Modélisation phonotactique de grandes classes phonétiques en vue d'une approche différenciée en identification automatique des langues

Jérôme FARINAS[1], Régine ANDRÉ-ORECHT[1]

[1]Institut de Recherche en Informatique de Toulouse

Équipe ART.ps – CNRS UMR 5505
118, route de Narbonne
31062 Toulouse Cedex 4

{jerome.farinas,obrecht}@irit.fr

Résumé – La plupart des systèmes d'Identification Automatique des langues accordent une grande importance au niveau phonotactique, en utilisant des modèles N-gram et des dictionnaires phonétiques de grande taille. Cependant, il est évident que l'introduction d'autres paramètres (acoustiques, phonétiques et prosodiques) améliorera les performances. Récemment, nous avons proposé un modèle phonétique alternatif qui exploite une discrimination voyelle/non-voyelle. Nous complétons ici cette étude en étudiant le niveau phonotactique et la différenciation en grandes classes phonétiques. Nous utilisons un modèle de langage n-multigramme sur des grandes classes phonétiques. Nous présentons ici une étude basée sur un décodeur de grandes classes phonétiques dans un but d'identification des langues.

Abstract *– Most systems of Automatic Language Identification give a great importance to the phonotactic level, by using N-gram models and relatively large phone-dictionary sizes. However, it is obvious that introducing other features (acoustic, phonetic and prosodic) will improve performances. Recently, we have proposed an alternative acoustic phonetic model which exploits the vowel / non-vowel distinction. Here we complete this preliminary system, by studying the phonotactical level and phonetic differentiation. We used a n-multigram model on broad phonetic categories. We present here a study based on an automatic broad phonetic decoder in an ALI task.*

1. Introduction

Parmi les différentes sources d'information disponibles pour identifier un langage donné, les informations phonotactiques, relatives aux règles qui gouvernent la combinaison des sons dans une langue, contribuent grandement à la décision d'identification [1]. Les systèmes actuels les plus performants en témoignent largement en privilégiant cette source de connaissances et sa modélisation [2].

Les études menées à l'IRIT en Identification Automatique des Langues, ont pour but d'exploiter le maximum de sources. C'est pourquoi nous portons nos efforts sur la modélisation acoustico-phonétique, la modélisation phonotactique, la modélisation prosodique et la fusion de ces informations. Une première étude nous a conduit à proposer une approche différenciée au niveau acoustico-phonétique. Pour prendre en compte les paramètres structuraux des systèmes phonologiques, deux espaces, l'espace vocalique et l'espace consonantique, sont modélisés par deux modèles distincts pour chacune des langues. L'identification est obtenue par fusion adéquate des scores ainsi obtenus [3]. Cette approche a montré une amélioration des résultats comparativement à une modélisation acoustique globale. Cette approche remet en cause la modélisation phonotactique. Après avoir étudié l'influence d'une telle séparation en classes phonétiques sur une modélisation phonotactique dans un cadre idéal [4], nous détaillons dans cet article le comportement de la modélisation sur de séquences de grandes classes phonétiques obtenues à partir d'un décodeur acoustico-phonétique automatique. L'architecture globale du système d'identification de la langue est détaillée en section 2, la modélisation du décodeur acoustico-phonétique grandes classes en section 3, le modèle multigramme en section 4. Le protocole expérimental en décrit en section 5 et les résultats sont présentés en section 6 et nous les discutons en section 7.

2. Système d'identification de la langue

Le système d'identification de la langue est ici réduit à deux parties : un décodeur acoustico-phonétique (DAP) grandes classes indépendant de la langue et une modélisation de la langue par n-multigrammes. La sortie du DAP alimente le modèle de langage. La décision d'identification est réalisée en comparant les scores de vraisemblance de chaque langue du modèle multigramme.

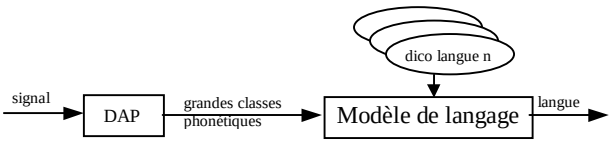

FIG. 1 : Schéma global du système d'identification.

3. Décodeur acoustico-phonétique

3.1 Détection de l'activité vocale

Pour pouvoir détecter les zones de non-activité vocale, nous avons développé un détecteur d'activité basé sur une analyse statistique du premier ordre du signal temporel.

Si l'on note N le nombre de segments issus de la segmentation automatique et $\{S_1,...,S_N\}$ la suite de ces segments.

Le seuil d'activité S_a est défini par :

$$S_a = \min_{S_i}(\mathrm{var}(S_i)) \quad (1)$$

Les segments ayant une variance inférieure à S_a sont étiquetés comme étant des silences. Un post-traitement permet de regrouper les segments de non-activité en cas de sur-segmentation : si plusieurs segments courts sont étiquetés comme étant des silences et ont une durée totale supérieure à 120 ms, on considère qu'il s'agit d'une zone de non-activité.

3.2 Modèles de grandes classes phonétiques

Le décodeur acoustico-phonétique est basé sur l'emploi de Modèles de Markov Cachés (MMC). 4 grandes classes phonétiques seront modélisées : voyelles, fricatives, occlusives et sonantes. Le voisement est considéré pour la modélisation des fricatives et des occlusives. Nous séparons le modèle des occlusives en deux pour distinguer la zone de silence avant explosion et l'explosion.

4. Modèle multigramme

Pour rendre compte des différentes règles qui gouvernent la combinaison des phonèmes d'une langue. Nous utilisons un modèle de langage multigramme [5] qui permet de détecter des motifs récurrents dans des suites d'observations. Ces motifs récurrents peuvent avoir une longueur variable.

La modélisation par multigrammes consiste à trouver la segmentation $S=(s_1,\dots, s_{n(S)})$ la plus probable d'une séquence d'observations $O=(o_1,\dots, o_T)$:

$$S^* = \arg\max \ \ell(O,S) \quad (1)$$

avec la vraisemblance $\ell(O,S) = \prod_{i=1}^{n(S)} P(z_i)$ (2)

où $Z_i = (o_{si},\dots, o_{s(i+1)-1})$ (3)

L'algorithme d'apprentissage est un algorithme itératif de type EM. A chaque itération, sont estimées les probabilités *a priori* d'une séquence d'observations Z_i :

$$P^{(k+1)}(z_i) = \frac{c(z_i / S^{*(k)})}{c(S^{*(k)})} \quad (4)$$

avec $S^{*(k)} = \arg\max_S \ \ell^{(k)}(O,S)$ (5)

où $\ell^{(k)}(O,S)$ est la vraisemblance de la séquence d'apprentissage O à l'itération k, $c(z_i / S^{*(k)})$ est le nombre d'occurrences de z_i dans la segmentation optimale $S^{*(k)}$. La segmentation la plus probable $S^{*(k)}$ est estimée en utilisant un algorithme de Viterbi. Au cours de ces itérations, les segmentations du corpus évoluent, faisant émerger les séquences d'observation les plus typiques.

Après apprentissage, un dictionnaire est créé contenant les séquences Z_i les plus probables et leur vraisemblance.

La phase de reconnaissance consiste à calculer la perplexité d'une séquence d'observation O en utilisant la segmentation la plus vraisemblable, suivant la formule :

$$PP_{VI}(O) = 2^{-\frac{1}{T}\log \ell^*(O)} \quad (6)$$

où T est le nombre d'observations de O et où la vraisemblance de cette même suite est :

$$\ell^*(O) = \arg\max_S \ \ell(O,S) \quad (7)$$

5. Expériences

5.1 Description du corpus

Les expériences sont menées sur six langues du corpus OGI Multi Language Telephone Speech : l'anglais, l'allemand, l'hindi, le japonais, le mandarin et l'espagnol. Il s'agit d'un corpus de parole téléphonique conversationnelle. Le corpus comprend une centaine de fichiers de 45 s de parole non contrainte par langue.

Les données d'apprentissage du DAP correspondent aux transcriptions phonétiques réalisées manuellement par des experts phonéticiens [6]. Ces transcriptions, réalisées au format international Worldbet [7], sont ensuite réduites en grandes classes phonétiques.

Le corpus est scindé en deux parties, l'une destinée à être utilisée pendant la phase d'apprentissage et l'autre pendant la phase de test. Il y a environ 50 locuteurs par parties. Les deux parties sont indépendantes, on ne retrouve pas de locuteur commun entre les deux sous-corpus.

5.2 Décodeur Acoustico-Phonétique

5.2.1 Paramètres

Pour extraire des vecteurs caractéristiques du signal de parole, nous avons utilisé un fenêtrage de Hamming sur 25 ms, avec décalage de 10 ms. Chaque vecteur est constitué de 9 Mel-Frequency Cepstral Coefficients (MFCC) et de leur dérivée première. L'énergie du signal a été normalisée sur tout le fichier audio.

L'extraction de paramètres n'est réalisée que sur les zones de parole repérées par le détecteur d'activité vocale (cf. 3.1) : les silences de plus de 120 ms sont rejetés.

5.2.2 Modèles de Markov Cachés

Trois types de MMC ont été utilisés :
- pour les sonantes (S), voyelles (V), fricatives voisées ou non-voisées (F) : MMC à trois états
- pour les zone d'occlusion (OP) : MMC à deux états
- pour l'explosion et le relâchement (P) : MMC à un seul état.

Il y a donc trois états pour chaque grande classe phonétique (voyelles, sonantes, fricatives et occlusives), ce qui revient a imposer une durée minimum de 45 ms minimum pour parcourir le MMC.

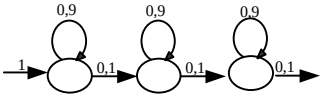

FIG. 2 : MMC à 3 états émetteurs (utilisé pour les voyelles, les fricatives et les sonantes).

5.2.3 Architecture du réseau global

Le réseau global permet de boucler indifféremment sur les grandes classes phonétiques (voyelles, fricatives, occlusives et sonantes). Les occlusives sont constituées de la séquence OP et P, c'est-à-dire la séquence d'occlusion et d'explosion. Notez que OP peut être voisé ou non-voisé, indifféremment du voisement ou non de P (par exemple on permet la séquence OP voisé / P non voisé). Les fricatives peuvent être voisées ou non.

5.2.4 Apprentissage

L'apprentissage des modèles est réalisé en trois temps :
- l'initialisation des paramètres des modèles en utilisant un algorithme de Viterbi
- la réestimation modèle par modèle en utilisant un algorithme de Baum-Welch
- la réestimation en une passe de tous les modèles simultanément par un algorithme de Baum-Welch.

5.2.5 Reconnaissance

La reconnaissance est réalisée en utilisant un algorithme de Viterbi : il fait correspondre à chaque fichier le réseau global de MMC, et extrait une transcription avec la vraisemblance de chaque mot.

5.3 Modèle de langage

Les transcriptions en grandes classes phonétiques des signaux de parole ont été utilisées en entrée du modèle de langage multigramme.

Nous avons utilisé des modèles n-multigrammes, avec n compris entre 3 et 5. N correspond à la longueur maximum de la séquence d'itération. Nous avons fait varier le nombre d'occurrences au-dessous desquelles une séquence de mots est incluse dans l'inventaire initial des séquences et celui de l'inventaire des séquences durant les itérations entre 2 et n. Nous avons utilisé 10 itérations pour l'apprentissage.

5.4 Décision d'identification

Pour chaque langue, on réalise un modèle phonotactique multigramme. Le problème d'identification consiste alors à trouver la langue qui maximise la probabilité d'observation de la séquence O :

$$L^* = \arg\max_L P(O/L) \quad (8)$$

soit

$$L^* = \arg\max_L PP_{VI}(O/L) \quad (9)$$

La décision est prise pour chaque segment du fichier de parole contenant une activité vocale et des silences inférieurs à 120 ms. La longueur moyenne des fichiers est de 5s.

6. Résultats

6.1 Décodeur Acoustico-Phonétique

Le score de reconnaissance moyen sur toutes les langues est de 63,8 %. L'*accuracy* (prise en compte des insertions dans les scores de reconnaissance) moyen est de 42,8 %. Si l'on ne considère plus le voisement (sur les fricatives et les occlusives), le score de reconnaissance et l'*accuracy* moyen sont respectivement de 70,4 % et de 50,0 %. Le détail par langues est listé dans TAB. 1 et TAB. 2.

TAB. 1 : détail par langue du taux de reconnaissance et *accuracy* pour le DAP grande classes complet.

	Angl.	**Allem.**	**Hindi**	**japo.**	**Mand.**	**Espa.**
Taux reco.	64,3 %	63,9 %	67,1 %	62,2 %	61,6 %	64,9 %
Accuracy	46,6 %	38,0 %	35,9 %	48,6 %	34,6 %	43,3 %

TAB. 2 : détail par langue du taux de reconnaissance et *accuracy* pour le DAP grande classes sans prise en compte du voisement.

	Angl.	**Allem.**	**Hindi**	**japo.**	**Mand.**	**Espa.**
Taux reco.	71,5 %	69,7 %	74,7 %	68,5 %	65,9 %	69,3 %
Accuracy	55,0 %	46,4 %	46,9 %	56,2 %	40,8 %	48,0 %

TAB. 3 : matrice de confusion des grandes classes phonétiques pour l'anglais.

	V	**F_NV**	**F_V**	**S**	**P_NV**	**P_V**	**OP_V**	**OP_NV**	**Silence**	**Omissions**
V	**6435**	208	207	277	172	157	54	77	76	*121*
F_NV	27	**1661**	67	37	78	47	15	34	29	*245*
F_V	34	152	**660**	45	72	103	36	35	20	*350*
S	204	160	253	**2777**	112	138	48	58	47	*116*
P_NV	31	148	37	21	**1440**	211	8	6	16	*242*
P_V	25	74	58	30	171	**746**	2	16	15	*232*
OP_V	16	22	23	16	15	0	**743**	204	10	*287*
OP_NV	16	34	21	16	7	23	237	**1647**	30	*377*
Silence	16	92	31	15	107	54	32	74	**677**	*148*
Insertions	*481*	*769*	*414*	*403*	*514*	*428*	*510*	*759*	*331*	

6.2 Identification de la langue

Le meilleur résultat en identification de la langue sur les 6 langues est 22,2 % d'identification correcte, en utilisant des 4-multigrammes.

TAB. 4 : Matrice de confusion de l'identification sur 6 langues

	angl.	**allem.**	**hindi**	**japo.**	**mand.**	**espa.**
anglais	**200**	172	145	140	169	156
allemand	115	**145**	70	50	74	64
hindi	47	42	**66**	83	56	64
japonais	25	38	31	**69**	38	32
mandarin	37	60	43	52	**71**	54
espagnol	112	117	95	78	132	**131**

7. Discussion et perspectives

Les résultats du décodeur acoustico-phonétique sont plutôt bons, particulièrement quand on ne considère pas le voisement des grandes classes phonétiques. Par contre les résultats en identification des langues sont assez moyens : l'information phonotactique n'a pu être correctement conservée : l'étude [4] qui portant sur l'identification des langues avec une modélisation multigramme sur des transcriptions manuelles présentait de bien meilleurs résultats. Le taux assez élevé d'insertion du DAP automatique grandes classes peut expliquer cette baisse de performance en identification.

Pour améliorer les résultats en identification, l'insertion d'un module de modélisation acoustique permettrait de filtrer les erreurs du DAP. Ce dernier pourrait également être perfectionné en rejetant des détections peu fiables, en se basant sur le score de vraisemblance et les n-meilleurs choix renvoyés par le DAP.

Remerciements

Nous remercions le groupe *Speech Vision and Robotics* de l'université de Cambridge pour la diffusion gratuite en *open source* du logiciel Hidden Markov Model Toolkit (HTK)[1]. Merci au *Centre for Speech Technology* au KTH à Stockholm de diffuser sous licence GNU le logiciel Wavesurfer[2] qui nous a permis d'analyser les signaux de parole.

Références

[1] Timothy J. Hazen, et Victor W. Zue. *Segment-based automatic language identification.* Journal of the Acoustical Society of America, Vol. 101, No. 4, pp. 2323-2331, 1997.

[2] Driss Matrouf et al. *Comparing different model configuration for language identification using a phonotactic approach.* Actes Eurospeech'99, pp 387-390, Budapest, 1999.

[3] François Pellegrino, Jérôme Farinas et Régine André-Obrecht. *Identification automatique des langues par une modélisation diférenciée des systèmes vocaliques et consonantiques.* Actes congrès Reconnaissances des Formes et Intelligence Artificielle, Paris, 2000.

[4] Jérôme Farinas et Régine André-Obrecht. *Variation sur les multigrammes,* XXIIIèmes Journées d'Etude sur la Parole, Aussois, 2000.

[5] Sabine Deligne. *Modèles de séquence de longueur variables : application au traitemant du langage écrit et de la parole.* Thèse de 3ème cycle, Ecole Nationale Supérieure des Télécommunications, Paris, 1996.

[6] Terri Lander. *The CSLU Labeling Guide.* Rapport interne, Center for Spoken Language Understanding, Oregon Graduate Institute, 1997.

[7] James L. Hieronymous. Ascii phonetic symbols for the world's languages: WolrdBet. rapport interne, Bell Labs, 1993.

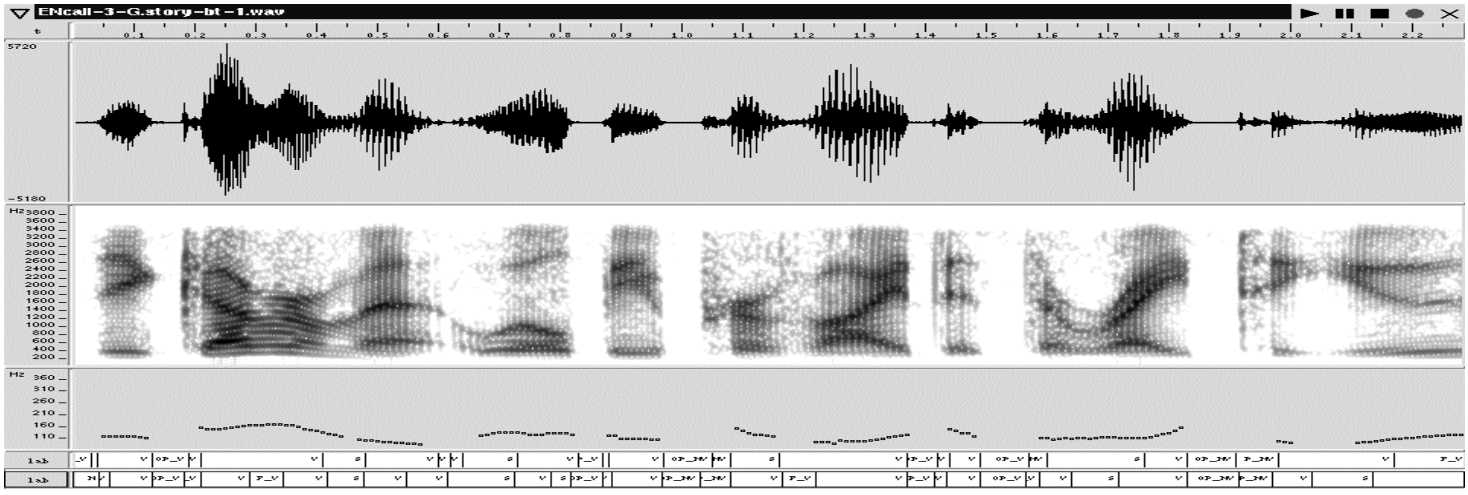

FIG. 3 : exemple de transcription grandes classes du DAP (de haut en bas : signal, spectrogramme, fréquence fondamentale, transcription manuelle, transcription

[1] http://htk.eng.cam.ac.uk/

[2] http://www.speech.kth.se/wavesurfer/