Identification automatique des langues: combinaison d'approches phonotactiques à base de treillis de phones et de syllabes

Dong Zhu, Martine Adda-Decker

LIMSI-CNRS, Université de Paris Sud,
BP 133, 91403 Orsay Cedex, France
dong.zhu@limsi.fr, madda@limsi.fr

ABSTRACT

This paper investigates the use of phone and syllable lattices to automatic language identification (LID) based on multilingual phone sets (73, 50 and 35 phones). We focus on efficient combinations of both phonotactic and syllabotactic approaches. The LID structure used to achieve the best performance within this framework is similar to PPRLM (parallel phone recognition followed by language dependent modeling) : several acoustic recognizers based on either multilingual phone or syllabe inventories, followed by language-specific n-gram language models. A seven language broadcast news corpus is used for the development and the test of the LID systems. Our experiments show that the use of the lattice information significantly improves results over all system configurations and all test durations. Multiple system combinations further achieve improvements.

1. INTRODUCTION

L'identification automatique des langues (IAL), qui consiste à déterminer la langue utilisée par un locuteur inconnu est un domaine de recherche très actif. Depuis une dizaine d'années, l'approche PPRLM [1] [2] [5] [9], qui consiste à mettre plusieurs décodeurs acoustiques en parallèle, utilise les modèles acoustiques à base de HMM (Hidden Markov Models), s'avère très performante pour l'IAL. Cette approche exploite de manière implicite un niveau acoustico-phonétique et de manière explicite un niveau phonotactique (via les modèles de langage n-grammes). D'autres niveaux peuvent contribuer à identifier la langue. Des recherches récentes montrent que la combinaison de différents niveaux d'information peut améliorer les performances d'identification. Ainsi les indices prosodiques peuvent contribuer à l'IAL [10], et l'approche PPRLM s'améliore avec l'intégration des modèles HMMs prosodiques [11] ; l'utilisation de SVM (Support Vector Machines) exploite la capacité de classification discriminante du système d'IAL [12] et pour l'évaluation du NIST (National Institut of Standards and Technology, les Etas-Unis) 2003, le meilleur système d'IAL est celui qui combine trois approches : PPRLM, GMM et SVM. Récemment, l'introduction de treillis de phones à la place de la meilleure séquence de phones décodée et l'utilisation de réseaux de neurones dans le module de décision [5] ont donné les meilleurs résultats sur le corpus NIST 2003.
Dans nos expériences récentes [7], nous avons présenté la modélisation de jeux de phones multilingues, l'introduction de l'approche syllabotactique pour l'IAL, ainsi que la supériorité des modèles acoustiques en contexte sur ceux indépendants du contexte. Les travaux antérieurs donnent aussi une comparaison préliminaire entre l'approche phonotactique et syllabotactique.
Les études présentées ici s'inscrivent dans la continuité de nos expériences précédentes. Elles essayent d'estimer l'apport de différentes configurations de système d'IAL et de combinaisons de plusieurs approches. Les travaux se concentrent sur plusieurs questions. D'abord, pour le système d'IAL à modèles acoustiques multilingues, l'utilisation de treillis peut-elle améliorer la représentativité des modèles de langage n-grammes ? Ensuite, concernant l'approche syllabotactique, est-ce que sa combinaison avec l'approche phonotactique peut améliorer les performances d'identification par rapport aux deux autres approches ? Si oui, pour la structure PPRLM, quelle est la meilleure combinaison de décodeurs phonétiques et syllabiques ?

2. CORPUS ET APPROCHE GÉNÉRAL

2.1. Corpus

Un corpus multilingue d'émissions de radio et de télévision a été collecté. Ce corpus offre plusieurs avantages pour l'IAL : la grande quantité de données est favorable à l'apprentissage du système ; la qualité du corpus est intéressante pour l'apprentissage des modèles acoustiques multilingues. Nous utilisons des corpus en sept langues : arabe classique, anglais américain, allemand, espagnol, français, italien, chinois mandarin, avec environ 20 heures par langue. Les corpus français et arabe sont des ressources fournies par la DGA. L'anglais, l'espagnol et le chinois sont extraits des corpus HUB4 du LDC. Les corpus allemands et italiens sont issus de divers projets européens FP5 LE (Olive, Alert) ou obtenus auprès d'ELDA. Pour ces corpus, des transcriptions orthographiques sont disponibles, et des lexiques de prononciation (phonémique, syllabique) ont été adaptés selon le choix de phones multilingues.
Le corpus d'apprentissage est divisé en deux parties. La moitié du corpus, soit environ 10 heures par langue, sert à l'apprentissage des modèles acoustiques multilingues. Le reste du corpus est utilisé pour l'apprentissage des modèles de langage spécifiques à chaque langue. Les corpus de test sont d'environ 30 minutes par langue, et ils sont divisés en segments de différentes durées : 3 secondes, 10 secondes, 20 secondes et 30 secondes.

2.2. *Inventaires multilingues phonétiques et syllabiques*

Concernant les jeux de phones multilingues, ils ont été définis à partir de huit jeux de phonèmes dépendants de la langue. Ici la huitième langue est le portugais, que nous n'avons pas gardé pour les tests, par manque de corpus adéquats. Des regroupements ont été effectués sur des critères linguistiques pour aboutir à un inventaire de soixante-treize phones multilingues (29 voyelles,40 consonnes). Pour la suite, nous avons exploré différents regroupements et les classifications des phones multilingues pour des études comparatives avec un nombre variable de phones. Ainsi, nous avons regroupé 29 voyelles en 6 classes en fonction de leurs valeurs de formants (F1, F2), cela donne un jeu de 50 phones gardant de nombreuses distinctions de consonnes, mais seulement 6 voyelles. Ensuite, la similarité des lieux d'articulation est la règle utilisée pour diminuer le nombre de consonnes. Finalement un jeu de 35 phones est obtenu qui contient 6 voyelles, 25 consonnes, et 4 phones spéciaux pour modéliser silence, hésitation, bruit, et respiration (table 1).
Le recensement des syllabes est effectué sur le corpus transcrit en syllabe d'après les règles de syllabation établies pendant les travaux précédents [7]. Les dictionnaires syllabiques ont été définis (9788 syllabes pour 73 phones, 9536 syllabes pour 50 phones, et 7712 syllabes pour 35 phones), en fonction de critères d'équilibre de répartition des syllabes sur les huit langues, et de taux de couverture. La fréquence d'occurrence comme critère de sélection permet d'éliminer syllabes rares, difficiles à modéliser avec une approche statistique, ainsi que des syllabes erronées.

2.3. *Approche générale*

Pour les trois jeux de phones multilingues décrits précédemment, des modèles acoustiques en contexte ont été estimés (3843 modèles acoustiques en contexte pour les 73 phones, 3455 modèles pour les 50 phones, 3415 modèles pour les 35 phones). Les modèles acoustiques sont des modèles HMMs à trois états, et chaque état contient trente-deux gaussiennes. Des modèles de langage phonétiques et syllabiques tri-grammes sont estimés à partir de données issues du décodage sans contrainte de modèles de langage. Les décodeurs acoustiques servent simplement à transformer les signaux acoustiques de parole en une suite de phones ou de syllabes. L'identité de la langue est ensuite obtenue en calculant le maximum de vraisemblance entre les modèles de langage spécifiques à chaque langue et les unités issues du décodage. Au lieu de se limiter à la meilleure solution, l'approche par treillis permet d'exploiter également les hypothèses sous-optimales générées lors du décodage acoustico-phonétique (acoustico-syllabique). L'utilisation de treillis permet ainsi d'améliorer la représentativité des hypothèses produites par les décodeurs acoustiques, en maximisant l'espérance du logarithme de la vraisemblance de la séquence des phones (ou syllabes) décodés (équation 1). L* représente la langue identifié, H représente une des séquences de phones (ou de syllabes) décodées et L représente la langue correspondante.

$$L^* \simeq \underset{L}{argmax}\ E_H[\log P(H|L)] \qquad (1)$$

Deux types de structures d'IAL sont proposés ici pour l'IAL : PRLM (Phone Recognition Followed by Language Modeling) et PPRLM. Dans le cadre du PRLM, neuf systèmes d'IAL ont été mis en place : (1) trois systèmes (73, 50 et 35 phones) d'IAL phonotactiques sans treillis, (2) trois systèmes d'IAL phonotactiques à base de treillis, (3) trois systèmes d'IAL syllabotactiques à base de treillis. Dans le cadre du PPRLM, les systèmes d'IAL bi-décodeurs, tri-décodeurs, quadri-décodeurs (figure 1) et penta-décodeurs sont mis en œvre pour isoler le système d'IAL le plus performant.

TAB. 1: Tableaux des jeux de phones multilingues.

Phones	73	50	35
Voyelles	a ʌ$_{gp}$ ʌ$_{ce}$ a : ã$_f$ ã$_p$	a	a
	e ɛ ẽ æ ə : ɛ : ɜˑ ɣ	e	e
	i ĩ ɪ ɨ	i	i
	o ɔ õ	o	o
	u ũ ʊ	u	u
	y ʏ ə ø ʔ	y	y
Semi-voyelles	w	w	w
	ɥ	ɥ	ɥ
	j	j	j
Consonnes	c ç	c ç	c
	ʧ	ʧ	ʧ
	s ʂ θ	s ʂ θ	s
	z ʐ ð	z ʐ ð	z
	ʃ	ʃ	ʃ
	ʒ	ʒ	ʒ
	f	f	f
	v	v	v
	m	m	m
	m̩	m̩	m̩
	n	n	n
	n̩	n̩	n̩
	l ɭ	l ɭ	l
	l̩	l̩	l̩
	R r χ h ħ r$_e$	R r χ h ħ r$_e$	r
	p	p	p
	b	b	b
	t ʈ	t ʈ	t
	d ɖ	d ɖ	d
	k q	k q	k
	g	g	g
	ɲ	ɲ	ɲ
	ŋ	ŋ	ŋ
spéciaux	.! & W	.! & W	.! & W

3. EXPÉRIENCES MENÉES

Des expériences ont été menées afin de comparer les performances de différentes combinaisons et approches. Au total, 19 systèmes d'IAL sont construits. Les tests se font sur les segments de 3, 10, 20 et 30 secondes, et les données de test distinctes des données d'apprentissage sont issues des mêmes types de sources que ces données. Les corpus du test, soit environ une heure par langue, sont divisés en segments de durées variées, 3 secondes (de 3s à 5s), 10 secondes (de 10s à 12s), 20 secondes (de 20s à 25s), 30 secondes (de 30s à 35s).

3.1. *Différents jeux de phones multilingues*

Comme le montrent les figure 2 et 3, parmi les systèmes d'IAL de différents jeux de phones multilingues, le sys-

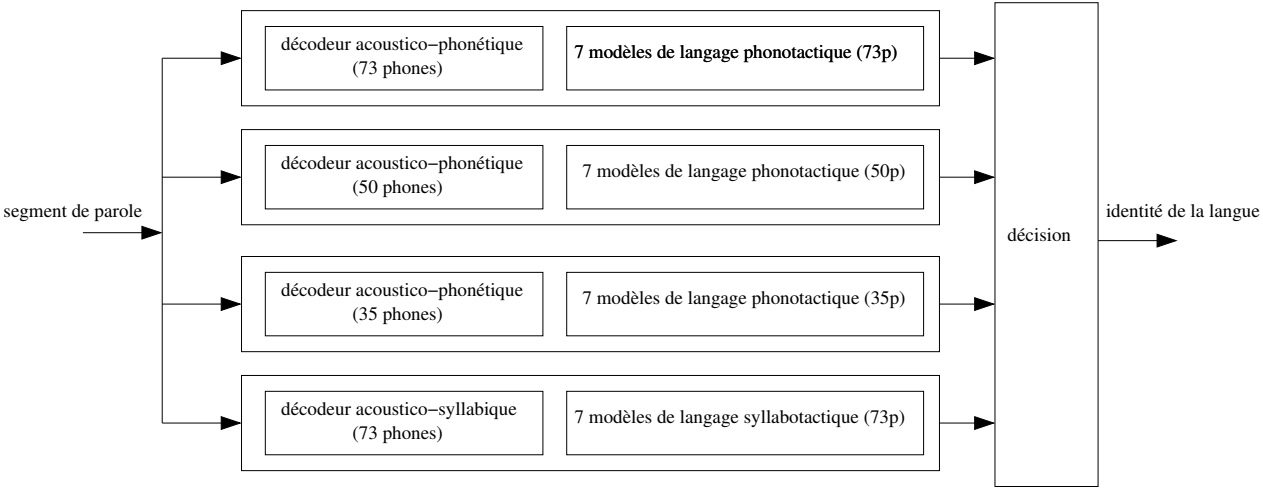

FIG. 1: Le système d'IAL (PPRLM) le plus performant combine l'approche phonotactique et l'approche syllabotactique utilisant les modèles acoustiques multilingues en

tème de 73 phones obtient le meilleur résultat (au niveau phonotactique aussi bien que syllabotactique). Les systèmes d'IAL de 73 et 50 phones sont généralement meilleurs que les systèmes de 35 phones, pour lequel les regroupements consonantiques doivent être remis en question.

3.2. Utilisation de treillis

Comme le montre la figure 2, sur tous les segments de test (3s, 10s, 20s, 30s) les systèmes phonotactiques d'IAL à base de treillis sont meilleurs que les systèmes sans treillis. On constate que sur les segments de test courts (3 secondes), le système d'IAL de 73 phones à base de treillis obtient un gain de 0,9% sur le système d'IAL de 73 phones sans treillis ; des gains peuvent également être mesurés pour le système d'IAL de 50 phones (0,2%) et le système de 35 phones (0.9%). Sur les segments de test longs (30 secondes), le gain est plus fort pour les systèmes avec moins de phones (2,5% pour 35 phones) que pour ceux avec plus de phones (0,3% pour 50 phones, 0,3% pour 73 phones). Les taux d'erreur d'identification sont présentés en détail dans la table 2.

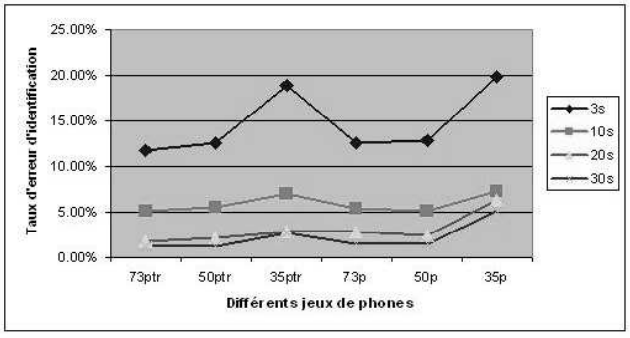

FIG. 2: Taux d'erreur d'identification des systèmes phonotactiques avec ou sans treillis, 73ptr signifie le système phonotactique de 73 phones avec treillis, 73p signifie 73 phones sans treillis.

3.3. Utilisation de l'approche syllabotactique

Le système d'IAL syllabotactique utilise les mêmes modèles acoustiques multilingues que le système phonotactique, mais le décodeur acoustico-syllabique produit des syllabes au lieu des phones, et les modèles de langage syllabotactiques se basent sur des milliers de syllabes. Nous avons illustré dans la figure 3 les résultats d'identification des systèmes syllabotactiques avec treillis et pour comparaison, nous avons ajouté les résultats des systèmes phonotactiques. Au niveau syllabotactique, le système d'IAL de 73 phones est le plus performant sur tous les segments de test (3s, 10, 20s et 30s). Il reste cependant un peu moins performant que le meilleur système phonotactique.

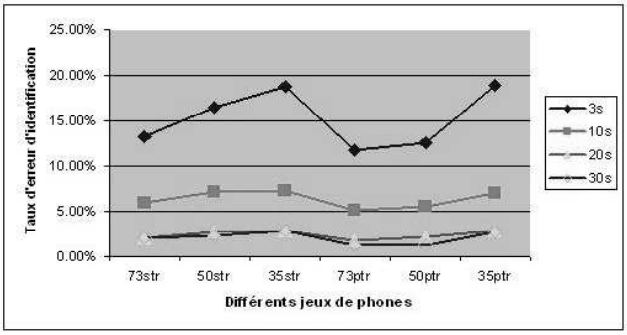

FIG. 3: Taux d'erreur d'identification des systèmes phonotactiques et syllabotactiques (décodage avec treillis). 73ptr signifie le système phonotactique de 73 phones, 73str signifie le système syllabotactique de 73 phones.

3.4. Combinaison de différents décodeurs

La mise en parallèle de plusieurs décodeurs donne une structure PPRLM, qui s'avère très efficace pour diminuer les erreurs. Différents systèmes d'IAL sont mis en œvre : système d'IAL de bi-décodeurs, de tri-décodeurs, quadri-décodeurs et penta-décodeurs. Comme le montre la table 2, la combinaison de plusieurs décodeurs améliore légèrement la performance d'IAL et tend à gommer la différence entre les différents systèmes. En effet pour les monodécodeurs et pour 3 secondes de test, les taux d'erreur d'iden-

tification varient entre 11,7% et 18,9% (13,2% et 18,8%) pour les systèmes phonotactiques avec treillis (syllabotactique avec treillis respectivement). Le système d'IAL qui met quatre décodeurs acoustiques (décodeurs acoustico-phonétiques de 73, 50 et 35 phones, décodeur acoustico-syllabique de 73 phones) en parallèle obtient le meilleur taux d'erreur d'identification de 0,9% sur les segments de 30 secondes. Au-delà, le système d'IAL avec cinq décodeurs n'arrive pas à améliorer davantage les résultats.

4. CONCLUSIONS ET PERSPECTIVES

Nous avons présenté différentes configurations de systèmes d'IAL, faisant toutes appel à des modèles acoustiques multilingues dépendant du contexte. Différents inventaires « phonémiques » multilingues, distinguant un nombre de (classes de) voyelles et de consonnes, varient globalement du simple au double ont été mis à l'épreuve. Des approches phonotactique et syllabotactique ont été utilisées de manière classique (en exploitant la meilleure séquence de phones/syllabes décodée) ou bien en exploitant la méthode à base de treillis, qui permet de mieux exploiter l'information produite lors du décodage phonétique/syllabique. Les inventaires plus grands (50, 73) produisent de meilleurs résultats pour l'approche phonotactique. Globalement l'utilisation de treillis de phones et de syllabes améliore les taux d'identification dans toutes les conditions et pour toutes les durées de test. De manière générale, l'approche syllabotactique n'est pas aussi performante que l'approche phonotactique, mais en situation de combinaison avec l'approche phonotactique elle permet d'obtenir des gains légers dans toutes les configurations impliquant le système syllabotactique à 73 unités de phones. Les expériences avec bi-décodeurs, tri-décodeurs, quadri-décodeurs et penta-décodeurs confirment la fiabilité de la structure PPRLM. Dans ce cadre-là, le système d'IAL fusionnant le décodeur acoustico-phonétique de 73 phones, 50 phones et 35 phones, et le décodeur acoustico-syllabique de 73 phones (figure 1) devient le système le plus performant sur les segments longs (supérieurs à 20s). Pour les travaux futurs, nous prévoyons de tester nos systèmes d'IAL sur plus de langues, et de les examiner sur des corpus de différents types comme la parole spontanée.

TAB. 2: Taux d'erreur d'identification pour 7 langues. 73p : décodage avec treillis, 73s : syllabotactique avec treillis ; la combinaison des systèmes est marqué par +.

Système avec treillis	3s	10s	20s	30s
73p+50p	11,6%	5,4%	2,0%	1,5%
73p+73s	13,9%	6,2%	2,5%	1,5%
35p+35s	15,7%	6,7%	2,7%	1,5%
73s+50s+35s	14,2%	6,4%	2,4%	2,3%
73p+50p+35p	11,6%	5,4%	2,1%	1,5%
73p+50p+35p+73s	11,7%	5,4%	1,8%	0,9%
73p+50p+35p+50s	11,7%	5,7%	2,1%	1,4%
73p+50p+35p+35s	12,0%	5,6%	2,4%	1,5%
73p+50p+35p+73s+35s	12,1%	5,4%	2,1%	1,2%

RÉFÉRENCES

[1] C. Corredor-Ardoy. et al. Language Identification with Language-independent acoustic models. *In Proc. Eurospeech*, Grèce, 1997.

[2] M.A. Zissman. Comparison of Four Approaches to Automatic Language Identification of Telephone Speech. *In IEEE Trans*, on Speech and Audio Processing 4(1), 1996.

[3] W.M. Cambell. et al. Language Recognition with Support Vector Machine. *In Proc. ODYSSEY*, Espagne, 2002.

[4] M. Adda-Decker. et al. Phonetic Knowledge, Phonotactics and Perceptual Validation for Automatic Language Identification. *In Proc. ICPHS*, Espagne, 2003.

[5] J.L. Gauvain. et al. Language Recognition using Phone Lattices. *In Proc. ICSLP*, Corée, 2004.

[6] N. Thangavelu. et al. Language Identification Using Parallel Syllable-like Unit Recognition. *In Proc. ICASSP*, Canada, 2004.

[7] D. Zhu. et al. Different Size Multilingual Phone Inventories and Context-Dependent Acoustic Models for Language Identification. *In Proc. Eurospeech*, Portugal, 2005.

[8] B. Ma. et al. An Acoustic Segment Modeling Approach to Automatic Language Identification. *In Proc. Eurospeech*, Portugal, 2005.

[9] P. Matejka. et al. Phonotactic Language Identification using High Quality Phoneme Recognition. *In Proc. Eurospeech*, Portugal, 2005.

[10] J.L. Rouas. Modeling Long and Short-Term Prosody for Language Identification. *In Proc. Eurospeech*, Portugal, 2005.

[11] Y. Obuchi. et al. Language Identification Using Phonetic and Prosodic HMMs with Feature Normalization. *In Proc. ICASSP*, Etats-Unis, 2005.

[12] C. White. et al. Discriminative classifiers for language recognition. *In Proc. ICASSP*, Etats-Unis, 2006.