Développement d'une technologie générique pour la reconnaissance de la parole indépendante de la tâche *

Fabrice Lefèvre, Jean-Luc Gauvain et Lori Lamel

Groupe Traitement du Langage Parlé, LIMSI-CNRS, FRANCE
{lefevre,gauvain,lamel}@limsi.fr

RÉSUMÉ

This work addresses issues in speech recognition portability via the development of generic core speech recognition technology. First, genericity of large domain reference models (designed for a task covering a large number of acoustic and linguistic events) is assessed trough their performance on various independent tasks. Then, new techniques based on a multi-source training are presented aiming at enhancing the level of genericity of the large domain models. Finally, methods for transparent adaptation of generic models to a particular task are studied.

1. INTRODUCTION

En dépit des progrès constants observés cette dernière décennie, la reconnaissance automatique de la parole ne peut toujours pas être considérée comme un problème résolu. Les systèmes de reconnaissance sont habituellement développés dans le contexte d'une application (ou d'une langue) particulière et le transfert vers une nouvelle tâche se révèle généralement coûteux en moyens humains et en temps.

L'objectif des travaux présentés dans cet article est le développement d'une technologie *générique* pour la reconnaissance de la parole. Un système de transcription est générique lorsqu'il fonctionne correctement sur un large panel de tâches, allant de la reconnaissance de chiffres à celle de conversations téléphoniques, sans nécessiter de développement coûteux pour chacune des tâches.

Au travers du développement d'une technologie générique, ce travail aborde le problème de la portabilité des systèmes de reconnaissance selon deux axes : développer une technologie fonctionnant sur une large gamme de tâches sans information particulière sur ces tâches; explorer des méthodes d'adaptation rapides et peu coûteuses qui partant d'un système générique robuste améliorent ses performances sur une tâche particulière. Une première série d'expériences a pour objectif de se faire une idée du niveau de la généricité de modèles large domaine. Pour cela, ils sont évalués en conditions inter-tâches, i.e. en traitant des données spécifiques aux nouvelles tâches à l'aide d'un système de transcription développé pour une autre tâche. Nous avons choisi d'évaluer les performances des modèles acoustiques et linguistiques de la tâche de transcription de journaux télédiffusés (Broadcast News, BN) sur trois tâches représentatives : la reconnaissance de petits vocabulaires (*TI-digits*), le dialogue oral homme-machine (*ATIS*) et la dictée de textes lus ou énoncés spontanément (*Wall Street Journal*). La tâche BN est relativement générale, englobant une grande variété d'événements acoustiques et linguistiques, offrant ainsi une couverture raisonnable des tâches visées. De plus, nous disposons de suffisamment de données acoustiques et linguistiques pour cette tâche pour permettre l'estimation de modèles précis prenant en compte un grand nombre de locuteurs et de caractéristiques linguistiques.

Des méthodes visant à augmenter le niveau de généricité des modèles large domaine ont été étudiées. Une voie possible est d'avoir recours à un apprentissage multi-source. Deux approches utilisant des données d'apprentissage spécifiques à la tâche ont été comparées : par regroupement (les modèles de référence sont adaptés avec l'ensemble des données des tâches candidates simultanément) et séquentielle (les modèles sont adaptés avec les données d'une tâche, les modèles résultants avec les données d'une autre tâche et ainsi de suite). L'objectif de l'apprentissage multi-source est d'obtenir des modèles génériques dont les performances sont comparables aux modèles spécifiques pour toutes les tâches considérées.

Une autre voie explorée est l'utilisation de méthodes transparentes permettant l'adaptation dynamique des modèles génériques à une nouvelle tâche. Actuellement, le transfert d'un système de reconnaissance vers une nouvelle tâche ou une nouvelle langue requiert la disponibilité d'une quantité suffisante de données d'apprentissage. Lorsque l'on s'intéresse à un nouveau domaine, il n'existe généralement pas de transcriptions détaillées des données et la production de données transcrites représente un coût important. L'approche testée consiste à utiliser les composants d'un système générique afin de transcrire automatiquement les données d'apprentissage spécifiques à la nouvelle tâche. Ces données permettent ensuite d'améliorer les performances des modèles génériques sur une tâche particulière.

Dans la section suivante le système de transcription du LIMSI est présenté. Dans la section 3, les expériences de reconnaissance inter-tâches nous permettent d'obtenir une première estimation du niveau de généricité des modèles. Les techniques d'apprentissage multi-source par regroupement et séquentielle sont présentées et testées dans la section 4. Puis des techniques d'adaptation transparente sont évaluées pour améliorer les performances en conditions inter-tâches (section 5).

* Le travail présenté dans cet article a été partiellement financé par la Commission Européenne dans le cadre du projet CORETEX.

2. DESCRIPTION DU SYSTÈME DE TRANSCRIPTION

Le système de reconnaissance de la parole du LIMSI pour la transcription d'informations télédiffusées utilise des modèles de Markov cachés à densités continues avec des mélanges de gaussiennes pour la modélisation acoustique et des modèles linguistiques de type n-grammes estimés sur de grands corpus de textes. Chaque modèle phonétique dépendant du contexte est un HMM gauche-droit à états liés obtenus grâce à un arbre de décision.

La reconnaissance est opérée en trois étapes : 1) génération d'une hypothèse initiale, 2) adaptation des modèles et génération d'un graphe de mots, 3) adaptation des modèles et génération de l'hypothèse finale. L'hypothèse initiale est utilisée pour l'adaptation des modèles acoustiques à l'aide de la technique du MLLR [7] préalablement à la génération du graphe de mots. Un modèle de langage de type 3-grammes avec back-off est utilisé lors des deux premières étapes. L'hypothèse finale est engendrée à partir d'un modèle 4-grammes et des modèles acoustiques adaptés lors de la seconde étape.

Dans le système de référence, les modèles acoustiques ont été entraînés avec environ 150 heures de données audio extraites du corpus Hub4 Broadcast News du LDC [4]. Des modèles acoustiques dépendants du genre sont estimés par une adaptation de type MAP [3] à partir des modèles initiaux pour la parole large bande et téléphonique. Le jeu de modèles acoustiques comprend 28.000 phones en contexte, inter-mots et dépendants de la position dans le mot, comptant pour 11.700 états liés et environ 360k gaussiennes [2].

Les modèles de langage sont obtenus par interpolation de modèles entraînés sur des textes d'origines diverses : journaux, câbles (*newswires*), transcriptions commerciales et transcriptions des données d'apprentissage. Le vocabulaire de reconnaissance contient 65.120 mots, avec en moyenne 1.2 prononciations par mot. Les prononciations reposent sur un ensemble de 48 phones, comprenant 3 unités pour représenter les silences, les hésitations (*euh*, *hum*...) et les bruits de respiration.

Ce système a obtenu un taux d'erreur de 17,1% lors de l'évaluation NIST en 1999 avec la contrainte de fonctionner en moins de 10 fois le temps réel. Il peut transcrire des données télédiffusées sans restriction avec un taux d'erreur aux alentours de 20% [2].

3. RECONNAISSANCES INTER-TÂCHES

Dans l'objectif de développer un moteur de reconnaissance de la parole générique, nous commençons par évaluer le système en conditions inter-tâches, i.e. en traitant des données spécifiques à une tâche à l'aide d'un système développé pour une autre tâche.

Pour la tâche de reconnaissance à petit vocabulaire, les expériences sont menées sur le corpus TI-digits [8] (17k énoncés de 225 locuteurs). Le vocabulaire comprend les chiffres de '0' à '9', plus 'oh' (utilisé pour zéro en anglais). La base de données contient 7 heures de parole, réparties à parts égales entre les ensembles d'apprentissage et de test. La parole, enregistrée dans un environnement calme, est de très bonne qualité. Les taux d'erreur les plus bas rapportés sur cette tâche se situent vers 0,2-0,3%. Le système que nous avons développé pour cette tâche n'a que 108 modèles phonétiques contextuels du fait de la faible couverture phonémique des chiffres. Le modèle linguistique pour cette tâche est une grammaire autorisant n'importe quelle séquence d'au plus 7 chiffres. Le taux d'erreur de notre système dépendant de la tâche est de 0,4%.

La tâche ATIS [1] d'information sur le transport aérien du DARPA est retenue comme représentante d'une tâche de dialogue homme-machine, et l'évaluation est réalisée sur les données de 1994. Les données de test comprennent 1h30 de parole venant de 24 locuteurs et enregistrées à l'aide d'un micro casque. Environ 40 heures de parole sont disponibles pour l'apprentissage. Les taux d'erreur en mots lors de l'évaluation de 1994 étaient pour la plupart compris entre 2,5% et 5%, ce que nous considérerons comme l'état de l'art pour cette tâche. Les modèles acoustiques utilisés dans notre système comprennent 1641 phones dépendants du contexte avec 4k états markoviens indépendants. Un modèle de langage 3-grammes avec back-off a été estimé sur les transcriptions des énoncés d'apprentissage. Le lexique comprend 1300 mots parmi lesquels les entités de la base de données (villes, aéroports, services...). Lorsqu'elles sont décrites par plusieurs mots, les entités sont considérées dans le lexique comme une seule entrée sous forme de mot composé. Le taux d'erreur de ce système est de 4,1%.

La tâche de dictée de textes est représentée par le corpus *Wall Street Journal* [10], et les conditions de test correspondent à l'évaluation ARPA Hub3 de 1995. Les données acoustiques d'apprentissage ont été prononcées par 355 locuteurs pour un total de 100 heures de parole. Le test Hub3 consiste en parole lue en studio par 20 locuteurs pour un total de 45 minutes. Le meilleur résultat, lors de l'évaluation, était 6,6% [11]. Une expérience contrastive est menée avec les données de l'évaluation WSJ93 Spoke 9 qui comprend 200 phrases dictées de façon spontanée par des journalistes [5]. Le meilleur résultat reporté lors de l'évaluation Spoke 9 de 1993 était de 19,1%, toutefois des taux plus faibles ont été rapportés depuis sur des ensembles de test comparables (par exemple 14,1% pour l'évaluation Spoke 9 de 94). 21k modèles dépendants du contexte et de la position dans le mot ont été entraînés pour le système WSJ, avec 9k états markoviens indépendants. Un vocabulaire de 65k mots a été sélectionné et un modèle 3-grammes avec back-off obtenu par interpolation de modèles entraînés sur différents ensembles de données (transcriptions des données d'apprentissage et textes de journaux). Ce système WSJ a un taux d'erreur de 7,6% sur les données lues et de 15,3% sur les données spontanées.

Pour la tâche de référence, c'est à dire la transcription de journaux télédiffusés, nous suivons les conditions de l'évaluation ARPA Hub4E de 1998 [9]. Les données d'apprentissage contiennent 150 heures de parole provenant d'émissions de télévisions et de radios nord-américaines. Le système du LIMSI a obtenu un taux d'erreur de 13,6% lors de l'évaluation de 1998.

Trois ensembles d'expériences sont rapportées dans le tableau 1. Le premier comprend les expériences inter-tâches réalisées en utilisant les modèles acoustiques et linguistiques de BN pour décoder les données de test des autres tâches. Le second jeu d'expériences utilise des modèles mixtes : les modèles acoustiques BN avec les modèles lin-

TAB. 1 – *Reconnaissances inter-tâches - Taux d'erreur sur les mots (%) pour les ensembles de test de BN, TI-digits, ATIS, WSJ lu et spontané après reconnaissance avec trois configurations différentes : (gauche) modèles acoustiques et linguistiques de BN, (milieu) modèles acoustiques BN combinés avec modèles linguistiques de la tâche et (droite) modèles acoustiques et linguistiques de la tâche.*

Tâche	*Modèles acoustiques & linguistiques BN*	*Modèles acoustiques BN & linguistiques de la tâche*	*Modèles acoustiques & linguistiques de la tâche*
BN	13.6	13.6	13.6
TI-digits	17.5	1.7	0.4
ATIS	20.8	4.7	4.1
WSJ lu	11.6	9.0	7.6
WSJ spontané	12.1	13.6	15.3*

* Modèles de WSJ lu

guistiques dépendants de la tâche. Nous pouvons aussi observer, en comparant les conditions dépendantes de la tâche (tableau 1, droite) et mixtes (tableau 1, milieu), que les modèles acoustiques BN sont relativement génériques. L'utilisation des modèles de langage spécifiques à la tâche, permet de mettre en évidence, pour les corpus TI-digits et ATIS, que l'écart de performance est principalement imputable à des différences linguistiques. Pour WSJ, les modèles linguistiques sont plus proches de ceux de BN et la réduction relative du taux d'erreur de 20%. Dans le cas de la dictée de textes spontanée, on observe même une amélioration des performances en utilisant les modèles linguistiques de BN. Cette amélioration peut être attribuée à une meilleure modélisation des phénomènes spontanés (comme les respirations ou les hésitations) dans les modèles BN.

4. APPRENTISSAGE MULTI-SOURCE

Dans cette section, nous étudions des méthodes pour améliorer la généricité des modèles au moyen d'un apprentissage multi-source. Ceci peut être réalisé de différentes façon – en groupant les données, en interpolant les modèles ou par une adaptation directe ou progressive. Notre objectif est d'obtenir des performances avec les modèles génériques qui soient comparables aux résultats des modèles spécifiques à la tâche pour chacune des tâches considérées.

La méthode la plus simple consiste à regrouper les données de toutes les tâches. Cet ensemble de données est alors utilisé pour adapter les modèles BN (les données de BN ne sont pas prises en compte). Au lieu de grouper les données, une méthode en plusieurs étapes peut être envisagée. Dans ce cas les modèles sont adaptés séquentiellement jusqu'à ce que les données de toutes les tâches aient été utilisées. Dans le travail présenté ici, les modèles acoustiques BN sont adaptés d'abord selon la séquence WSJ lu, ATIS et TI-digits puis selon la séquence TI-digits, ATIS et WSJ lu. Pour ces expériences seule l'adaptation MAP (en condition supervisée) a été appliquée.

Les résultats pour les deux types de ré-apprentissage sont donnés dans le tableau 2 en utilisant un lexique et des modèles linguistiques spécifiques à la tâche. De façon prévisible, l'ordre retenu pour l'adaptation séquentielle influence fortement les performances des modèles. De façon générale, le ré-apprentissage multi-source permet d'augmenter les performances des modèles de référence, les taux d'erreur sont comparables (BN et TI-digits) ou meilleurs (ATIS et WSJ) que ceux obtenus avec les modèles spécifiques des tâches.

Les résultats obtenus sur le test WSJ spontané confirme une augmentation du niveau de généricité des modèles de référence. L'apprentissage multi-source ne prenant en compte aucune donnée spécifique à cette tâche, l'amélioration des performances vient de la mise en commun de caractéristiques propres à WSJ lu d'une part et à BN d'autre part.

5. ADAPTATION INCRÉMENTALE NON-SUPERVISÉE

Une seconde voie pour faciliter le transfert d'un système vers une nouvelle tâche consiste à réaliser une adaptation transparente des modèles génériques permettant une mise à jour dynamique des modèles avec de nouvelles données. Dans cette section, nous étudions une procédure d'adaptation non-supervisée et incrémentale utilisant une système de reconnaissance pour transcrire des données audio brutes qui servent à l'adaptation des modèles. De nouvelles données sont ensuite transcrites avec les modèles adaptés permettant une nouvelle passe d'adaptation. Ainsi de suite, tant que de nouvelles données sont disponibles.

La procédure est appliquée sur le tâche ATIS. Les modèles acoustiques sont adaptés par une combinaison de MLLR et MAP [6]. L'adaptation des modèles de langage est obtenue par une interpolation entre le modèle 3-grammes de BN et un modèle 3-grammes estimé sur les transcriptions automatiques. Les résultats sont présentés dans le tableau 3 en fonction de la quantité de données d'adaptation. Environ un tiers (15 heures) du corpus d'apprentissage d'ATIS a été transcrit à l'aide des modèles BN, et les 26 heures restantes ont été transcrites en utilisant les modèles BN adaptés avec les 15 premières heures. L'adaptation des modèles conduit à une réduction relative du taux d'erreur de 74% (15,3% absolu). Ainsi, en dépit d'un taux d'erreur initial élevé, l'adaptation transparente des modèles de référence permet d'atteindre des performances comparables aux modèles spécifiques à la tâche.

6. CONCLUSIONS

Dans cet article, un aperçu de la généricité des systèmes actuels de reconnaissance de la parole a été obtenu, en testant un système relativement large domaine sur des données de trois tâches de complexités variées. Les modèles venant de la tâche BN ont été choisis comme modèles de référence car ils couvrent un large éventail de conditions acoustiques et linguistiques. Ces modèles acoustiques sont relativement indépendants de la tâche dans la mesure où ils conduisent à une faible dégradation des performances

TAB. 2 – *Apprentissage multi-source - Taux d'erreur sur les mots sur les ensembles de test de BN, TI-digits, ATIS, WSJ lu et spontané après reconnaissance avec trois configurations différentes, incluant toutes un lexique et des modèles linguistiques spécifiques à la tâche combinés avec des modèles acoustiques BN adaptés : (gauche) apprentissage multi-source par regroupement des données, (milieu) apprentissage multi-source par adaptation séquentielle et (droite) adaptation supervisée dépendante de la tâche. L'évaluation BN a été réalisée en conditions 10xTR, le taux d'erreur de référence est alors de 14.2%.*

	Modèles acoustiques			
	Multi-source			*Spécifiques à la tâche*
	Regroupement	*Séquentielle*		
Tâche		BN→WSJ→ATIS→TI	BN→TI→ATIS→WSJ	
BN (10×TR)	14.9	15.8	15.3	14.2
TI-digits	0.7	0.6	1.3	0.4
ATIS	3.1	3.6	3.2	4.1
WSJ lu	6.7	7.4	6.7	7.6
WSJ spontané	11.8	12.4	11.5	15.3*

* Modèles de WSJ lu

TAB. 3 – *Adaptation incrémentale non-supervisée - Taux d'erreur sur les mots (%) pour ATIS en fonction de la quantité de données utilisée pour l'adaptation non-supervisée : (gauche) quantité et taux d'erreur en mots des données d'adaptation, (droite) taux d'erreur sur le test après adaptation des modèles acoustiques seuls (1ère colonne) du modèle de langage seul (2ème colonne) et des deux (3ème colonne). Les 26h de données sont transcrites à l'aide des modèles acoustiques adaptés avec 15h de données.*

Données d'adaptation		*Adaptation non-supervisée des modèles BN*		
Quantité	*Taux d'erreur*	*Acoustiques*	*Linguistique*	*Acoustiques & Linguistique*
0	-	20.8	20.8	20.8
15h	27.8	13.5	8.7	6.9
26h	14.0	8.7	6.8	5.5

comparées aux taux d'erreur obtenus avec des modèles acoustiques et linguistiques spécifiques à la tâche.

Des expériences ont été menées pour améliorer la généricité des modèles acoustiques. Il a été montré que l'apprentissage multi-source améliore la précision des modèles génériques, permettant d'obtenir des résultats comparables ou meilleurs que ceux obtenus avec des modèles dépendants de la tâche.

L'adaptation incrémentale non-supervisée des modèles acoustiques et linguistiques BN se révèle très efficace pour la transcription de dialogues spontanés. Une réduction relative du taux d'erreur de 74% a pu être obtenue sur la tâche ATIS. Le taux d'erreur du système adapté utilisant uniquement des données audio brutes est de 5,5%, à comparer au 4,1% obtenus par le système dédié à la tâche entraîné sur la même quantité de données annotées manuellement.

RÉFÉRENCES

[1] D. Dahl *et al.* Expanding the scope of the atis task: The atis-3 corpus. In *Proc. of the ARPA Spoken Language Systems Technology Workshop*, pages 3–8, March 1994.

[2] J.-L. Gauvain and L. Lamel. Fast decoding for indexation of broadcast data. In *ICSLP*, volume 3, pages 794–798, Beijing, 2000.

[3] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio Processing*, 2(2):291–298, 1994.

[4] D. Graff. The 1996 broadcast new speech and language-model corpus. In *1997 DARPA Speech Recognition Workshop*, Chantilly, 1997.

[5] F. Kubala *et al.* The hub and spoke paradigm for csr evaluation. In *Proc. of the ARPA Spoken Language Systems Technology Workshop*, pages 9–14, 1994.

[6] F. Lefevre, J.-L. Gauvain, and L. Lamel. Genericity and adaptability issues for task-independent speech recognition. In *ISCA International Tutorial and Research Workshop on Adaptation Methods for Speech Recognition*, Sophia-Antipolis, 2001.

[7] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, 9:171–185, 1995.

[8] R.G. Leonard. A database for speaker-independent digit recognition. In *ICASSP*, 1984.

[9] D.S. Pallett *et al.* 1998 broadcast news benchmark test results. In *Proc. of the DARPA Broadcast News Workshop*, pages 5–12, 1999.

[10] D.B. Paul and J.M. Baker. The design for the wall street journal-based csr corpus. In *ICSLP*, Banf, 1992.

[11] P.C. Woodland *et al.* The htk large vocabulary recognition system for the 1995 arpa h3 task. In *Proc.of ARPA Speech Recognition Workshop*, 1996.