Le système de dictée du LIMSI pour l'évaluation AUPELF'97

G. Adda, M. Adda-Decker, J.L. Gauvain, L.F. Lamel,

Groupe Traitement du langage parlé
LIMSI-CNRS, BP 133
91403 Orsay cedex, FRANCE
{gadda,madda,gauvain,lamel}@limsi.fr

Résumé

Dans le cadre du projet de *Dictée Vocale* de l'ARC-B1 de l'AUPELF nous avons apporté différentes améliorations à notre système de dictée en langue française. Les travaux réalisés visaient en premier lieu à augmenter la couverture lexicale pour le français afin de minimiser les erreurs du système dues aux mots inconnus. Ensuite la précision des modèles de langage a été accrue par une augmentation significative des corpus d'apprentissage (*Le Monde 87-96*, *Le Monde Diplomatique 89-96*, *Agence France Presse 94-96* totalisant environ 270 M mots) et par les choix de leur utilisation. La modélisation acoustique a pu être affinée par les techniques de partage d'états entre modèles, permettant de générer un nombre plus élevé de modèles contextuels. Pour le décodeur différents ensembles de modèles de langage et modèles acoustiques plus ou moins complexes sont développés. Le décodeur intègre ces sources de connaissance de manière progressive lors de trois étapes successives. Le système de dictée est évaluée ici sur les données de développement (609 phrases) de la campagne AUPELF'97 correspondant à des textes lus du journal *Le Monde* datant de mai 1996. Des résultats sont présentés avec notre système utilisant les listes officielles de 20k et 64k mots et une liste libre de 65k mots. L'utilisation de cette dernière fournit les meilleurs résultats de reconnaissance, avec un taux d'erreur de 12,7% sur l'ensemble des données (8,7% sur l'ensemble réduit T').

Introduction

Dans le cadre du projet de *Dictée Vocale* de l'ARC-B1 de l'AUPELF nous avons apporté différentes améliorations à notre système de dictée en langue française [Gauvain J.-L. et al (1994)], [Lamel L.F. et al (1996)]. Le système de reconnaissance repose sur des principes de modélisation markovienne tant au niveau acoustique, qu'au niveau linguistique. Les connaissances spécifiques à la langue française sont apportées par les bases de données (parole et texte), le choix d'un ensemble de phonèmes, et un lexique donnant pour chaque mot graphémique une ou plusieurs transcriptions phonétiques utilisant les symboles phonétiques fixés auparavant. Le système de reconnaissance utilise des modèles acoustiques de phones en contexte, chaque phone en contexte étant représenté par une chaîne de Markov cachée à 3 états avec des densités d'observations multi-gaussiennes. Les modèles de langage utilisés sont des statistiques N-grammes (bigrammes et trigrammes) estimées sur des textes journalistiques.

Nous décrivons dans cet article nos travaux de développement pour l'évaluation AUPELF'97. Ces travaux concernent le choix du lexique et des transcriptions phonétiques, le choix des corpus d'apprentissage pour les modèles de langage et les modèles acoustiques et de l'estimation de ces modèles. Dans la présente campagne d'évaluation nous participons aux deux catégories de test P0 et Q0. Dans la catégorie P0 on limite la sortie du système de reconnaissance à des suites de mots issus de listes de mots prédéfinies et communes à tous les participants de la campagne (20k mots, 64k mots). La catégorie Q0 laisse au système la liberté de définir la liste la mieux adaptée au test prévu. Nous donnons les résultats obtenus dans ces différentes catégories.

Modélisation lexicale

Le système de reconnaissance fait appel à un ensemble de mots (liste de mots), chaque mot étant représenté acoustiquement par un graphe de modèles acoustiques de phones en contexte, graphe qui est défini par la (les) transcription(s) phonétique(s) associée(s) à ce mot dans le lexique de transcription.

Avec une telle approche la taille et le contenu de la liste de mots influent directement sur les performances de reconnaissance par le biais du pourcentage de MHV (mots hors vocabulaire) présents dans les phrases de test. Ainsi il est important d'avoir une liste de mots de taille importante pour garder un taux de MHV faible. Pour une taille fixée le taux de MHV peut être minimisé par un choix approprié des mots à inclure dans la liste. La définition de la liste de mots se fait suite à une première étape de nettoyage et de normalisation des textes, pendant laquelle la notion de mot est définie pour le système.

Normalisation graphémique

Les textes d'apprentissage contiennent une part de bruit dans le codage graphémique: formatage, codage des caractères accentués, symboles de ponctuation (apostrophe, trait d'union, point), majuscules en début de phrase, et fautes de frappe. La normalisation graphémique vise à réduire ce bruit permettant ainsi de diminuer le nombre de formes différentes observées, ce qui conduit à une meilleure couverture lexicale, et un meilleur apprentissage des modèles de langage. Pour définir la notion du mot à l'intérieur du système, les règles et traitements suivants ont été appliqués

[Adda G. et al., (1997)]:

1. traitement des majuscules en début de phrase (Dorénavant → dorénavant),
2. traduction des chiffres (y compris des chiffres romains) en mots (110 → cent dix),
3. éclatement des sigles non acronymes (ADB → A. D. B.),
4. détection des mots composés (arc-en-ciel, aujourd'hui) avant éclatement d'unités liées par des marqueurs de ponctuation (prends-le, j'aurai).

Définition de la liste de mots

Le choix des mots se fait de manière automatique en utilisant un seuil sur les comptes d'occurrences de mots dans les textes d'apprentissage. L'adéquation entre textes d'apprentissage et de test peut être optimisée en pondérant l'utilisation entre textes d'apprentissage récents et sources plus anciennes. L'adéquation progressive entre textes d'apprentissage et mots du test se traduit par un taux de MHV décroissant. L'utilisation des textes récents est particulièrement importante pour couvrir les noms (essentiellement noms propres) se rapportant à l'actualité immédiate (par exemple concernant l'actualité politique en Birmanie: birmanisation, Aung San Suu Kyi). Retenir des textes sur une période très longue permet de sélectionner les mots d'un intérêt plus général, ayant une durée de vie plus longue.

La liste de 65400 mots pour la catégorie de test Q0 a été élaborée à partir des textes du *Monde* des années de janvier 1987 à avril 1996 (185 M de mots) et des textes du *Monde Diplomatique* de janvier 1989 à avril 1996 (6 M de mots). Un sous-ensemble de ces textes a été choisi en optimisant le taux de couverture de la liste de mots sur les données de développement (textes du *Monde* de mai '96). Les textes des mois de mai à septembre 1996 n'ont pas été utilisés pour faire ce choix. La meilleure couverture a été obtenue en retenant les mots les plus fréquents des textes du *Monde* des années 1991 à 1996 (sans utiliser le *Monde* diplomatique) et en appliquant un coefficient 2 aux textes de 1996. Le taux de mots hors vocabulaire en utilisant cette liste est de 1,34% sur les données de développement.

Lexique de transcriptions phonémiques

Le lexique de transcriptions phonémiques a été développé au LIMSI, à partir d'un logiciel de conversion automatique graphème-phonème [Prouts B., (1980)], [Boula de Mareüil P., (1996)]. Des corrections manuelles sont effectuées et des variantes de prononciations ajoutées (exemples dans table 1). Les variantes servent essentiellement à modéliser des phénomènes comme: le e-muet optionnel en fin de mot ou morphème finissant par une ou plusieurs consonnes (*Budapest*, *vingt-deux*), les réductions de groupes consonantiques finals (*autres*, *décembre*) et les liaisons inter-mot. Toutes ces variantes permettent de générer un nombre variable de phonèmes par mot. D'autres types de variantes autorisent différents phonèmes à une position donnée (*était*, *désertions*). Enfin les mots posant le plus de problèmes pour la transcription phonétique sont les mots d'origine étrangère (essentiellement noms propres), comme par exemple: *squatter*, *Morgan*, *Wonder*. Ces mots peuvent être plus ou moins « francisés » suivant leur historique et/ou leur contexte sémantique (*Morgan*: `Michèle Morgan` ou `la banque Morgan Stanley`), ou suivant la culture linguistique du locuteur.

```
était        [eE]tE [eE]tEt(V)
vingt-deux   vIt{x}d@ vI{n}d@
              vIt{x}d@z(V) vI{n}d@z(V)
autres       otrx otr(V.) otrxz(V) ot
décembre     desAbrx desAbr(V.) desAb(C)
squatter     skwate skwater(V) skwat[XE]r
désertions   dezEr[st]jO
Budapest     bydapEst bydapEstx(C.)
Morgan       mcrgA mcrgan
Wonder       wcndXr wcndEr vOdEr
```

Figure 1: Exemples de transcriptions phonétiques multiples pour la reconnaissance. Les symboles phonétiques entre crochets sont au choix, entre accolades ils sont optionnels. Les parenthèses codent des contraintes d'enchaînement avec **V** désignant les voyelles, **C** les consonnes et **.** le silence.

Modèles de langage

Les modèles de langage N-grammes permettent de tenir compte localement (sur des suites de N mots) de la syntaxe de la langue, ou plus précisément d'un langage décrivant les corpus d'apprentissage. Deux types de modèles de langage sont construits: les modèles bigrammes et les modèles trigrammes. Les modèles de bigrammes sont utilisés pour la génération des graphes de mots, étape servant essentiellement à définir un sous-langage plausible, étant donné le signal acoustique. Pour des raisons matérielles la taille de ces modèles doit rester limitée ($\sim$ 2 M bigrammes). Les modèles de trigrammes sont utilisés pour l'étape de reconnaissance proprement dite. Ici les modèles peuvent être significativement plus importants et doivent être les plus précis possible.

Différents modèles sont construits pour chacun des deux types, utilisant des configurations différentes pour les corpus d'apprentissage. Les choix de ces configurations sont liés à la disponibilité des textes normalisés au moment de la création des modèles.

Les perplexités montrées dans les tableaux 1 et 2 sont obtenues uniquement sur les phrases de développement. Les textes utilisés pour l'apprentissage vont jusqu'à la date du texte de développement, c'est-à-dire avril 96 inclus.

Modèle bigramme

Dans le tableau 1 nous comparons des modèles de bigrammes comprenant environ 2 M de bigrammes, construits à partir de 3 ensembles variant dans des proportions inférieures à 40% :

1. *Le Monde 91-96*, *Le Monde Diplomatique 89-96*,

2. *Le Monde 87-96*,

3. *Le Monde 87-96*,*Le Monde Diplomatique 89-96*.

On peut voir que le corpus d'apprentissage augmente d'environ 35% en passant de 1) à 2). Mais les données supplémentaires correspondent à des sources relativement anciennes (années 87-90), et n'apportent pas d'amélioration significative de la perplexité sur le test de développement. Entre 2) à 3) on n'a rajouté que 3% de données supplémentaires, mais incluant des données plus récentes (*Monde Diplomatique 89-96*). On peut noter ici un gain de perplexité plus important. Ce dernier modèle de langage a été retenu pour la première passe du décodage sur les données de développement.

Modèle trigramme

Au moment de la construction des modèles de trigrammes l'ensemble des années de 87-96 du journal *Le Monde* a été normalisé. Les différents ensembles de textes d'apprentissage correspondent ici aux 3 ensembles :

1. *Le Monde 87-96*,

2. *Le Monde 87-96*, *Le Monde Diplomatique 89-96*

3. *Le Monde 87-96*, *Le Monde Diplomatique 89-96*, *Agence France Presse 94-96*

variant dans des proportions inférieures à 30%. Dans le tableau 2 sont spécifiés le nombre de mots dans chacun de ces ensembles, la taille du modèle de langage estimé (nombres de trigrammes et de bigrammes) et la perplexité obtenue sur le développement. On peut voir que l'adjonction du corpus *AFP* (65 M mots) entraîne une augmentation notable de la taille du modèle de langage, indiquant que la taille du corpus d'apprentissage est probablement toujours insuffisante. La proportion de bigrammes augmente de 10%, les trigrammes d'environ 25%. La perplexité baisse légèrement sur le corpus de développement. C'est donc ce dernier modèle de langage qui est retenu pour la deuxième passe du décodage sur les données de développement.

Modèle biclasse

Une des conséquences de la taille insuffisante du corpus d'apprentissage, est l'absence dans le modèle de langage d'un certain nombre de bigrammes et trigrammes de mots réalisant des accords en genre et en nombre. La seule information utilisée alors est la probabilité de l'unigramme. Il en découle un nombre important d'erreurs dues à des problèmes d'accords à court terme,

Pour pallier ce défaut, nous avons interpolé le trigramme utilisé en dernière passe avec un bigramme de classes de mots. Les classes sont obtenues automatiquement par recuit simulé et minimisation de la perplexité, un mot n'admettant qu'une seule classe [Jardino M. (1996)].

	LeM 91-96 + LeMD	*LeM 87-96*	*LeM 87-96 + LeMD*
nb. mots	120 M	185 M	191 M
seuil d'occ.	3	4	4
nb. bigram.	1,83 M	2,09 M	2,17 M
perplexité	232,3	232,0	228,8

Table 1: Modèles bigrammes estimés sur les 3 corpus d'apprentissage suivants:
1) *Le Monde 91-96 + Le Monde Diplomatique 89-96* (*LeM 91-96 + LeMD*),
2) *Le Monde 87-96,* (*LeM 87-96*)
3) *Le Monde 87-96 + Le Monde Diplomatique 89-96* (*LeM 87-96 + LeMD*).
Ce tableau contient pour chaque corpus la perplexité d'un modèle bigramme, le nombre de mots dans le corpus, le seuil du nombre minimum d'occurrence d'un bigramme et le nombre de bigrammes retenus dans le modèle de langage.

	LeM	*LeM + LeMD*	*LeM+ LeMD + AFP*
nb. mots	185 M	191 M	255 M
seuil bg.	0	0	0
seuil tg.	1	1	1
nb. bigram.	11,9 M	12,1 M	13,5 M
nb. trigram.	13,9 M	14,3 M	18,1 M
perplexité	137,7	137,3	135,2

Table 2: Modèles trigrammes estimés sur les 3 corpus d'apprentissage suivants :
1) *Le Monde 87-96* (*LeM:*),
2) *Le Monde 87-96 + Le Monde Diplomatique 89-96* (*LeM + LeMD*),
3) *Le Monde 87-96 + Le Monde Diplomatique 89-96 + Agence France Presse 94-96* (*LeM + LeMD + AFP*).
Ce tableau contient pour chaque corpus la perplexité du modèle trigramme, le nombre de mots dans les corpus d'apprentissage, le seuil du nombre minimum d'occurrences de bigrammes et trigrammes, le nombre de bigrammes et de trigrammes retenus dans le modèle de langage.

Modèles acoustiques

Les paramètres des modèles ont été estimés en utilisant l'algorithme MAP [Gauvain J.-L., Lee C.H., (1994)]. Pour évaluer l'importance de la taille du corpus d'apprentissage pour la qualité des modèles acoustiques, nous avons construit trois familles de modèles, suivant trois ensembles d'apprentissage croissant, mais de même nature:

- **Bref80**: 5,3k phrases de 80 locuteurs
- **Bref**: 66,6k phrases de 120 locuteurs (incluant Bref80),
- **Bref+Bref2**: 85,9k phrases de 420 locuteurs (incluant Bref), les 300 locuteurs supplémentaires, enregistrés au LIMSI ces dernières années, ont lu chacun environ 60 phrases du *Monde*.

La précision de la modélisation acoustique est fonction du nombre de contextes phonétiques différents que l'on peut correctement modéliser. La technique de partage d'états permet d'augmenter le nombre de contextes modélisables.

L'adaptation des modèles au locuteur permet de gagner en précision à nombre de contextes fixes.

Modèles de triphones

Le corpus **Bref80** permet de construire environ 600 modèles de triphones en imposant un nombre minimum d'occurrences de 250 pour un contexte donné. L'utilisation de **Bref** permet d'augmenter le nombre de triphones à 4000 contextes. En séparant les données de **Bref** pour les locuteurs masculins et féminins, on arrive avec le même seuil à 2500/2800 pour modèles hommes/femmes. Le corpus **Bref+Bref2** permet d'étendre l'ensemble des triphones à 3100/3200 modèles homme/femme. Les gains observés sur les données de développement, en passant de Bref80 à Bref, se situent autour de 12% en relatif, en passant de Bref à Bref+Bref2 le gain relatif est inférieur à 3%.

Modèles de triphones à partage d'états

Le partage d'états [S. Young, P. Woodland, (1994)] est une technique efficace pour augmenter le nombre de contextes modélisables à base de données d'apprentissage constante. Au lieu de considérer un état lié à un seul modèle acoustique, il peut contribuer à modéliser différents contextes. Il est intéressant d'observer que le nombre d'états reste en gros constant (comparé aux modèles de triphones simples) dans la configuration optimale, mais le nombre de contextes peut être plus que doublé. Les gains relatifs en rajoutant le partage d'états en configuration optimale des modèles reste néanmoins inférieur à 5% (relatif).

Modèles de triphones adaptés au locuteur

L'adaptation au locuteur se fait de manière non-supervisée à partir de la sortie du système de reconnaissance (à l'issue de la première étape du décodage à précision maximale) en utilisant la méthode MLLR [Legetter C.J., Woodland P.O., (1995)]. Les gains observés par l'adaptation tournent autour de 5% en relatif, et paraissent d'autant plus faibles que le nombre de contextes modélisés est élevé (et le modèle de langage est puissant).

Décodeur

Le décodeur intègre les différents modèles de manière progressive. Lors d'une première étape, le but du décodeur consiste simplement à restreindre l'espace de recherche à explorer lors de la phase de reconnaissance proprement dite. L'espace de recherche, composé a priori de l'ensemble des phrases possibles du français, peut être réduit de manière significative étant donnés le signal acoustique et des modèles simples. A l'issue de cette étape on dispose pour chaque phrase d'un graphe de mots hypothèses. Lors d'une deuxième étape, qui est la phase de reconnaissance proprement dite, les modèles les plus précis sont utilisés sur l'espace de recherche restreint. La restriction de l'espace de recherche, via la génération de graphes de mots, entraîne le risque de décisions erronnées qui doivent être minimisées en gardant l'espace de recherche suffisamment large.

Génération de graphes de mots

Le but ici est de générer des graphes de mots réalisant un compromis entre taille des graphes et taux d'erreur dans les graphes. Par taux d'erreur du graphe, nous entendons le pourcentage de mots de la transcription de référence manquant dans le chemin optimal à travers le graphe (optimal par rapport à la transcription graphémique de référence). Pour la génération des graphes on utilise un modèle de langage bigramme (2,2 M de bigrammes) et comme modèle acoustique des triphones dépendants de leur position dans le mots, comprenant environ 7900 états (modélisant environ 2600 triphones).

Décodage à précision maximale

Lors de la phase de reconnaissance proprement dite il s'agit d'intégrer les modèles les plus précis et les mieux adaptés aux phrases à reconnaître. Le décodage à précision maximale se fait en deux étapes. D'abord les phrases sont décodées avec des triphones indépendants de leur position (environ 9000 états partagés entre environ 5200 triphones) et un modèle de langage trigramme (13.5M de bigrammes et 18.1M de trigrammes) en limitant l'espace de recherche aux graphes de mots générés lors de la première passe. Ensuite les hypothèses produites par le système sont utilisées pour adapter les modèles acoustiques au locuteur. Une dernière étape de reconnaissance utilise les modèles acoustiques adaptés pour chaque locuteur. Le modèle biclasse est optionnellement utilisé lors de cette dernière passe.

Expériences et résultats

Nous décrivons ici essentiellement les expériences concernant la catégorie Q0, où chaque participant a la liberté de construire le système le plus performant possible, sans contraintes sur les corpus d'apprentissage et sur la liste des mots que le système peut reconnaitre. Les résultats fournis en catégorie P0 (listes de mots 20k et 64k imposées) sont à interpréter de la manière suivante: comme le système permet de gérer jusqu'à 65k mots, l'ensemble des mots inconnus est modélisé par l'ensemble des mots les plus fréquents absents de la liste officielle jusqu'à la limite de 65k mots. L'approche que nous avons retenue, permet de garder le même modèle de langage pour les tests P0 et Q0. Les mots identifiés mais non présents dans les listes officielles sont retirés de l'hypothèse pour la catégorie P0. Avec le lexique de 65k mots on obtient une couverture de 98,6% des mots du test de développement. Ceci correspond à environ 200 mots inconnus pour un texte d'environ 15000 mots (taille des données de développement).

Nous donnons une synthèse des résultats sur le corpus de développement dans le tableau 3. De manière générale on peut noter un gain important en passant de Bref80 à Bref (réduction d'erreur de 15% relatif), et des différences non significatives entre Bref et Bref+Bref2. En fait on observe avec l'utilisation de Bref2 une réduction significative du taux d'erreur (de l'ordre de 4%) lorsque les textes

	MHV (%)	*Bref80*	*Bref*	*BREF+B2*
20k-P0 T'	3,67	12,46	10,79	10,72
64k-P0 T	1,89	15,39	13,32	13,39
65k-Q0 T'	0,45	10,79	8,83	8,75
65k-Q0 T	1,34	15,03	12,86	12,92

Table 3: Résultats obtenus en fonction des différents ensembles d'apprentissage utilisés pour les modèles acoustiques. Les résultats correspondent aux taux d'erreur de mots obtenus avec les systèmes 20k, 64k, 65k+ sur les ensembles de test de développement T et T' (600 et 300 phrases).

de l'AFP ne sont pas utilisés, c'est-à-dire lorsque le modèle de langage est moins précis. Cette observation a motivé notre décision d'utiliser les modèles Bref+Bref2 pour notre système primaire lors de l'évaluation.

Comparant les résultats de Q0 et de P0-64k sur l'ensemble de développement complet T, on constate que la différence entre les taux d'erreur respectifs est légèrement inférieure à la différence entre les deux taux de MHV. La dégradation observée en passant de Q0 à P0-64k est d'environ 3% (relatif).

Comparant les résultats de Q0-T et Q0-T'(T': 300 phrases, sous-ensemble de T, à taux de MHV contrôlé) on peut observer que le gain est largement supérieur à la réduction du taux de MHV. Ceci peut s'expliquer par une homogénisation des thèmes traités lorsqu'on contrôle le taux de MHV et par une réduction importante de la perplexité. Quand on élimine les paragraphes contenant beaucoup de mots nouveaux, on a tendance à éliminer des sujets nouveaux. Le contrôle du taux de MHV semble donc avoir comme conséquence un choix de textes, non seulement mieux couvert par la liste de mots, mais également par le modèle de langage.

Comparant les résultats de Q0-T' et P0-20k-T' on peut constater que l'écart entre les taux d'erreur est encore inférieur à l'écart entre les taux de MHV respectifs. Ceci indique que le taux d'erreur sur des mots rares (MHV en condition P0) doit être assez élevé. Il est intéressant de noter que le taux d'erreur pour P0-20k-Bref-T' est identique au taux d'erreur pour Q0-65k-Bref80-T'.

L'utilisation d'un modèle biclasse lors de la dernière passe du décodeur permet de réduire légèrement le taux d'erreur en évitant quelques erreurs d'accord en genre et en nombre. Avec les modèles entraînés sur Bref, et l'interpolation des modèles trigramme et biclasse, on obtient un taux d'erreur de 12,72% sur les 609 phrases (8,71% sur les 300 phrases T').

Analyse d'erreurs

Les erreurs sont dues en très large partie aux mots homophones (et quasi-homophones, par exemple `est et`), et, en proportion moindre aux noms propres. Les homophones (quasi-homophones) proviennent des formes fléchies pour les verbes, (`aller`, `allé`, `allée`, `allés`, `allées`, `allai`, `allais`, `allait`, `allaient`), des accords en genre et en nombre pour les noms et les adjectifs. Les formes verbales plus rares, donc moins bien apprises par le modèle de langage, sont souvent reconnues comme une suite de mots:

`alerta` → `alerte à`,
`plaça` → `place à`.

Les noms propres admettent eux aussi souvent des formes homonymes:

`Joseph Shuster` → `Josef Schuster`,
`Action Comics` → `action comme X`,
`Daily Planet` → `délit planète`.

Le problème des homophones est à résoudre au niveau du modèle du langage. Notre étude a d'ailleurs montré l'intérêt d'augmenter les corpus d'apprentissage pour le modèle de langage.

Des problèmes de reconnaissance se posent également sur des séquences de mots peu observées dans les textes (`papillonnant de coeur en coeur` → `papiers au nom de quelle rancoeur`).

En dehors des différents types d'erreurs observés globalement sur le corpus, on peut remarquer que les taux d'erreurs varient de manière importante d'un locuteur à l'autre (voir tableau 4). Ainsi on peut observer que pour presque la moitié des locuteurs le taux d'erreurs est inférieur à 10% (le meilleur résultat étant 5,7%, avec un taux de MHV de 0,4%), alors que le plus élevé est 30,4% (avec un taux de MHV de 4,8%). Ces variations sont dues d'une part à la conformité de la prononciation (par exemple le locuteur l16m n'est pas de langue maternelle française), et d'autre part à la difficulté des textes à lire. Un texte est d'autant plus difficile qu'il est mal appris par le modèle de langage et mal couvert par le lexique. Ceci est malheureusement le cas pour le locuteur le plus difficile, ayant à lire un résumé du film *FARGO* contenant un récit peu typique des textes du *Monde* avec une répétition des noms de personnages non répertoriés dans le lexique: *Marge Gunderson, Showalter, Grimsrud, Ludegaard*. La conjonction de difficultés de différents types conduit ainsi à des taux d'erreurs largement supérieurs à la moyenne.

Discussion

Grâce à ce travail, nous avons pu faire différents constats quant à l'apport supplémentaire de parole et de textes pour l'apprentissage des modèles. En ce qui concerne la modélisation acoustique, nos travaux ont montré, qu'en partant du corpus BREF, l'apport de données d'apprentissage supplémentaires ne permet plus d'améliorer les performances de reconnaissance de manière significative, au moins avec notre stratégie actuelle de construction de modèles. Concernant les modèles de langage, un ajout de textes d'apprentissage est d'autant plus intéressant, qu'il s'agit de données proches (dans le temps, et probablement aussi proche dans la thématique) des données de test. Les erreurs de reconnaissance du français journalistique proviennent pour une part assez faible des niveaux acoustico-phonétiques (modèles acoustiques, lexique de transcriptions phonétiques), la majeure partie étant due à la modèlisation du langage (allant de la syntaxe à la sémantique).

locuteur	*nb. phrases*	*err. mots (%)*	*MHV (%)*
l01f	45	8,3	0,4
l02f	36	8,1	1,3
l03m	36	17,9	2,7
l04f	40	9,9	1,6
l05f	52	8,4	0,9
l06m	18	5,7	0,4
l07m	17	8,9	0,0
l08m	33	16,5	1,1
l09f	20	24,1	2,6
l10m	21	10,1	1,2
l11m	38	12,1	1,4
l12m	38	12,9	2,1
l13m	16	13,1	2,2
l14m	51	15,3	1,8
l15m	15	8,0	0,5
l16m	19	30,4	4,8
l17f	52	20,5	1,6
l18m	15	14,2	1,1
l19f	32	11,9	2,4
l20f	15	6,7	1,0
Moyenne	609	12,72	1,5

Table 4: Résultats pour chacun des 20 locuteurs des données de développement dans les conditions suivantes: vocabulaire de 65k mots, apprentissage sur le corpus Bref, modèle de langage trigramme et biclasse.

Remerciements

Nous remercions Michèle Jardino pour son aide dans la construction des classes automatiques de mots.

Une partie de ce travail a été effectuée dans le cadre des Actions de Recherche Concertées « Linguistique, Informatique et Corpus Oraux » , financées par l'AUPELF-UREF.

Références

[Adda G. et al., (1997)] "*Ressources pour l'apprentissage, le développement et l'évaluation des systèmes de dictée vocale en français : corpus de texte, de parole et lexical*", JST97, avril 1997, Avignon.

[Boula de Mareüil P., (1996)] "*Pour une approche par règles en transcription graphème-phonème*", Séminaire GDR-PRC CHM Lexique et communication parlée (pp. 203-209). Toulouse 1996.

[Gauvain J.-L., Lamel L.F., Adda G., Adda-Decker M., (1994)] "*Speaker-Independent Continuous Speech Dictation*", Speech Communication, Vol. 15, No.1-2, Oct. 1994.

[Gauvain J.-L., Lee C.H., (1994)] "*Maximum A Posteriori for Multivariate Gaussain Mixture Observation of Markov Chains,*" IEEE Trans. on Speech and Audio Processing, pp. 291-298, 1994.

[Gauvain J.-L., Lamel L.F., Adda G., Matrouf D., (1996)] "*Developments in Continuous Speech Dictation using the 1995 ARPA NAB News Task,*" Proceedings ICASSP, 1996.

[Jardino M. (1996)] "*Multilingual stochastic n-gram class language models*", Proc. ICASSP'96.

[Lamel L.F., Gauvain J.-L., (1995)] "*A Phone-based Approach to Non-Linguistic Speech Feature Identification*", Computer Speech and Language, Vol. 9, pp. 87-103, janvier 1995.

[Lamel L.F., Adda-Decker M., Adda G., Gauvain J.-L., (1996)] "*Reconnaissance Multilingue de Grands Vocabulaires,*" livre AUPELF-UREF 1996).

[Legetter C.J., Woodland P.O., (1995)] "*Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models*", Computer Speech & Language, Vol. 9, pp. 171-185, 1995.

[Prouts B., (1980)] "*Contribution à la synthèse de la parole à partir du texte: Transcription graphème-phonème en temps réel sur microprocesseur*, Thèse de docteur-ingénieur, U. Paris XI, Nov. 1980.

[S. Young, P. Woodland, (1994)] "State clustering in hidden Markov model-based continuous speech recognition," Computer Speech & Language, Vol. 8, 1994.