Transcription de la parole conversationnelle

J.-L. Gauvain, G. Adda, L. Lamel, F. Lefèvre et H. Schwenk

Groupe Traitement du Langage Parlé
LIMSI-CNRS, BP 133
91403 Orsay Cedex, FRANCE
{gauvain,gadda,lamel,lefevre,schwenk}@limsi.fr

RÉSUMÉ. *Cet article décrit le développement d'un système de reconnaissance de la parole conversationnelle, à partir d'un système à l'état de l'art pour la transcription d'émissions d'information. Nous décrivons les principales améliorations apportées aux modèles acoustiques, aux modèles linguistiques et au décodeur. Pour la modélisation acoustique, nos travaux ont porté sur l'introduction d'une normalisation par locuteur, le recours à des techniques d'apprentissage adaptatif et d'apprentissage discriminant, et une meilleure prise en compte des variantes de prononciation. Pour la modélisation linguistique, la principale difficulté vient de la faible quantité de données d'apprentissage disponible. Nous introduisons deux techniques permettant de diminuer l'impact de cette situation sur les performances du système : la sélection de textes de nature conversationnelle et un modèle représentant les mots dans un espace continu. La transcription est obtenue en effectuant un décodage par consensus sur un treillis de mots. Ces améliorations ont permis de réduire le taux d'erreur de 51% à 21%.*

ABSTRACT. *This paper describes the development of a speech recognition system for the processing of conversational speech, starting with a state-of-the-art broadcast news transcription system. We identify major changes and improvements in acoustic and language modeling, as well as decoding, which are required to achieve good performance on conversational speech. Some major changes on the acoustic side include the use of speaker normalizations (VTLN and SAT), a better pronunciation modeling and the use of discriminative training (MMIE). On the linguistic side the primary challenge of the limited amount of language model training data is addressed through the use of a data selection technique, and a smoothing technique based on a neural network language model. At the decoding level, lattice rescoring and minimum word error decoding are applied. On the development data, the improvements yield an overall word error rate of about 21% whereas the original BN transcription system had a word error rate of 51% on the same data.*

MOTS-CLÉS : *parole conversationnelle, modélisation acoustique, modélisation linguistique*

KEYWORDS: *conversational speech, acoustic modeling, language modeling*

1. Introduction

Au LIMSI nous travaillons sur la transcription d'émissions d'information (radio et télévision) dans plusieurs langues depuis 1996. Plus récemment, nous avons abordé le problème de la transcription de la parole conversationnelle. La transcription de conversations est une tâche bien plus difficile que la transcription d'émissions d'information, difficulté due principalement au caractère spontané de la parole conversationnelle. Dans cet article nous décrivons les travaux conduits au LIMSI pour faire évoluer un système de transcription d'émissions d'information (*Broadcast News*, BN) vers un système de transcription de conversations. Cette tâche est depuis plusieurs années au coeur des campagnes annuelles d'évaluation de systèmes de reconnaissance de parole organisées par le NIST, utilisant la famille des corpora SwitchBoard (SWB) collectés par le LDC (Godfrey *et al.*, 1992). Ces évaluations ont permis de mettre en évidence les principales difficultés posées par le traitement automatique de la parole conversationnelle (Hain *et al.*, 1999; Ljolje & *et al*, 2000; Stolcke & *et al*, 2000; Matsoukas *et al.*, 2002).

Le système de transcription de parole conversationnelle utilise les mêmes composants que le système de transcription d'émissions d'information. Les ajouts principaux sont : la normalisation de la longueur du conduit vocal (VTLN), une adaptation MLLR contrainte et une adaptation MLLR non contrainte avec plusieurs classes de régression, un apprentissage adaptatif, un apprentissage discriminant, l'utilisation de probabilités de prononciation, un modèle de langage neuronal, et un décodage par consensus. Certaines de ces techniques qui s'étaient révélées peu efficaces sur les données BN, en particulier la normalisation VTLN et les probabilités de prononciation, s'avèrent très utiles pour la parole conversationnelle.

Après une description succincte du système de transcription d'émissions d'information (section 2) qui constitue le point de départ pour nos développements, nous décrivons les changements effectués pour traiter la parole conversationnelle. Ces changements concernent les modèles acoustiques (section 3), le modèle linguistique (section 4) et la procédure de décodage (section 5). Nous précisons l'impact de ces améliorations sur le taux d'erreur.

2. Système de référence

Le système de transcription BN du LIMSI repose sur deux composants principaux : un segmenteur audio et un décodeur lexical (Gauvain *et al.*, 2002). La segmentation audio est effectuée de manière itérative, par un algorithme de segmentation-agglomération utilisant des mélanges de gaussiennes. Le résultat est un ensemble de segments acoustiquement homogènes correspondant aux tours de parole des locuteurs intervenant dans le document. Le décodeur lexical utilise des modèles de Markov cachés avec densités de probabilité continues (sommes pondérées de gaussiennes) pour les modèles acoustiques, et des statistiques n-grammes obtenues sur de grands corpora de textes pour le modèle linguistique. Les modèles de Markov cachés représentent des

allophones contextuels avec une structure gauche-droite à états liés. Ils modélisent des séquences de trames centisecondes représentées par 39 composantes : 12 coefficients cepstraux (PLP) et le logarithme de l'énergie à court-terme, avec leurs dérivées d'ordre 1 et 2.

Le décodage en mots est effectué en trois passes. La première passe produit une hypothèse qui est utilisée pour réaliser l'adaptation MLLR (Leggetter & Woodland, 1995) non supervisée des modèles acoustiques. Les modèles adaptés sont utilisés dans la seconde passe pour générer un graphe de mots. Ces deux passes utilisent un modèle de langage trigramme. L'hypothèse finale est générée avec un modèle quadrigramme et les modèles acoustiques adaptés lors de la seconde passe. L'ensemble du décodage est effectué en moins de 5 fois le temps réel[1]. La première passe utilise un jeu d'allophones représentant environ 5500 contextes avec 6300 états liés et 16 gaussiennes par état. Les passes 2 et 3 utilisent des modèles plus gros, représentant 11000 contextes phonétiques avec 11700 états liés, et respectivement 16 et 32 gaussiennes par état.

Le regroupement des états est réalisé en créant un arbre de décision pour chaque état de chaque phonème de façon à maximiser la vraisemblance des données d'apprentissage pénalisée par le nombre d'états liés. Nous utilisons un ensemble de 184 questions relatives à la position du phonème dans le mot, et aux caractéristiques acoustiques du phonème et de ses voisins immédiats.

Pour l'anglais-américain, le système de transcription BN a un taux d'erreur sur les mots inférieur à 20%[2]. Ce système a été adapté à cinq autres langues (Allemand, Arabe, Espagnol, Français et Mandarin) avec des performances comparables (Gauvain & Lamel, 2003).

Considérant le niveau de développement de nos modèles BN, il paraissait intéressant d'évaluer ce système sur de la parole conversationnelle avant toute modification. Cette évaluation a été menée sur les données de test de l'évaluation NIST Hub5 2001 (Eval01). Le taux d'erreur initial avec les modèles BN était de 51%. En utilisant les transcriptions de données conversationnelles (corpus SWB du LDC) pour construire le modèle de langage, le taux d'erreur est réduit à 47% (le vocabulaire n'a pas été modifié car le taux de mots hors-vocabulaire reste inférieur à 1%). En combinant ce modèle de langage avec les modèles acoustiques estimées sur les données SWB, on obtient un taux d'erreur de 36%. Ces résultats montrent qu'une part importante de l'écart entre les modèles BN et les données SWB réside dans la modélisation acoustique, et que réestimer les modèles BN sur les données SWB en appliquant les techniques développées pour les données BN n'est pas suffisant pour obtenir des performances acceptables sur des données conversationnelles.

1. Il faut donc environ 5 heures pour transcrire une émission d'une heure.
2. Le taux d'erreur varie selon le type de données choisi. Sur les jeux de test du NIST, le taux d'erreur de ce système varie de 10% à 15% selon le test.

3. Modélisation acoustique

Les données audio BN sont principalement de type studio, la part des données de qualité téléphonique y étant très réduite (environ 5%). Les conversations du corpus SWB sont de qualité téléphonique et ont été enregistrées sur 2 canaux correspondant chacun à un coté de la conversation. Comme pour les données BN, les vecteurs centisecondes comprennent 39 coefficients cepstraux obtenus à partir de spectres en échelle Mel. Ils sont estimés sur une bande réduite à 0-3,8kHz (à comparer à 0-8kHz pour les données BN). La moyenne et la variance de chaque coefficient cepstral sont normalisées pour chaque côté de la conversation. Pour les données BN, la normalisation est réalisée par groupe de segments supposés correspondre à un seul locuteur.

Les modèles phonétiques SWB ont la même topologie et sont construits de la même manière que les modèles BN. Ils sont estimés sur toutes les données transcrites disponibles des corpora SwitchBoard (3606 conversations) et CallHome[3] (120 conversations) collectés en téléphonie filaire. Environ 3% des données CallHome et 10% des données SwitchBoard ont été rejetées durant l'alignement forcé entre le signal et la transcription manuelle. Au total nous utilisons 430 heures de données, comprenant une quantité à peu près égale de locutrices et de locuteurs.

Ces données permettent de modéliser 32k contextes phonétiques avec environ 12k états liés. Les modèles acoustiques ont été estimés dans les mêmes conditions que les modèles BN (mêmes algorithmes et même jeu de questions). Les contextes phonétiques les plus fréquents sont modélisés, les contextes inter et intra-mots étant comptabilisés séparément. Deux ensembles de modèles acoustiques dépendants du sexe du locuteur sont obtenus après une adaptation MAP (Gauvain & Lee, 1994) des modèles indépendants du sexe.

Tous les résultats mentionnés ci-après ont été obtenus sur les données de test de l'évaluation Hub5 2001 du NIST (Eval01), composées de 3 sous-ensembles de 20 conversations chacun provenant des corpora SWB-I, SWB-II (téléphonie filaire) et SWB-II cellulaire pour un total d'environ 6 heures de signal.

3.1. *Normalisation spectrale*

La normalisation de la longueur de conduit vocal (VTLN) (Andreou *et al.*, 1994) est une technique de normalisation du locuteur intervenant au niveau des paramètres acoustiques. Elle est largement répandue dans les systèmes de reconnaissance à grands vocabulaires. Nous avions précédemment tenté d'appliquer sans succès cette technique aux données BN. Toutefois, au vu des gains significatifs observés sur les systèmes SWB développés par d'autres équipes (Hain *et al.*, 1999), nous avons reconsidéré notre position. Cette normalisation repose sur une modification linéaire de l'échelle des fréquences afin de compenser les différences de longueur de conduit

3. Dans la suite de l'article, nous utilisons le terme SWB pour désigner l'ensemble des ces deux corpora.

vocal entre les locuteurs. Le spectre de puissance en échelle Mel est estimé à partir d'un banc de filtres modifié selon une fonction linéaire par morceaux (pour ne pas sortir de la bande 0-3,8kHz). Le coefficient de normalisation est sélectionné parmi un ensemble de valeurs (0,8 à 1,25) pour maximiser la vraisemblance des données de test, en utilisant une transcription obtenue avec un décodage rapide.

Bien que la procédure d'estimation par recherche du maximum de vraisemblance permette de réduire significativement le taux d'erreur, les itérations de cette procédure ne convergent pas correctement pour les données d'apprentissage et peut donc être sous optimale. Ceci peut être attribué au fait que le Jacobien de la transformation est ignoré, même si la normalisation de la variance des paramètres devrait compenser en partie cela. Une prise en compte correcte du Jacobien suppose de construire un modèle acoustique par valeur possible du coefficient VTLN, ce qui doublerait le temps de calcul nécessaire à l'estimation de ce coefficient. Nous proposons une procédure originale permettant de s'affranchir de la compensation explicite du Jacobien.

Les coefficients VTLN sont estimés pour chaque côté de la conversation en alignant les segments audio avec leur transcription pour des valeurs allant de 0,8 à 1,25, mais nous utilisons des modèles mono-gaussiens dépendants du sexe (un jeu de modèles pour chaque sexe) pour estimer deux coefficients par locuteur. Avec des modèles cibles estimés sur les données d'un seul sexe, les histogrammes des coefficients VTLN (sur l'ensemble des locuteurs d'apprentissage et pour les deux jeux de coefficients) deviennent unimodaux et beaucoup plus compacts. La nature unimodale des distributions et le recours à des modèles mono-gaussiens réduisent grandement le besoin de compenser le Jacobien et rendent la procédure d'estimation très stable.

Bien que les modèles servant à l'estimation des coefficients VTLN (pour les données d'apprentissage et de test) soient appris séparément sur les données des locutrices et des locuteurs, les modèles utilisés lors de la reconnaissance sont estimés sur l'ensemble des données (cf. figure 1), puis ils sont adaptés avec les données d'un seul sexe. Les résultats expérimentaux sont regroupés dans le tableau 1 pour des modèles sans normalisation VTLN, des modèles avec une estimation des coefficients indépendante du sexe et dépendante du sexe, et avec et sans adaptation MLLR. On peut observer que, sans adaptation des modèles acoustiques, la normalisation VTLN réduit le taux d'erreur d'environ 2%. Ce gain est réduit à 1,5% après adaptation des modèles. Un gain supplémentaire de 0,4% est obtenu par la procédure d'estimation dépendante du sexe.

On peut aussi noter que les taux d'erreurs sont significativement différents pour les trois sous-ensembles de données (SWB1, SWB2, et SWB2-CELL). La différence entre les corpora SWB2 et SWB2-CELL est due au type de téléphonie (filaire versus cellulaire), tandis que la différence entre SWB1 et SWB2 tient au fait que les données de test SWB1 ont été collectées dans les mêmes conditions que les données d'apprentissage (SWB1) avec des locuteurs en commun.

VTLN	MLLR	SWB1	SWB2	CELL	total
non	non	28,0	36,1	42,2	35,6
GI	non	26,7	33,5	40,4	33,7
non	oui	26,1	32,0	38,1	32,2
GI	oui	24,4	30,2	36,9	30,7
GD	oui	24,2	30,1	36,2	30,3

Tableau 1. *Taux d'erreur pour les 3 sous-ensembles du corpus Eval01 avec des modèles indépendants du sexe (GI) et dépendants du sexe (GD), avec et sans normalisation VTLN.*

3.2. *Apprentissages adaptatif et discriminant*

L'apprentissage adaptatif SAT (Anastasakos *et al.*, 1996) vise à réduire l'impact des variations inter-locuteurs lors de l'estimation des modèles acoustiques. Pour ce faire, une transformation linéaire est estimée pour chaque locuteur du corpus d'apprentissage en maximisant la vraisemblance des données transformées étant donné le modèle multilocuteur. Un nouveau modèle est alors construit en utilisant les données d'apprentissage transformées. Ce modèle canonique est utilisé lors du décodage pour faciliter l'adaptation non-supervisée. La transformation linéaire est obtenue par le biais de la technique d'adaptation contrainte CMLLR (Gales, 1998), la même transformation étant appliquée aux vecteurs moyens et aux matrices de covariances des gaussiennes. On peut montrer que cette transformation contrainte des modèles est pour le décodage équivalente à une transformation des données (vecteurs de coefficients cepstraux). La réduction absolue du taux d'erreur par rapport à une adaptation MLLR sans modèle SAT est de l'ordre de 0,5%. L'utilisation d'une adaptation MLLR contrainte lors du décodage apporte un gain additionnel de 0,5%.

L'importance de l'apprentissage discriminant a été clairement montrée dans le cas de la parole conversationnelle (Woodland & Povey, 2000). Il s'agit d'une technique coûteuse dans la mesure où elle nécessite d'estimer la probabilité a priori des données. En pratique, cette probabilité est estimée à partir des treillis de mots obtenus sur l'ensemble des données d'apprentissage. Nous avons développé une procédure d'apprentissage de type MMIE (Woodland & Povey, 2000) compatible avec l'apprentissage de Viterbi utilisé dans notre système. Cette procédure commence par la génération (avec un modèle bigramme) d'un treillis de mots pour chaque tour de parole du corpus d'apprentissage. Ces treillis sont ensuite redécodés avec un modèle unigramme de façon à réduire l'influence du modèle de langage et à augmenter le nombre de confusions. Hormis la génération initiale des treillis, l'algorithme de réestimation n'est pas plus lent que l'algorithme non discriminant. Il permet de réduire de 1,9% absolu le taux d'erreur sans adaptation et de 1,2% après adaptation.

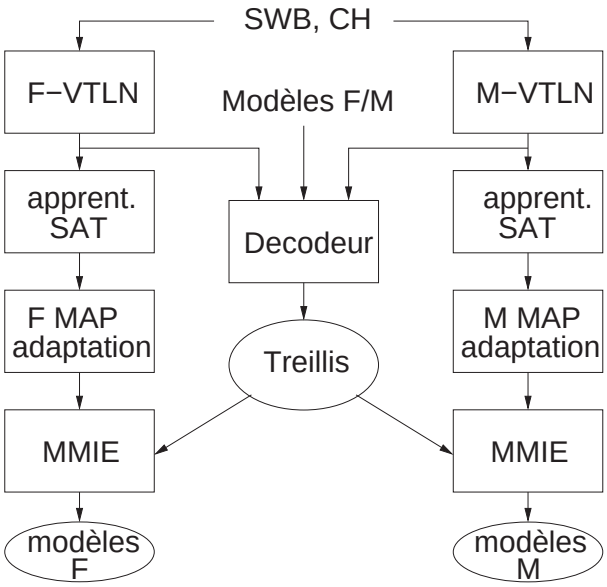

Figure 1. *Procédure d'apprentissage des modèles acoustiques*

4. Modélisation linguistique

Les difficultés posées par la modélisation linguistique de la parole conversationnelle résident d'une part dans le caractère spontané de la parole qui conduit à une syntaxe relâchée avec de nombreuses hésitations et reprises, et d'autre part dans la faible quantité de données d'apprentissage disponible. Pour la tâche BN, il est relativement aisé de trouver une grande variété de textes pertinents pouvant servir de données d'apprentissage. Pour la parole conversationnelle, la seule source disponible est la transcription des données audio d'apprentissage, c'est-à-dire environ 4,5 millions de mots pour 430 heures de données. Nous proposons deux approches permettant de compenser en partie ce problème. La première approche consiste à inclure dans les données d'apprentissage des textes d'autres sources, sélectionnés pour leur nature conversationnelle. La seconde approche consiste à utiliser un réseau de neurones pour lisser les probabilités *n*-grammes.

4.1. *Vocabulaire et prononciations*

Le vocabulaire de reconnaissance contient 51k mots comprenant les mots apparaissant au moins deux fois dans les données SWB (4,5M de mots), complétés avec les mots les plus fréquents dans un corpus de transcriptions BN (320M de mots). L'utilisation des transcriptions BN permet d'obtenir une meilleure couverture de la langue anglaise. Une différence importante avec le vocabulaire BN est l'incorporation des in-

terjections “uh-huh” et “mhm” (signifiant oui) et “uh-uh” (signifiant non) qui n’étaient pas considérées comme des éléments lexicaux. Comme pour le système BN, les 300 séquences lexicales les plus fréquentes, sujettes à des phénomènes de réduction importants, sont modélisées sous forme de mots composés. Par contre, les acronymes, nombreux dans le corpus BN, sont rares dans le corpus SWB, et ne sont pas traités comme des mots. La couverture lexicale atteint 99.7% sur le jeu de test NIST Eval01.

Les prononciations sont basées sur les mêmes 48 phones utilisés dans le système BN (dont 3 pour représenter les pauses, les hésitations et les respirations). Un graphe de prononciation est associé à chaque mot afin d’autoriser des variantes telles que les réductions. Les prononciations de base sont extraites du lexique anglais-américain du LIMSI. Les formes les plus fréquemment et fortement modifiées de par la nature conversationnelle des données ont été vérifiées et adaptées aux données SWB. Le dictionnaire de prononciations a un total de 60586 transcriptions phonétiques pour 51075 mots. Les probabilités des prononciations ont été évaluées à partir des fréquences d’occurrences relevées dans les alignements forcés des transcriptions sur les données d’apprentissage. La prise en compte des probabilités des prononciations n’avait jamais donné de gain significatif sur les performances du système BN. Pour les données SWB, le gain absolu sur le taux d’erreur est de 1,9% avant adaptation et de 0,4% après adaptation MLLR.

4.2. *Sélection de données conversationnelles*

Afin d’augmenter la quantité de données conversationnelles utilisée pour la construction du modèle, nous avons sélectionné dans le corpus de transcriptions BN les données les plus proches des données conversationnelles au moyen d’un critère de perplexité (Iyer & Ostendorf, 1999). Plus précisément, les phrases sélectionnées dans le corpus BN doivent être à la fois proche des données conversationnelles (en mesurant la perplexité pour un modèle estimé uniquement sur les données SWB) et éloignées des données journalistiques de la même période (en mesurant la perplexité pour un modèle estimé sur des textes de journaux). Ces deux critères de sélection sont combinés au moyen de la divergence de Kullback-Leibler pour les deux modèles (SWB et journaux) calculée sur les données BN. Pour chaque année représentée dans le corpus BN, on retient un tiers des phrases ayant la divergence la plus grande entre le modèle conversationnel et le modèle journalistique de la même année. Au total on obtient un corpus pseudo-conversationnel de 65M de mots. Un second corpus pseudo-conversationnel de 180M de mots sélectionné sur le WEB par l’Université de Washington a également été inclus dans les données d’apprentissage (Bulyko *et al.*, 2003).

Un modèle *n*-grammes a été construit pour chaque source de données (SwitchBoard, CallHome, BN, données sélectionnées) au moyen du logiciel SRI LM (Stolcke, 2002). Nous avons obtenu les meilleurs résultats en perplexité avec le lissage de Kneser-Ney modifié (Chen & Goodman, 1999). Ces modèles ont ensuite été combinés par interpolation en utilisant l’algorithme EM pour estimer les coefficients d’in-

Modèle	SWB+BN	+ BNselect & WEB	+ RN
perplexité	58,5	56,9	54,5
taux d'erreur	21,7%	21,5%	21,1%

Tableau 2. *Comparaison des différents modèles de langage sur le test Eval01. SWB+BN : corpus SWB et BN, BNselect : sélection de données conversationnelles dans le corpus BN, WEB : corpus conversationnel extrait du WEB, RN : modèle de langage neuronal entraîné sur les transcriptions (4,5M mots).*

terpolation qui minimisent la perplexité du modèle résultant sur notre ensemble de développement.

Les résultats en terme de perplexité et de taux d'erreur obtenus avec ces modèles sont donnés dans les deux premières colonnes du tableau 2. Les taux d'erreur sont obtenus avec les meilleurs modèles acoustiques en utilisant la procédure de décodage décrite plus loin dans la section 5. Le modèle linguistique utilisant seulement les transcriptions des données conversationnelles (SWB) et des émissions d'information (BN) a une perplexité de 58,5 et un taux d'erreur de 21,7%. L'ajout des données sélectionnées (BNselect) et des données WEB permet des réduire la perplexité à 56,9 et le taux d'erreur à 21,5%. Il est à noter que l'apport des données BNselect devient minime, voire négligeable lorsqu'on dispose de suffisamment de données WEB. La réduction du taux d'erreur due à ces données est de 0.15% sans les données WEB et n'est plus que de 0.03% avec les données WEB.

4.3. *Lissage des probabilités n-grammes*

Dans un modèle n-gramme, les mots sont habituellement représentés dans un espace discret (indices dans le vocabulaire) rendant toute interpolation difficile. L'idée à la base du modèle neuronal consiste à projeter les mots dans un espace continu et à estimer les probabilités *n*-grammes dans cet espace (Bengio & Ducharme, 2001). Ainsi les distributions de probabilités sont des fonctions continues des représentations des mots, permettant une meilleure généralisation à des *n*-grammes inobservés.

La figure 2 résume l'approche pour le cas d'un modèle quadrigramme, c'est-à-dire pour un contexte de trois mots. La projection des mots dans un espace continu de dimension P=50 est effectuée par la première couche du réseau. Pour cela un mot est codé à l'entrée du réseau par un vecteur binaire de dimension N (taille du vocabulaire). Lorsque le i-ème mot du vocabulaire est présenté au réseau, la i-ème composante de ce vecteur vaut 1 et toutes les autres 0. La multiplication de ce vecteur avec la matrice de projection (poids de la première couche) donne la représentation continue du mot. La matrice de projection est la même pour tous les mots du contexte. Les deux autres couches du réseau servent à estimer les probabilités n-grammes pour un contexte h_j donné (entrée du réseau) et pour l'ensemble du vocabulaire :

$$P(w_j = i|h_j), \forall i = 1...N$$

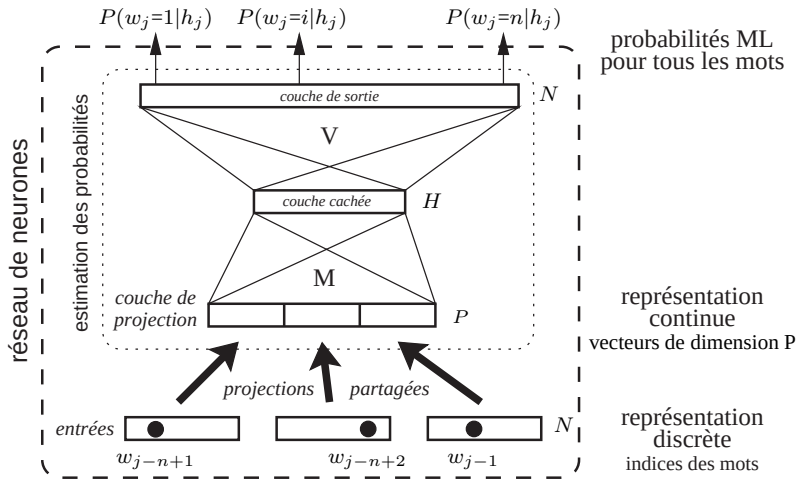

Figure 2. *Architecture du modèle de langage neuronal quadrigramme.* h_j *dénote le contexte* $w_{j-n+1}, ..., w_{j-2}, w_{j-1}$ *et* N *et la taille du vocabulaire.*

. Le réseau de neurones apprend conjointement, par rétropropagation du gradient, la projection des mots dans l'espace continu et les probabilités *n*-grammes en minimisant la perplexité des données d'apprentissage (Schwenk & Gauvain, 2002; Schwenk & Gauvain, 2004).

La complexité de cette architecture étant élevée, plusieurs améliorations et optimisations ont été apportées pour la rendre utilisable dans un système de reconnaissance de la parole (Schwenk, 2004). Notamment, la dimension de la couche de sortie est limitée aux 2000 mots les plus fréquents, les autres étant traités par un modèle de langage à repli. Le décodage est effectué en générant un treillis de mots avec un modèle à repli, ce treillis est ensuite réévalué par le modèle neuronal. Ceci permet d'obtenir les mêmes performances qu'un décodage complet mais avec un gain en vitesse considérable. La réévaluation des treillis avec le modèle neuronal demande moins de 0,1xRT sur un processeur Intel Xeon à 2.8GHz.

Le modèle neuronal est estimé sur les transcriptions des données conversationnelles (4,5M de mots), et il est interpolé avec le modèle n-gramme classique lors du décodage. Le tableau 2 contient les perplexités et les taux d'erreur obtenus avec ce modèle et avec le modèle n-gramme de référence. L'utilisation du modèle neuronal réduit le taux d'erreur de 0,4%. Tous ces résultats correspondent à une réévaluation des treillis de la dernière passe de décodage (cf. section 5).

Passe	MA	VTLN	MLLR	ML	CN	RTF	Err.
1	MMIE	non	non	3g	non	1,9	36,1
2	SAT, MMIE	oui	2	3g	non	13,9	23,6
		oui	2	4g	non	0,0	22,9
		oui	2	4g	oui	0,0	22,1
3	SAT, MMIE	oui	5	4g + RN	oui	3,1	21,1

Tableau 3. *Taux d'erreur en mots sur les données Eval01 pour chaque passe de décodage. MLLR : le nombre de classe de régression est spécifié pour chaque passe. CN : décodage par réseau de consensus avec probabilités de prononciation. RTF : facteur temps réel sur un processeur Intel Xeon à 2.8GHz.*

5. Décodage

La procédure de décodage a été largement modifiée. Les principales modifications portent sur l'estimation du coefficient VTLN, la procédure d'adaptation au locuteur et le décodage par consensus utilisant des probabilités de prononciation. Le décodage est réalisé en 3 étapes. Dans la première passe, le sexe du locuteur est déterminé pour chaque côté de la conversation (au moyen de 2 GMM) et un décodage trigramme rapide permet de générer une transcription initiale (cf. tableau 3, ligne 1). Cette transcription sert d'une part à estimer les coefficients VTLN pour chaque côté des conversations et d'autre part à adapter les modèles SAT qui sont utilisés dans la seconde passe. Les passes 2 et 3 utilisent les données normalisées. Chacune génère un treillis de mots avec un modèle trigramme. Ce treillis est ensuite réévalué avec un modèle quadrigramme et est transformé en réseau de consensus en intégrant les probabilités de prononciation. Les probabilités *a posteriori* des arcs du treillis sont estimées par l'algorithme avant-arrière. Les réseaux de consensus sont obtenus en fusionnant de façon itérative les noeuds des treillis et en dupliquant les arcs jusqu'à ce qu'un graphe linéaire soit obtenu pour chaque treillis. Cette procédure permet d'obtenir des résultats comparables à ceux obtenus avec l'algorithme de regroupement d'arcs proposé dans (Mangu *et al.*, 1999) mais elle se révèle significativement plus rapide. La transcription finale est obtenue en prenant le mot le plus probable pour chaque ensemble de confusions.

Les transcriptions des passes 1 et 2 sont utilisées lors de la passe suivante pour les adaptations MLLR (contrainte et non contrainte). Une seule classe de régression est utilisée pour l'adaptation contrainte, alors que pour l'adaptation non contrainte nous utilisons deux classes de régression (parole/non parole) en deuxième passe, et cinq classes (non parole, consonnes non voisées, consonnes voisées, et deux classes de voyelles) pour la dernière passe. Afin d'accélérer le décodage, l'espace de recherche pour la dernière passe est restreint au treillis de la seconde passe après transformation en graphe de mots.

Le taux d'erreur obtenu après chaque passe est donné dans le tableau 3, pour le jeu de test NIST Eval01. La réduction importante du taux d'erreur entre la première

et la seconde passe est due à la combinaison de la transformation VTLN, des adaptations acoustiques, du modèle quadrigramme, des probabilités de prononciation et du décodage par réseau de consensus (la contribution de chaque amélioration est donnée dans le tableau 3). Le gain de la troisième passe vient de l'adaptation acoustique avec 5 classes de régression (-0,6%) et de l'utilisation du modèle à réseau de neurones (-0,4%).

Le temps de calcul pour chaque étape est indiqué dans la colonne RTF du tableau 3. L'étape la plus longue est la génération des treillis de mots qui nécessite en moyenne un temps égal à environ 14 fois la durée des données à décoder. La réévaluation des treillis avec un modèle quadrigramme et le décodage par consensus représentent un temps négligeable. Au total le temps de décodage est égal à 18,9 fois la durée du signal.

6. Conclusions

Nous avons décrit le travail mené au LIMSI pour développer un système de transcription de conversations téléphoniques à partir d'un système de transcription d'émissions d'information. Il apparaît que le traitement de la parole conversationnelle requiert des modifications importantes tant pour les modèles acoustiques et linguistiques que pour le processus de décodage. Le taux d'erreur initial du système BN sur les données conversationnelles était de l'ordre de 50%. La simple réestimation des paramètres des modèles sur les données SWB ne conduit pas à un taux d'erreur satisfaisant. Il a donc fallu revoir l'ensemble des composants du système. Les éléments suivants on été ajoutés : une normalisation spectrale (VTLN), un apprentissage adaptatif SAT et un apprentissage discriminant des modèles acoustiques, une adaptation MLLR contrainte, l'utilisation de classes phonémiques pour l'adaptation MLLR, un modèle de langage neuronal, et un décodage par réseau de consensus utilisant des probabilités de prononciation. Tous ces raffinements ont permis d'obtenir un taux d'erreur de 21% avec un facteur temps-réel inférieur à 20.

7. Bibliographie

ANASTASAKOS T., MCDONOUGH J., SCHWARTZ R. & MAKHOUL J. (1996). A compact model for speaker-adaptive training. In *Proceedings ICSLP*, volume 2, p. 1137–1140, Philadelphia, PA.

ANDREOU A., KAMM T. & COHEN J. (1994). Experiments in vocal tract normalisation. In *CAIP Workshop Frontiers in Speech Recognition II*.

BENGIO Y. & DUCHARME R. (2001). A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13: Morgan Kaufmann.

BULYKO I., OSTENDORF M. & STOLCKE A. (2003). Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In *Human Language Technology Conference*.

CHEN S. F. & GOODMAN J. T. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, **13**(4), 359–394.

GALES M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer, Speech and Language*, **12**(2), 75–98.

GAUVAIN J.-L. & LAMEL L. (2003). Structuring broadcast audio for information access. *EURASIP journal on Applied Signal Processing*, **2003**(2), 140–150.

GAUVAIN J.-L., LAMEL L. & ADDA G. (2002). The limsi broadcast news transcription system. *Speech Communication*, **37**(1-2), 89–108.

GAUVAIN J.-L. & LEE C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Trans. on Speech and Audio Processing*, **2**(2), 291–298.

GODFREY J., HOLLIMAN E. & MCDANIEL J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, p. 517–520, San Francisco.

HAIN T., WOODLAND P., NIESLER T. & WHITTAKER E. (1999). The 1998 htk system for transcription of conversational telephone speech. In *Proceedings IEEE ICASSP*.

IYER R. & OSTENDORF M. (1999). Relevance weighting for combining multi-domain data for n-gram language modeling. *Computer Speech and Language*, **13**, 267–282.

LEGGETTER C. & WOODLAND P. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, **9**, 171–185.

LJOLJE A. & *et al* (2000). The at&t lvcsr-2000 system. In *Proc. NIST Speech Transcription Workshop*.

MANGU L., BRILL E. & STOLCKE A. (1999). Finding consensus among words: Lattice-based word error minimization. In *Proceedings ESCA Eurospeech*, p. 495–498.

MATSOUKAS S., COLTHURST T., KIMBALL O., SOLOMONOFF A., RICHARDSON F., QUILLEN C., GISH H. & DONGIN P. (2002). The 2001 byblos english large vocabulary conversational speech recognition system. In *Proceedings IEEE ICASSP*, volume I, p. 721–724.

SCHWENK H. (2004). Efficient training of large neural networks for language modeling. In *IEEE joint conference on neural networks*, p. 3059–3062.

SCHWENK H. & GAUVAIN J.-L. (2002). Connectionist language modeling for large vocabulary continuous speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, p. I: 765–768.

SCHWENK H. & GAUVAIN J.-L. (2004). Neural network language models for conversational speech recognition. In *International Conference on Speech and Language Processing*, p. 1283–1286.

STOLCKE A. (2002). Srilm - an extensible language modeling toolkit. In *Proceedings IEEE ICASSP*, volume 2, p. 901–904.

STOLCKE A. & *et al* (2000). The sri march 2000 hub-5 conversational speech transcription system. In *Proceedings NIST Speech Transcription Workshop*.

WOODLAND O. & POVEY D. (2000). Large scale discriminative training for speech recognition. In *Proceedings ISCA ITRW ASR'00*, p. 7–16.