20.70	22.41
Au-vol (locuteur)	20.65	12.61	13.40	17.09	28.38	40.51	20.56	**22.11**

Table 5.2 – Résultats de reconnaissance avec des différents systèmes pour chaque des six accents régionaux de l'Anglais quand l'accent cible n'est pas représenté dans le modèle acoustique. 'Global' correspond au TEM (%) sur tout les jeu de test, pendant que 'Moy' au TEM moyen quand chaque accent est pondéré également. Ces résultats ne se comparent pas avec ceux du Tableau 5.1 étant donné que les conditions d'entraînement sont différentes.

ligne du Tableau 5.1. De manière générale, les modèles obtenus via interpolation conduisent à des meilleurs résultats que les modèles obtenus par regroupement d'échantillons ou par adaptation MAP. Cela reste valable même quand la réduction des mélanges est appliquée.

Les expériences rapportées ci-dessus ont fait l'usage de tout le corpus d'entraînement. Nous avons évalué des scénarios qui simulent des tâches peu-dotées plus extrêmes. Dans un 2ème scénario, nous avons considéré que seulement 2 ou 3 heures de données de l'accent cible étaient disponibles pendant l'entraînement. Ces données n'ont pas été utilisées pour l'estimation des modèles, mais pour l'estimation des coefficients d'interpolation ou pour l'adaptation MAP. Dans un 3ème scénario, nous avons considéré qu'aucune donnée était disponible pour l'entraînement. Les résultats de ces expériences sont rapportés au Tableau 5.2. La première ligne montre le modèles obtenus par échantillons regroupés, la référence. La deuxième partie du tableau montre les résultats issus du 2ème scénario. Ici, l'interpolation de modèles conduits aux meilleurs résultats. La troisième partie du tableau montre les résultats issus du 3ème scénario, où seul l'interpolation au-vol (ou le regroupement d'échantillons) est possible. Dans cette condition, l'interpolation se montre spécialement utile, conduisant à des gains importants par rapport à la référence, soit le système indépendant de l'accent.

6 Conclusions

Dans cette thèse, nous avons étudié des différentes approches dont l'objectif était de réduire les coûts de développement de systèmes de reconnaissance automatique de la parole. Une grande partie de ces coûts sont dus à l'effort humain nécessaire pour produire manuellement une **grande quantité** de transcriptions audio, qui sont fondamentales pour la construction de modèles acoustiques et pertinentes pour la construction des modèles de langue.

Pour des tâches dont les données sont **rares**, des efforts supplémentaires sont demandés pour produire les corpora d'entraînement et raffiner le système, augmentant le coût de développement. Parmi d'autres tâches, se distinguent la reconnaissance des langues et dialectes peu dotés. Deux axes de recherche ont été explorés dans cette thèse : 1) l'utilisation de méthodes d'apprentissage non-supervisés comme un moyen de réduire le besoin de transcriptions manuelles, et 2) l'utilisation d'interpolation des modèles acoustiques comme moyen de réduire le besoin de données audio spécifiques de la tâche ciblée.

Les approches non-supervisées ont été appliquées à la construction de trois des principales composantes des systèmes de RAP à l'état de l'art : les modèles acoustiques, le Perceptron multicouche utilisé pour extraire des attributs acoustiques et les modèles de langue. Plusieurs conclusions peuvent être tirées de ces travaux.

Des approches non-supervisées pour entraîner des modèles acoustiques peuvent être utilisés pour améliorer considérablement les performances des systèmes de RAP sans exiger des transcriptions manuelles. L'apprentissage non-supervisé des MAs peut être encore amélioré via l'application des méthodes de filtrage ou de pondération basés sur une mesure de confiance, visant à diminuer l'impact des erreurs de reconnaissance présentes dans les transcriptions automatiques. La stratégie d'entraînement joue également un rôle important. Nous avons observé que la manière comme les sous-ensembles des données d'apprentissage sont utilisés à chaque itération peut permettre d'éviter la propagation des erreurs de reconnaissance et d'améliorer considérablement l'efficacité du modèle.

Une nouvelle approche pour l'apprentissage non-supervisé des MAs a été proposé, à savoir, l'utilisation de plusieurs hypothèses de décodage (au lieu de la meilleure) pour guider l'estimation des paramètres des modèle acoustiques. Dans les expériences rapportées, l'utilisation de plusieurs hypothèses de décodage a apporté des meilleurs résultats que les approches standards, surtout lorsqu'une grande quantité d'erreurs est présenté dans les transcriptions automatiques utilisées pour l'apprentissage. L'utilisation de plusieurs hypothèses de décodage a été justifié théoriquement comme étant une approximation aux vraies transcriptions plus souple que l'approche standard (1-meilleur).

Dans cette thèse, nous avons également proposé d'étendre l'usage des méthodes non-supervisées pour la construction des Preceptron multicouche. Nous avons montré que les données audio non-transcrites peuvent aussi être utilisées pour estimer efficacement les paramètres des PCM appliquées à l'extraction d'attributs acoustiques. La méthode proposée a été utilisé pour créer des modèles acoustiques performants basée sur des attributs PCM d'une manière complètement non-supervisée. Les résultats obtenus sont compétitifs aux PCM cross-lingues. Un PCM entraîné sur 316 heures de données non-transcrites venant de la langue cible obtient des performances similaires à un PCM cross-lingue entraîné sur 600 heures de données transcrites.

Des transcriptions audio générées automatiquement ont également été explorés comme

un moyen d'améliorer les estimations des modèles de langue. En ajoutant une quantité relativement faible de transcriptions automatiques dans le corpus d'apprentissage des ML (seulement 3M de 639M mots), nous avons rapporté des gains de performance pour la reconnaissance d'émissions diffusés. Le travail expérimental réalisé a montré que l'apprentissage non-supervisé de MLs est une tâche bien plus difficile que l'apprentissage non-supervisé de MAs, comme observé précédemment par d'autres groupes de recherche. Néanmoins, nous avons obtenus des gains additifs en utilisant conjointement l'apprentissage non-supervisé des MLs et MAs. Cela dit, si un certain jeu de données a été automatiquement transcrit pour l'apprentissage des MAs, le coût supplémentaire pour utiliser ces transcriptions pour l'apprentissage des MLs est négligeable. Les méthodes non-supervisés ont été évaluées sur des modèles de langue n-grammes à repli (*backoff*) standards et des modèles de langue neuronaux. Les méthodes appliquées sur ces deux modèles ont conduit à des gains relatifs de performance similaires à l'égard de leurs modèles de référence. Cependant, les MLNs surpassent les modèles à repli en valeurs absolues. Les gains obtenus avec les deux types de modèles, à repli et neuronaux, ne sont pas complémentaires.

Inspiré par des pratiques appliquées couramment dans la modélisation de la langue, nous avons proposé l'utilisation de deux méthodes qui visent à prendre en considération la pertinence des différents sous-ensembles de données d'apprentissage pendant l'estimation des modèles : la pondération des données et l'interpolation des modèles de mélanges de Gaussiennes. Les approches théoriques proposées ont été empiriquement validées pour deux tâches : la reconnaissance des données diffusées du Portugais Européen et la reconnaissance de données multi-accentuées de l'Anglais.

Ces deux méthodes proposées ont été comparées théoriquement avec les approches d'apprentissage adaptative. En particulier, nous avons montré que la pondération des données peut être vue comme un type restreint d'apprentissage adaptative par locuteur (Anastasakos et al., 1996). L'interpolation de modèles et l'apprentissage adaptative par regroupements de locuteurs (Gales, 1998) présentent des similarités dans le sens que les deux méthodes utilisent de coefficients scalaire afin de pondérer plusieurs modèles. Néanmoins, les méthodes de pondération de données et d'interpolation visent à construire des modèles acoustiques pour une tâche donnée, tandis que les méthodes d'apprentissage adaptative visent à réduire la variabilité de la parole parmi les locuteurs.

Dans les expériences rapportées, la pondération des données et l'interpolation des mélanges de Gaussiennes ont obtenus des meilleures résultats de reconnaissance en comparaison avec les modèles de référence, à savoir ceux obtenus à partir d'échantillons regroupés et ceux obtenus à partir de l'adaptation maximum *a posteriori*. De plus, nous avons montré que des coefficients optimales peuvent être estimés automatiquement sur un petite quantité de données, même si des transcriptions ne sont pas disponibles. L'interpolation de modèles s'est avérée une solution particulièrement utile pour des tâches peu dotées. Dans les tests faits pour la reconnaissance des données multi-accentuées en Anglais, des meilleurs performances ont été obtenus avec l'interpolation de modèles lorsqu'une faible (seulement 2 à 3 heures) ou même nulle quantité de données d'un accent cible étaient disponibles pour l'apprentissage.

L'interpolation de mélanges de Gaussiennes telle que proposée engendre une augmentation des paramètres du modèle. Afin de revenir à des taux de complexités plus bas, un algorithme de réduction de mélange a été proposé et justifié théoriquement. La solution, basée sur une version restreinte de l'estimation à maximum de vraisemblance a générée

des meilleurs résultats comparé à d'autres algorithmes connus (Runnalls, 2007; Davis and Dhillon, 2007).

7 Références

Anastasakos, T., McDonough, J., Schwartz, R., and Makhoul, J. (1996). A compact model for speaker-adaptive training. In *Proc. International Conference on Spoken Language (ICSLP)*, volume 2, pages 1137–1140.

Davis, J. V. and Dhillon, I. (2007). Differential entropic clustering of multivariate Gaussians. In *Advances in Neural Information Processing Systems*, volume 19, pages 337–344. MIT Press.

Fousek, P., Lamel, L., and Gauvain, J.-L. (2008). Transcribing broadcast data using MLP features. In *Proc. Interspeech*, pages 1433–1436.

Gales, M. J. (1998). Cluster adaptive training for speech recognition. In *Proc. International Conference on Spoken Language (ICSLP)*, volume 1998, pages 1783–1786.

Runnalls, A. R. (2007). Kullback-leibler approach to gaussian mixture reduction. *IEEE Transactions on Aerospace and Electronic Systems*, 43(3) :989–999.

Vergyri, D., Lamel, L., and Gauvain, J.-L. (2010). Automatic speech recognition of multiple accented English data. In *Proc. Interspeech*, pages 1652–1655.