Reconnaissance de la parole pour des locuteurs non natifs en présence de bruit

D. Fohr[1], O. Mella[1], I. Illina[1], F. Lauri[1], C. Cerisara[1], C. Antoine[2]

(1) LORIA, 615 rue du jardin botanique – 54602 Villers-lès-Nancy, France
(2) DIALOCA, 6 rue d'Uzès – 75002 Paris, France
Mél : {cerisara, fohr, illina, lauri, mella, antoinec}@loria.fr

ABSTRACT

In real world applications, speech recognition is confronted with two main difficulties : the non native speakers and the background noise. The aim of this paper is to compare on the same noisy database different methods in order to increase the robustness of our HMM-based automatic speech recognition system. To deal with the non native speakers, we have tested two solutions: multi-models and adaptation techniques. For noisy speech, we have evaluated two types of methods: the first one (PMC and MLLR) adapts the initial models, trained in clean speech, with a few noisy sentences. The second one (RATZ and MCR) tries to remove the noise from the signal without modifying the HMM models.

1. INTRODUCTION

Le développement de la mobilité des personnes et la portée de plus en plus internationale des applications de reconnaissance automatique de la parole (serveurs vocaux téléphoniques, contrôle aérien, location de voiture équipée de commandes vocales) font que ces applications sont confrontées simultanément à deux problèmes : celui des locuteurs non natifs et celui de la reconnaissance en milieu bruité. L'objectif de l'étude présentée dans cet article est l'évaluation de plusieurs méthodes de compensation ou d'adaptation permettant de résoudre ces problèmes.

Dans les deux premières sections, nous présentons brièvement le moteur de reconnaissance et les conditions de nos expérimentations. Puis, deux solutions pour mieux prendre en compte les locuteurs non natifs sont décrites dans la section 4. Enfin, quatre méthodes permettant la compensation ou l'adaptation au bruit ambiant sont évaluées dans la section 5.

2. CONDITIONS EXPÉRIMENTALES

Notre expérimentation a porté sur un corpus de test enregistré par quatre locuteurs français qui énoncent en anglais cent commandes vocales comportant en moyenne 4,5 mots. Le vocabulaire comprend les dix chiffres et 104 mots de commande. La perplexité de la grammaire associée à cette tâche est de l'ordre de 9.

A ce corpus enregistré à 16kHz a été ajouté un bruit stationnaire de moteur avec deux niveaux de rapport signal/bruit : un bruit à 17 dB appelé *bruit faible* dans la suite de l'article et un bruit à 7 dB appelé *bruit fort.*

Les applications réelles fondées sur des commandes vocales nécessitent la reconnaissance de la commande complète. Aussi avons-nous choisi de présenter les résultats en pourcentage de commandes correctement reconnues plutôt qu'en mots reconnus même si les écarts entre les résultats des méthodes sont amplifiés. Afin de pouvoir comparer les différents résultats obtenus, tous les tests sont effectués sur 50 phrases par locuteur, les 50 autres étant utilisées pour l'adaptation ou la compensation lorsque cela est nécessaire.

3. MOTEUR DE RECONNAISSANCE

Toutes les expérimentations décrites dans cet article ont été effectuées à l'aide du logiciel ESPERE. Cette boîte à outils que nous avons conçue autour d'un moteur de reconnaissance markovien permet l'apprentissage de modèles de Markov cachés et le test d'algorithmes de reconnaissance [1]. Pour nos expérimentations, nous avons choisi la paramétrisation acoustique suivante : 35 coefficients (12 MFCC et leurs coefficients de régression du premier et deuxième ordres, en omettant C0).

Les modèles de Markov choisis pour la reconnaissance sont des modèles phonétiques à 3 états multigaussiens indépendants du contexte. En effet, bien que le corpus soit constitué de commandes assez brèves, l'application visée doit pouvoir rester flexible et autoriser la modification du vocabulaire.

4. PROBLÈMES DES LOCUTEURS NON NATIFS

4.1 Introduction

Les locuteurs n'utilisant pas leur langue maternelle pour énoncer ces commandes, nous avons, dans une première étape, voulu tester la pertinence des modèles phonétiques utilisés dans la phase de reconnaissance.

Trois solutions sont envisageables. Le corpus étant en langue anglaise, la plus simple consiste à utiliser un ensemble de modèles phonétiques anglais associés à un lexique standard. Les locuteurs étant français, une autre solution est de fonder la reconnaissance sur un ensemble de modèles phonétiques français associé à un lexique contenant les mots du vocabulaire représentés avec les phonèmes français acoustiquement les plus proches des phonèmes anglais : ainsi le mot « *three»* est phonétisé [s R i]. Mais chaque locuteur peut plus ou moins bien parler anglais, voire même être anglophone ou bilingue, la meilleure solution ne serait-elle donc

pas de mixer les deux ensembles de phonèmes ?

Pour ce mixage, nous avons décidé de laisser, pour chacun des phonèmes à reconnaître, le moteur ESPERE choisir entre les modèles anglais et français. En effet, nous supposons, qu'au sein d'un mot, un locuteur peut articuler correctement un phonème anglais (par exemple une diphtongue) puis prononcer le phonème suivant « à la française » (par exemple « th »). Une illustration de ce mixage est donnée à la figure 1.

Toujours dans l'optique de mieux prendre en compte la prononciation des locuteurs non natifs, une deuxième étape consiste à adapter les différents ensembles de modèles phonétiques à chacun des locuteurs. Le paragraphe 4.2 présente les résultats de l'étude des trois solutions décrites ci-dessus et le paragraphe 4.3 ceux de l'adaptation au locuteur.

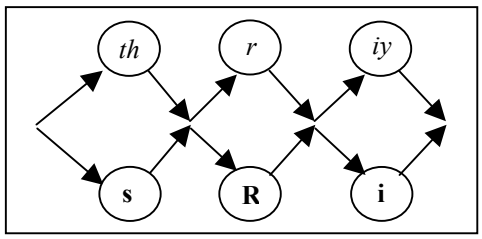

Figure 1: Représentation de « three » : en **gras** les phonèmes français, en *italique* les phonèmes anglais.

4.2 Comparaison des ensembles de modèles

Nous avons testé les trois solutions : modèles anglais, modèles français et modèles mixtes avec les lexiques associés. Les 49 modèles phonétiques anglais ont été appris sur le corpus anglo-américain TIMIT, les 31 modèles phonétiques français sur le corpus BREF80. La table 1. donne les taux de reconnaissance globaux des phrases (commandes vocales) en fonction de la solution choisie et du nombre de gaussiennes par état des modèles. Les mauvais scores obtenus avec les modèles phonétiques anglais s'expliquent par le fait que les locuteurs ont un mauvais accent et une prononciation « à la française ». C'est également pour cette raison que l'utilisation des modèles mixtes n'améliore pas les scores de reconnaissance par rapport aux modèles français.

Table 1 : Taux de reconnaissance des commandes vocales en % en fonction du nombre de gaussiennes et des modèles.

	1G	2G	4G	8G	16G
Français (FRA)	71	73	81	81	85
Anglais (ENG)	59	59	60	62	63
Mixtes (MIX)	68	67	71	82	84

En ce qui concerne le nombre de gaussiennes par état, l'utilisation de 16 gaussiennes apporte peu d'améliorations et nous avons décidé de nous limiter à 8 gaussiennes dans les expérimentations suivantes, ce qui est un bon compromis entre le taux de reconnaissance et le temps de calcul.

4.3 Adaptation au locuteur

Dans la seconde étape de notre étude, nous avons choisi d'adapter les modèles phonétiques avec la méthode MLLR (Maximum Likelihood Linear Regression) [2] en utilisant une matrice pleine avec biais. L'adaptation est effectuée en mode supervisé : la séquence de phonèmes prononcés est fournie au module d'adaptation.

Les modèles anglais et français sont adaptés séparément et les modèles adaptés mixtes sont l'union des modèles adaptés dans chaque langue.

La table 2. présente les taux de reconnaissance par locuteur sans adaptation et avec adaptation en utilisant 50 phrases par locuteur. Le test de reconnaissance est effectué sur les 50 autres phrases.

Table 2 : Taux de reconnaissance des commandes vocales en % avec et sans adaptation au locuteur.

Loc.	LOC 1		LOC 2		LOC 3		LOC 4		Tous	
	sans	avec	sans	avec	sans	avec	sans	avec	sans	avec
FRA	86	92	86	94	82	94	72	76	81	89
ENG	80	88	52	76	74	78	44	80	62	81
MIX	90	100	82	88	88	94	72	86	82	92

4.4 Discussion des résultats

L'adaptation au locuteur donne d'excellents résultats sur les modèles anglais ce qui s'explique par l'écart entre la prononciation des locuteurs français et les modèles anglais issus de TIMIT. Notons que l'adaptation est particulièrement efficace pour le plus « mauvais » locuteur (loc4) à condition, toutefois, d'avoir une quantité suffisante de données d'adaptation.

Nous l'avons vérifié en étudiant l'influence de la quantité de parole utilisée pour adapter les modèles sur les performances de reconnaissance. Quel que soit l'ensemble de modèles, l'utilisation de moins de 5 secondes de données d'adaptation dégrade les performances. Il y a en effet trop de paramètres à estimer (matrice de transformation pleine) avec trop peu de données. En revanche, à partir de 15s de parole, le taux de reconnaissance stagne. Cette asymptote peut s'expliquer par le fait que nous avons utilisé une seule matrice de transformation pour toutes les gaussiennes

Enfin, notre étude montre que l'application de l'adaptation au locuteur sur les modèles mixtes permet d'obtenir le meilleur taux de reconnaissance (92%). Nous pensons que les modèles mixtes permettent donc de prendre en compte les locuteurs non natifs ayant un niveau d'anglais très variable (locuteurs 1 et 4). Toutefois, nous allons enregistrer quelques locuteurs natifs pour nous assurer que les modèles mixtes permettent

aussi d'obtenir de bons résultats avec ceux-ci.

5. ADAPTATION AU BRUIT

5.1 Introduction

Un second problème souvent rencontré en reconnaissance de parole embarquée est la présence de bruit (habitacle de voiture, cockpit d'avion, téléphone mobile). Nous avons donc étudié différentes méthodes pour augmenter la robustesse au bruit. Ces méthodes se divisent en deux classes : celles de la première classe cherchent à débruiter le signal comme le font les méthodes de normalisation MCR et RATZ et les secondes à modifier les modèles markoviens appris en parole « propre » à partir de quelques secondes de parole bruitée, comme le réalisent les méthodes PMC et MLLR. Nous allons décrire dans les paragraphes suivants ces différentes méthodes avant de comparer leurs résultats dans le paragraphe 5.6.

5.2 RT-MCR

La normalisation RT-MCR (Real Time Mean Cepstre Removal) consiste à soustraire à chaque composante cesptrale sa valeur moyenne. Lors de la phase d'apprentissage, cette moyenne est calculée *a posteriori* sur toute la phrase alors que, pendant l'étape de reconnaissance, elle est évaluée de façon incrémentale en temps réel :

$M(t) = \frac{M(t-1)*99 + C(t)}{100}$ où C(t) est la composante cepstrale et M(t) sa moyenne à l'instant t.

5.3 RATZ

La méthode RATZ (Multivariate Gaussian Based Cepstral Normalisation) [3] permet de compenser les bruits additifs ou de convolution.

RATZ suppose que l'effet du bruit sur le signal non bruité peut être modélisé dans le domaine cepstral par des facteurs additifs. Ces facteurs sont estimés selon le maximum de vraisemblance puis sont soustraits aux vecteurs de cepstre de la parole bruitée afin de rapprocher ces derniers de la parole sans bruit. Les modèles acoustiques issus de l'apprentissage ne sont pas modifiés. Pour estimer de façon fiable ces facteurs, la parole non bruitée du corpus d'apprentissage est modélisée par un modèle de Markov simplifié (Gaussian Markov Model) qui est un modèle à un seul état. Cet état est représenté par une combinaison de K distributions gaussiennes :

$P(x)=\sum_{k=1}^{K} w_k \mathrm{N}(\mu_{k,x},\sigma_{k,x})$

où N est une distribution gaussienne. L'ensemble de paramètres $\Lambda=\{w_k,\mu_{k,x},\sigma_{k,x}\}$ de ce GMM est ensuite estimé selon le maximum de la vraisemblance.

Étant donné l'ensemble estimé Λ, la parole bruitée $Y=\{y_1,\ldots,y_T\}$ est supposée produite par la densité de probabilité suivante $P(y) = \sum_{k=1}^{K} w_k \mathrm{N}(\mu_{k,y},\sigma_{k,y})$

La méthode RATZ lie la parole bruitée à la parole non bruitée de la façon suivante :

$\mu_{k,y}=\mu_{k,x}+n_k$ et $\sigma_{k,y}=\sigma_{k,x}+R_k$

$\{n_1,\ldots,n_K\}$ et $\{R_1,\ldots,R_K\}$ sont les vecteurs représentant les facteurs correctifs. Le but de la méthode RATZ est de trouver les valeurs de ces facteurs qui maximisent la fonction de vraisemblance pour la parole bruitée. Cela est habituellement effectué en utilisant l'algorithme EM (Expectation Maximisation) [4].

Les facteurs estimés sont ensuite utilisés pour compenser le bruit présent dans l'ensemble Y et en déduire la parole débruitée $\overline{X}=\{\overline{x}_1,\ldots,\overline{x}_T\}$. Pour cela l'estimation au minimum de l'erreur (Minimum Mean Squared Error) est utilisée. Enfin cette parole débruitée $\overline{X}$ est passée au moteur de reconnaissance.

Nous avons appliqué cette méthode sur le corpus de test par lot de 50 phrases : les facteurs correctifs sont estimés sur 50 phrases, puis ces 50 phrases sont débruitées avant la reconnaissance.

5.4 PMC

PMC (Parallel Model Combination) est une méthode d'adaptation des modèles acoustiques à l'environnement. Proposée par M. Gales et S. Young en 1993, elle a subi par la suite plusieurs améliorations [5]. Le principe général de PMC consiste à construire des modèles des différents bruits affectant la parole dans l'environnement de test, puis à combiner ces modèles de bruit avec les modèles de parole issus de l'apprentissage.

Nos choix d'implémentation ont privilégié la rapidité de l'algorithme d'adaptation, aussi bien en temps de calcul qu'en terme de quantité de données d'adaptation requise: en particulier, nous nous sommes limités à une unique passe sur le signal de test, ce qui explique pourquoi nous ne considérons pas le bruit convolutif, qui lui, doit être estimé sur plusieurs secondes de parole.

De plus, nous avons choisi un modèle très simple pour le bruit additif, c'est-à-dire un vecteur spectral. Ceci nous permet notamment d'estimer ce modèle avec très peu de données et simplifie considérablement la combinaison des modèles. Celle-ci est réalisée en transformant les coefficients statiques des vecteurs moyennes des modèles dans le domaine spectral, puis en ajoutant dans ce même domaine le vecteur de bruit estimé, et, enfin, en projetant le vecteur résultat à nouveau dans le domaine cepstral. Cette transformation nécessite l'utilisation du coefficient C0 du cesptre ce qui a exigé d'apprendre de nouveaux modèles phonétiques.

Dans notre étude, ces nouveaux modèles phonétiques seront adaptés au bruit avec 0.2s de parole d'adaptation.

5.5 MLLR

Nous avons appliqué la méthode MLLR du paragraphe 4.3 pour adapter les modèles phonétiques de la parole « propre » au locuteur et au bruit. Pour un rapport signal/bruit donné, pour chacun des locuteurs, nous avons utilisé 50 phrases pour adapter les modèles et les 50 autres phrases pour le test de reconnaissance. Nous avons effectué 3 tests différents : (a) : les modèles sont adaptés uniquement au locuteur avec la parole non bruitée (MLLR loc dans la table 3.), (b) : l'adaptation est effectuée en une seule passe avec la parole du locuteur prononcée dans les mêmes conditions de bruit que celles du test (MLLR global), (c) : l'adaptation est effectuée en deux passes, les modèles sont adaptés au locuteur sur de la parole propre puis au bruit (MLLR 2 passes).

5.6 Résultats et discussion

La table 3. récapitule les taux de reconnaissance au niveau commande des différentes méthodes de compensation ou d'adaptation au bruit en utilisant les modèles phonétiques français.

La première ligne de la table donne les résultats obtenus lorsqu'aucune méthode n'est utilisée. Par ailleurs, dans le cas de PMC, nous avons utilisé les 200 premières millisecondes de chaque fichier de parole pour estimer le bruit. Cette transformation nécessitant l'utilisation du coefficient cepstral C0, nous avons ajouté dans la table les résultats obtenus par le système de base avec des modèles incluant le coefficient C0 (36 coefficients).

Table 3 : Taux de reconnaissance en % avec les modèles phonétiques français sur les 3 conditions de test.

	Sans bruit	Bruit faible	Bruit fort
Système de base	81	52	11
RT-MCR	84	65	20
RATZ+RT-MCR	82	68	35
Système de base avec C0	86	27	4
PMC avec C0	85	80	33
MLLR loc	89	53	14
MLLR global	89	79	31
MLLR 2 passes	89	80	38

Nous pouvons constater que la normalisation cepstrale (RT-MCR) apporte une amélioration en parole bruitée bien que cette méthode soit principalement conçue pour le bruit convolutif. L'application de RATZ améliore très légèrement les résultats précédents mais au détriment d'une quantité non négligeable de parole d'adaptation.

Dans le bruit faible, les méthodes PMC et MLLR global donnent des résultats satisfaisants et comparables (de l'ordre de 80%). Mais si on considère la quantité de parole nécessaire à l'adaptation, la méthode PMC est nettement plus performante. En effet, elle ne nécessite que 0,2s de parole d'adaptation et peut donc fonctionner en temps réel. Par ailleurs, adapter les modèles uniquement au locuteur (MLLR loc) ne rend pas plus robuste la reconnaissance en parole bruitée par rapport au système de base. En revanche, l'adaptation globale en deux passes améliore légèrement les résultats notamment pour le bruit fort.

Dans le bruit fort aucune des méthodes proposées n'atteint un score exploitable dans une application réelle (reconnaissance de la commande complète).

De plus, notons que si l'ajout du coefficient C0 augmente légèrement le taux de reconnaissance dans le cas de la parole propre, il le dégrade fortement dans le cas de la parole bruitée.

La table 4. montre que toutes ces conclusions restent valables avec des meilleurs scores de reconnaissance lorsqu'on utilise les modèles mixtes.

Table 4 : Taux de reconnaissance en % avec les modèles phonétiques mixtes

	Sans bruit	Bruit faible	Bruit fort
Système de base	82	63	22
RT-MCR	87	69	31
Système de base avec C0	85	18	2
PMC avec C0	84	78	29
MLLR loc	92	75	22
MLLR global	92	90	47

6 CONCLUSION

En ce qui concerne le problème des locuteurs non natifs, la méthode MLLR appliquée aux modèles mixtes permet d'obtenir un bon taux de reconnaissance en parole propre (92%). Pour traiter le problème du bruit additif, la même méthode appliquée à la fois au bruit et au locuteur permet d'obtenir un score comparable dans le bruit faible (90%), En conclusion, nous pouvons constater que l'application de MLLR sur les modèles mixtes permet de traiter simultanément le problème du locuteur non natif et du bruit faible stationnaire additif.

Une autre solution possible est de combiner les meilleures méthodes : utilisation des modèles mixtes, puis adaptation au bruit par PMC, et enfin, adaptation MLLR au locuteur et au bruit. Cette solution permet d'atteindre le score de 92% en bruit faible. Ce bon résultat cache cependant une forte disparité entre les bons locuteurs qui atteignent 98% de taux de reconnaissance et le « mauvais » locuteur qui plafonne à 74%.

BIBLIOGRAPHIE

[1] D. Fohr, O. Mella, C. Antoine : *The Automatic Speech Recognition Engine ESPERE : experiments on telephone speech*, ICSLP, pp 246-249, 2000.

[2] C.J. Leggetter and P.C. Woodland : *Maximum Likelihood Linear Regression for Speaker Adaptation of Continous*

Density Hidden Markov Models, Computer Speech. and Language, 9 : 171-185, 1995.

[3] P.J. Moreno and R.M. Stern : *Multivariate Gaussian-Based Cepstral Normalisation*, ICASSP, 1995.

[4] P.J. Moreno : *Speech Recongition in Noisy Environments* PhD Thesis, Carnegie Mellon University, 1996.

[5] M. Gales and S. Young, *Robust speech recognition in additive and convolutional noise using parallel model combination*, Computer Speech and Language 9, pp. 289-307, 1995.