Etude de la cohabitation entre la bande large et la bande étroite en reconnaissance automatique de la parole

Mohamed-Ali Ben Salah, Jean Monné*, Denis Jouvet* & Régine André-Obrecht***

Orange Lab., 2 avenue Pierre Marzin 22300 Lannion, France*
IRIT- Université Paul Sabatier, 118 route de Narbonne 31062 Toulouse Cedex 9**

mohamedali.bensalah@orange-ftgroup.com

ABSTRACT

In this paper we study the problem of cohabitation between narrowband and wideband speech for automatic speech recognition task. A series of speaker-independent, continuous speech phone recognition experiments have been carried out using the BREF80 and ESTER French corpus. Experiments performed used wideband speech typically sampled at 16 kHz and narrowband speech sampled at 8 kHz. The Aurora Advanced Front-End algorithm was taken as reference. This cohabitation study involves us addressing the problem of pseudo wideband signals where data are over-sampled from 8 kHz to 16 kHz. The strategy developed for the purpose of cohabitation shows its efficiency and leads to a better interaction of the automatic speech recognition system with the different speech data.

Keywords: wideband, narrowband, speech recognition, Aurora Advanced Front-end.

1. INTRODUCTION

Bien que la demande accrue pour les signaux codés en bande large et typiquement échantillonnés à 16 kHz (WB) commence à prendre du terrain et touche plusieurs applications et divers domaines des télécommunications [9], l'intérêt pour les signaux codés en bande étroite et spécifiquement échantillonnés à 8 kHz (NB) reste de nos jours d'actualité. En effet, il n'est pas réaliste à un horizon de quelques années que des applications comme la téléphonie IP en bande large aient totalement remplacé la téléphonie sur le réseau commuté en bande étroite. Dans un futur très proche, les systèmes de reconnaissance automatique de la parole (ASR) vont donc interagir avec de la parole codée en WB mais aussi en NB. Ainsi traiter la question de la cohabitation entre la bande large et la bande étroite en reconnaissance automatique de la parole s'avère d'une grande utilité et cela dans le but de garantir une réponse optimale des ASR face aux divers types de données parole.

Des travaux de la littérature précisent que contrairement à l'anglais, la WB n'est pas très bénéfique pour la langue française et que l'impact observé suite à divers expérimentations sur le corpus BREF80 [4] est très faible [5]. Gauvain et Lamel observent sur ce dernier une amélioration relative non significative du taux d'erreur phonème de 0,8% par rapport au même corpus échantillonné à 8 kHz. Aussi, afin de faciliter l'étude de la cohabitation entre signaux WB et NB, nous avons en premier lieu mesuré l'apport que la bande large peut introduire pour la reconnaissance automatique de la parole.

L'étude de cette cohabitation nous a amenés à poser le problème de la fausse bande large (FWB) où les données présentées comme des données WB sont en réalité issues d'un codage ou un transcodage bande étroite. Pour des problèmes d'architecture réseaux aucune signalisation concernant la FWB n'est fournie, or ce codage influence largement les performances des ASR: la plateforme de reconnaissance n'a aucun moyen de connaître l'origine de données traitées et ne peut donc pas répondre d'une manière optimale en présence de la fausse bande large. Nous proposons d'insérer dans l'ASR une détection de ces signaux FWB.

Le papier est organisé de la façon suivante. Après une description du contexte expérimental dans la deuxième partie, nous détaillons le système de reconnaissance développé pour gérer le problème de la cohabitation entre la bande large et la bande étroite. Les résultats expérimentaux obtenus ainsi que leur analyse, sont rassemblés dans la troisième partie.

2. CONTEXTE EXPÉRIMENTAL

2.1. Généralités

Afin de garantir une indépendance à toute tâche spécifique de reconnaissance, les expériences décrites dans la suite de l'article sont réalisées dans le cadre d'un décodage phonétique. Un jeu de 36 phonèmes est considéré:. Le système de reconnaissance automatique utilisé est le système RecoSoft, basé sur des modèles de Markov cachés (MMC) [8]. La modélisation acoustique prend en compte l'influence contextuelle des phonèmes et repose sur une modélisation GMM à 8 gaussiennes. L'analyse acoustique de référence est effectuée par l'intermédiaire de l'algorithme ETSI ES 202 050 [10] connu sous le nom de la paramétrisation Aurora WI008 et spécialement conçu pour les systèmes de reconnaissance distribués. Cet algorithme propose un module d'extraction des trames acoustiques pour les signaux échantillonnés à 8 kHz et à 16 kHz. Dans la norme Aurora, après débruitage du signal grâce à un filtrage de Wiener, la trame acoustique fournie par cet algorithme est constituée de 13 coefficients cepstraux, et du logarithme de l'énergie. Dans notre système les vecteurs acoustiques de référence sont composés de 33 paramètres: les 10 premiers coefficients MFCCs, le log-énergie plus leurs

dérivées premières et secondes. Nous appelons Aurora_8kHz et Aurora_16kHz les analyses spécifiques aux signaux échantillonnés à 8 kHz et à 16 kHz. Nous précisons que nous ne faisons appel à aucun modèle de langage dans nos expérimentations de décodage phonétique.

2.2. Description des corpus de parole

Les expériences sont réalisées sur les corpus BREF80 et ESTER. Toutes les données audio sont initialement échantillonnées à 16 kHz : ces signaux seront qualifiés de signaux WB

BREF80

Le corpus d'apprentissage est d'une durée de 2h42. 30 locuteurs et 29 locutrices prononcent chacun 67 phrases. Le corpus de test, est d'une durée de 12 minutes, il fait intervenir 4 hommes et 4 femmes ne figurant pas dans la base d'apprentissage.

ESTER

Le corpus d'apprentissage est construit a partir de la phase 1 et de la phase 2 de la campagne ESTER [4], il est composé de 69h11 de parole issue d'émissions radiophoniques prononcés par 1607 locuteurs. Le corpus de test représente 12h46 de parole.

Filtrage des signaux 16 kHz : les signaux FWB et NB

Afin de construire des données filtrées en bande étroite tous les signaux audio sont filtrés dans la bande [0-4kHz]. Nous appelons ces signaux filtrés les signaux NB. Ces derniers sont ensuite sur échantillonnés à 16 kHz dans le but de simuler des bases présentant de la parole en fausse bande large, les données FWB. Nous résumons dans le tableau 1 les caractéristiques de chacun des signaux utilisés dans nos expérimentations:

Tableau 1 : Caractéristiques des signaux

Type de Signal	Fréquence d'échantillonnage	Bande passante
WB	16 kHz	[0-8kHz]
NB	8 kHz	[0-4kHz]
FWB	16 kHz	[0-4kHz]

Construction d'une phonétisation de référence

La phonétisation de référence qui nous sert à calculer les taux d'erreurs est obtenue par alignement forcé du signal WB sur l'ensemble des variantes correspondant aux prononciations possibles des mots. Nous avons retenu pour chaque phrase parmi les variantes possibles, la phonétisation ayant la vraisemblance la plus élevée pour le modèle adapté sur le signal WB.

Evaluation des performances

Dans les expériences, nous donnons les performances de nos systèmes de reconnaissance en termes de PER (Phoneme Error Rate), le taux d'erreur en phonème (taux de substitution, d'insertion et d'omission) [5]. Les pauses introduites pour marquer l'instant où le locuteur a introduit un silence entre deux mots ne sont pas prises en compte dans cette mesure.

2.3. Initialisation et adaptation des modèles

Le laboratoire dispose de MMC entraînés sur des larges corpus de parole téléphonique. Ces modèles servent de modèles initiaux Mi_{NB} pour apprendre les modèles traitant les signaux NB. Le problème d'initialisation des modèles spécifiques aux signaux WB, que les données soient de vraies données ou de fausses données bande large, est plus délicat puisqu'on ne dispose pas de modèles entraînés spécifiques à la bande large. La procédure suivie afin d'initialiser et adapter les modèles est la suivante :

- adaptation des modèles Mi_{NB} aux données NB pour obtenir les modèles M_{NB},

- alignement des signaux avec les modèles M_{NB},

- initialisation des modèles WB à l'aide ces alignements,

- adaptation avec les signaux WB pour obtenir les modèles M_{WB},

- adaptation avec les signaux FWB pour obtenir les modèles M_{FWB}

Le même principe utilisant des techniques d'adaptation bayésienne et décrit dans [2] est respecté pour les processus d'adaptation.

3. RÉSULTATS ET DISCUSSIONS

3.1. Mesure de l'apport de la bande large

Les variations du taux d'erreur, au cours de ces expériences sont toujours calculées par rapport au taux PER obtenu sur les signaux NB avec les modèles M_{NB}. Tous les résultats sont rassemblés dans le tableau 2.

A la lecture du bloc 1 du tableau 2, où figurent les performances du système de référence 'signaux NB, modèle M_{NB}' et du système 'signaux WB, modèle M_{WB}', il apparait clairement que l'utilisation des signaux WB améliore les performances de reconnaissance. Sur le corpus BREF80, une réduction relative d'environ 10% du taux d'erreur en phonème est observée. Le même constat se confirme aussi pour le corpus ESTER puisque le taux d'erreur en phonème est réduit de 2,1 points ce qui correspond à une réduction relative du PER de 5%. Pour justifier l'apport de la bande large mesuré par notre système de reconnaissance et se positionner vis-à-vis des ASRs utilisant un modèle de langage notamment aux systèmes de Gauvain et Lamel [5], de Boite et Ris [1] et de Leblouch et Collen [6] [7], nous avons effectué quelques évaluations avec un trigram phonétique comme modèle de langage appris sur le Corpus "Le Monde".

Tableau 2: Comparaison pour les différents signaux et méthodes sur BREF80 et ESTER (sans modèle langage)

			BREF80		ESTER	
	Signal	**Modèle**	PER	Réduction PER	PER	Réduction PER
1	**NB**	M_{NB}	26.1 *[25.8 - 26.4]*	-	38.9 *[38.7 – 39.0]*	-
	WB	M_{WB}	23.6 *[23.3 - 23.8]*	10%	36.7 *[36.6 - 36.9]*	5%
2	**FWB**	M_{WB}	42.4 *[42.0 - 42.7]*	-64%	43.8 *[43.6 - 43.9]*	-13%
		M_{FWB}	28.1 *[27.8 - 28.4]*	-9%	39.3 *[39.2 - 39.4]*	-1%
3	**FWB**	M_{WB+FWB}	28.3 *[28.0 - 28.5]*	-9%	39.9 *[39.8 – 40.0]*	-3%
	WB		25.3 *[25.0 - 25.5]*	2%	38.7 *[38.6 - 38.9]*	0%
4	**FWB**	Sélection WB / FWB M_{WB} ou M_{NB}	26.1 *[25.8 - 26.4]*	0%	39.3 *[39.1 - 39.4]*	-1%
	WB		23.6 *[23.3 - 23.8]*	10%	36.7 *[36.6 – 36.9]*	5%

Dans [5], le système proposé combine des modèles de phonèmes en contexte, un modèle de langage bigram phonétique et une adaptation au genre du locuteur; ce système atteint un taux d'erreur en phonème de 21,3% sur BREF80. Boîte et Ris proposent une reconnaissance associant HMM et MLP et obtiennent un taux d'erreur de 25% sur ce même corpus. Dans [6] et [7], l'information syllabique est utilisé afin de guider le processus du décodage phonétique. Avec le système syllabique et un modèle de langage en bigram syllabique le taux d'erreur obtenu est de 15,8% sur BREF80 et 22,34% sur ESTER. L'analyse de ces résultats montre clairement que les performances de notre ASR sur la bande large sont comparables avec ces travaux et que contrairement à ce qui est annoncé dans [5] l'apport de la WB est aussi significatif pour la langue française.

Les résultats reportés dans le tableau 3 montrent que, même avec le trigram phonétique, la réduction du PER apportée par la bande large est du même ordre (10%). De ce fait, pour toutes les comparaisons entre la bande large et la bande étroite, aucun modèle de langage ne sera utilisé dans toutes les expérimentations décrites par la suite.

Tableau 3 : performance de l'ASR sur BREF80 avec un trigram phonétique

Corpus	Modèle	PER	Réduction PER
BREF80 (8 kHz)	M_{NB}	22.7 [22.4 - 22.9]	-
BREF80 (16 kHz)	M_{WB}	20.5 [20.2 - 20.7]	10%

3.2. Problème de la fausse bande large

Dans le bloc 2 du tableau 2, sont données les performances obtenues avec notre ASR sur les signaux FWB, en utilisant les modèles M_{WB} et M_{FWB}. Les résultats montrent une dégradation importante des performances avec les modèles M_{WB}. Sur le corpus BREF80 une détérioration relative du PER de 64% est observée. La dégradation relative du taux d'erreur en phonème sur le corpus de test ESTER est de 13%. La différence de comportement sur les deux corpus étudiés

peut être expliquée par le fait que les données ESTER sont issues des transcriptions d'émissions radiophoniques et elles présentent divers conditions d'enregistrement (studio, téléphone) contrairement à BREF80 où les signaux sont en pure bande large. Le modèle M_{WB} adapté sur ce dernier corpus et contrairement à celui adapté sur ESTER est testé dans des conditions totalement différentes de celles d'apprentissage. C'est pour cela que nous avons pensé à construire des modèles propre à la FWB, et évaluer leurs performances. Toutefois, même avec des modèles de reconnaissance adaptés sur la fausse bande large M_{FWB}, on n'arrive pas à atteindre les taux d'erreurs de référence surtout sur le corpus BREF80 où la dégradation est de 9%.

Afin d'améliorer la réponse du système et garantir une réponse plus optimale devant les signaux en fausse bande large, une première solution consiste à mélanger les bases de parole en vraie et fausse bande large dans le but d'adapter le système à des données représentant les différentes conditions de test. Cette technique a été utilisée pour traiter le problème de la cohabitation entre les données parole provenant d'un réseau téléphonique commuté RTC et les données GSM. On note par la suite M_{WB+FWB} le modèle adapté sur un corpus d'adaptation mixte.

Dans le bloc 3 du tableau 2, il apparaît nettement que l'adaptation des modèles sur des corpus de parole mixte permet une amélioration des résultats en ce qui concerne les données en fausse bande élargie. Toutefois, les performances restent non satisfaisantes puisqu'on perd plus de la moitié du gain apporté par la bande large.

3.3. Détection parole en fausse bande large

La stratégie que nous avons développée repose sur l'identification de l'origine du signal présenté à l'entrée de la reconnaissance. De cette manière, après identification de l'origine, il est possible de mettre en adéquation paramétrisation et modèle : un signal WB

sera reconnu par le modèle WB et le signal FWB sera filtré dans la bande [0 - 4kHZ] pour être traité par le modèle NB. Précisons le déroulement de cette procédure :

Le problème de la détection des signaux échantillonnés à 16 kHz peut être formalisé simplement selon le test d'hypothèses suivant:

$$\begin{cases} H_0 : x \text{ est un signal bande large} \\ H_1 : x \text{ est un signal en fausse bande large} \end{cases}$$

Le signal x est classé comme un signal vraie bande large si et seulement l'hypothèse H_0 est vraie. L'hypothèse validée correspond au maximum de vraisemblance, à savoir l'hypothèse H_0 est vraie si et seulement si:

$$P(H_0|x) > P(H_1|x) \quad (1)$$

En supposant les deux hypothèses équiprobables, et en utilisant la formule de Bayes, on se ramène à

$$P(x|H_0) > P(x|H_1) \quad (2)$$

Ce qui est équivalent à :

$$P(x_{WB}|M_{WB}) > P(x_{WB}|M_{FWB}) \quad (3)$$

où x_{WB} est l'observation obtenue par l'analyse Aurora_16kHz de x. Il s'en suite que :

$$\sum_{\phi} P(x_{WB},\phi|M_{WB}) > \sum_{\phi} P(x_{WB},\phi|M_{FWB}) \quad (4)$$

où ϕ désigne de manière générale une suite de phonèmes, Si on approxime la somme sur toutes les suites possibles par le terme prédominant qui correspond à la suite la plus probable, l'hypothèse H_0 est vraie si et seulement si:

$$\underset{\phi}{Max}\, P(x_{WB},\phi|M_{WB}) > \underset{\phi}{Max}\, P(x_{WB},\phi|M_{FWB}) \quad (5)$$

Finalement la règle de décision s'écrit :

$$\ln\left(P(x_{wB},\hat{\phi}_{WB}|M_{WB})\right) > \ln\left(P(x_{wB},\hat{\phi}_{FWB}|M_{FWB})\right) \quad (6)$$

avec $\hat{\phi}_{WB}$ et $\hat{\phi}_{FWB}$ les suites les plus probables respectivement pour les deux modèles M_{WB} et M_{FWB}

Les résultats obtenus donnent un taux de bonne détection atteignant 100% sur BREF80 et 89% sur ESTER WB. L'investigation des performances données par le bloc 4 du tableau 2 montre qu'avec cette technique on arrive à égaler les PER de référence sur les deux corpus (bloc 1 tableau 2).

4. CONCLUSION

Dans ce papier, nous étudions le problème de la cohabitation entre la bande large et la bande étroite en reconnaissance automatique de la parole, et de manière plus appronfondie la question de la fausse bande large. Une technique permettant au système de reconnaissance de mieux interagir avec des signaux parole en vraie bande large et des signaux en fausse bande large, i.e. des issus d'un codage NB à 8kHz mais présentés comme des données échantillonnées à 16 kHz est proposée. Les résultats obtenus sur les deux corpus de test BREF80 et ESTER montrent la validité de l'approche consistant à détecter l'origine des données parole WB au moyen de la comparaison des vraisemblances de décodage obtenus par des modèles de reconnaissance concurrents, afin d'utiliser le modèle le plus adéquat en fonction du type du signal à reconnaître.

BIBLIOGRAPHIE

[1] J.M. Boite et C. Ris. Development of a French speech recognizer using a hybrid HMM/MLP system. ESANN, 1999.

[2] L. Delphin-Poulat. Comparison of Techniques for Environment/Application Adaptation in a Telephony Context. ITRW, August 2001.

[3] G. Gravier, J-F. Bonastre, E. Geoffrois, S.Galliano, K. McTait and C. Khalid. The Ester Evaluation Campaign for the Rich Transcription of French Broadcast news, Proc. Language Evaluation and Resources Conference, 2004.

[4] L.F. Lamel, J.L. Gauvain, M. Eskénazi. BREF, a large vocabulary Spoken Corpus for French. EUROSPEECH, 1991.

[5] L.F. Lamel, J.L. Gauvain. High Performance Speaker-Independent Phone Recognition Using CDHMM. EUROSPEECH, 1993.

[6] O. Leblouch, P. Collen. Reconnaissance Automatique de phonèmes guidée par les Syllabes. JEP, 2006.

[7] O. Leblouch, P. Collen. Automatic Syllabe-Based Phoneme Recognition Using ESTER Corpus. ISCGAV, 2007.

[8] L.R. Rabiner, B.H. Juang. An Introduction to Hidden Markov Models. IEEE ASSP Magazine, January 4-15, 1986.

[9] M.L. Seltzer, A. Acero. An EM Algorithm for Training Wideband Acoustic Models from Mixed-Bandwidth Training Data", Proc. IEEE ASRU 2005, pp. 197-202, November 2005.

[10] Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm. ETSI ES 202 050, 2003.