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 approach 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 achieves 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 utilise les modèles acoustiques à base de HMM (Hidden Markov Models), s'avère très performante pour l'IAL. Mais plusieurs niveaux peuvent être exploités pour identifier la langue, des recherches récentes montrent bien que la combinaison de différents niveaux d'information peut améliorer beaucoup la performance d'IAL. Les indices prosodiques contribuent effectivement 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], 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 la configuration d'identification la plus performante sur le corpus NIST 2003.

Nos expériences récentes [7] présentent la modélisation des jeux de phones multilingues, l'introduction de l'approche syllabotactique pour l'IAL, ainsi que la supériorité des modèles acoustiques en contexte sur ceux indépendant 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 dont les modèles acoustiques sont multilingues, l'utilisation des treillis peut-elle améliorer la représentativité des hypothèses des modèles de langage n-grammes? Ensuite, l'approche syllabotactique étant une solution d'IAL fiable, sa combinaison avec l'approche phonotactique améliore-t-elle encore la performance d'IAL 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 GÉNÉRALE

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 d'IAL; la grande qualité du corpus est aussi favorable pour l'apprentissage des modèles acoustiques multilingues. Nous utilisent des corpus multilingues pour 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 langages 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. Approche générale

Trois jeux de phones multilingues [7] sont définis à partir de ces langues pour l'apprentissage des modèles acoustiques, ainsi que trois jeux de syllabes multilingues correspondant à ces phones (9788 syllabes pour 73 phones, 9536 syllabes pour 50 phones, et 7712 syllabes pour 35 phones). Les modèles acoustiques en contexte sont des modèles HMMs à trois états, et chaque état contient trente-deux gaussiennes. Les modèles de langage phonétiques et syllabiques sont tri-grammes, et estimés à partir de données issues du décodage sans la contrainte de modèles de langage. Pour nos systèmes d'IAL, les décodeurs acoustiques servent 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.

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 sont 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 aussi évalués pour isoler le système d'IAL le plus performant.

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 sont issues des mêmes sources que les données d'apprentissage. Les corpus du test sont extraits d'un corpus de cinq heures. En moyenne, 126 segments par langue sont utilisés pour le test de 3 secondes, 135 segments par langue pour le test de 10 secondes, 86 segments par langue pour le test de 20 secondes, 60 segments par langue pour le test de 30 secondes.

3.1. Utilisation de treillis

L'utilisation de treillis vise à améliorer la représentativité des hypothèses produites par les décodeurs acoustiques. Cela consiste à maximiser l'espérance du logarithme de la vraisemblance de la séquence des phones (ou syllabes) décodés (1). L^* représente la langue identifié, H représente une séquence des phones (ou syllabes) et L représente la langue correspondante.

$$L^* \cong \arg\max_L E_H \left[\log P\left(H|L\right)\right] \qquad (1)$$

Comme la figure 2 le montre, sur tous les segments de test (3s, 10s, et 20s) les systèmes phonotactiques d'IAL à base de treillis sont toujours 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 1,6% sur le système d'IAL de 73 phones sans treillis ; par contre le gain est plus fort pour le système d'IAL de 50 phones (11,1%) et le système de 35 phones (7%). Sur les segments de test longs (30 secondes), le gain est aussi plus fort pour les systèmes d'IAL avec moins de phones (2,8% pour 50 phones, 6,2% pour 35 phones) que pour ceux avec plus de phones (0,2% pour 73 phones). Les taux d'identification sont présentés en détail dans la table 1.

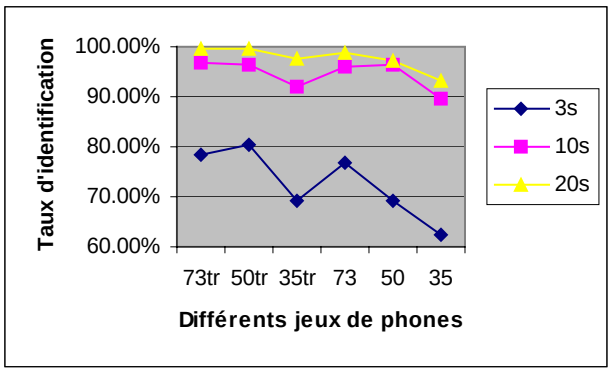

Figure 2 : Taux moyens d'identification des systèmes phonotactiques avec ou sans treillis, 73tr signifie 73 phones et treillis.

3.2. Utilisation de l'approche syllabotactique

Le système d'IAL syllabotactique utilise les mêmes modèles acoustiques multilingues que le système phonotactique. Mais dans le système syllabotactique, 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. Au niveau syllabotactique, le système d'IAL de 35 phones est le plus performant sur tous les segments de test (3s, 10, 20s et 30s). Pour les segments de courte

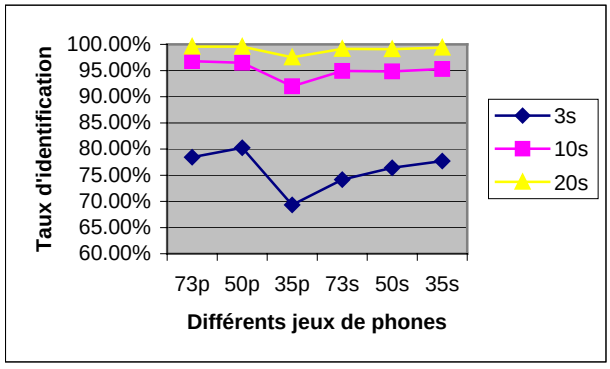

Figure 3 : Comparaison des approches phonotactiques et syllabotactiques (décodage avec treillis). 73p signifie 73 phones et l'approche phonotactique, 73s signifie 73 phones et l'approche syllabotactique.

durée, l'approche phonotactique est plus performante que l'approche syllabotactique. Par contre, sur les segments de test plus longs, comme 30 seconde, le système d'IAL syllabotactique de 35 phones obtient le même taux d'identification de 100% que le système phonotactique. Il est intéressant de combiner ces deux approches différente pour améliorer l'IAL.

3.3. Combinaison de différents décodeurs

Depuis une décennie, la structure PPRLM s'avère la plus performante pour l'IAL. Différents systèmes d'IAL sont estimés : système d'IAL de bi-décodeurs, de tri-décodeurs, quadri-décodeurs et penta-décodeurs. Comme le montre la table 1, la combinaison de plusieurs décodeurs améliore légèrement la performance d'IAL et tend à gommer la différence entre les différents systèmes : pour 3 secondes de test les taux d'identification varient entre 76,9% et 79,6%. Alors qu'ils sont compris entre 62,3% et 80,2% (74,1% et 77,7%) pour les systèmes phonotactiques avec et sans treillis (syllabotactique avec treillis respectivement). Par exemple, sur le segment de 10 secondes, le système d'IAL, qui combine le système de 73p (73 phones et phonotactique) et 50p, obtient un gain de 0,43% sur le système de 73p, et de 0,74% sur celui de 50p. En plus sur le segment de 20 secondes, ce système de bi-décodeurs obtient seulement un gain de 0,13%. Plus les segments de test sont longs, moins la combinaison de systèmes gagne en performance. Le système d'IAL qui met quatre décodeurs acoustiques (décodeurs acoustico-phonétiques de 73, 50 et 35 phones, décodeur acoustico-syllabique de 35 phones) en parallèle obtient le meilleur taux d'identification de 100% sur les segments de 20 et 30 secondes. Au-delà, le système d'IAL avec cinq décodeurs n'arrive pas à améliorer encore la performance d'IAL.

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 qui varie en gros 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 petits inventaires de phones (ici 35) sont mieux adaptés pour l'approche syllabique, alors que les inventaires plus grands (50, 73) produisent de meilleurs résultats pour l'approche phonotactique. L'approche syllabotactique donne les meilleurs résultats pour le jeu de 35 phones multilingues, avec des résultats variant en gros de 78% à 100% d'identification correcte pour des durées de test allant de 3 secondes à 30 secondes. Globalement l'utilisation de treillis de phones et de syllabes améliore significativement les taux d'identification dans toutes les conditions et pour toutes les durées de test, avec des gains absolus plus importants pour les durées courtes. Dans la plupart des cas, 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 35 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 35 phones (figure 1) devient le système le plus performant sur les segments longs (supérieurs à 20s). Pour les travaux futurs, il reste à tester nos systèmes d'IAL fusionnés sur plus de langues, et il est aussi important de les examiner sur les corpus du différent type comme la parole spontanée.

Table 1 : Taux moyen d'identification pour 7 langues. 73p : approche phonotactique à 73 phones à base ; 73p-tr : plus décodage treillis, 73s-tr syllabotactique et treillis ; la combinaison des systèmes est marqué par +.

	3s	10s	20s	30s
Systèmes phonotactiques avec/sans treillis				
73p-tr	78,4	96,8	99,6	100,0
73p	76,8	95,9	99,0	99,8
50p-tr	**80,2**	**96,5**	**99,6**	**100,0**
50p	69,1	96,3	97,1	97,2
35p-tr	69,3	92,0	97,6	100,0
35p	62,3	89,7	93,4	93,8
Systèmes syllabotactiques avec treillis				
73s-tr	74,1	94,9	99,1	99,3
50s-tr	76,4	94,8	99,1	99,8
35s-tr	**77,7**	**95,3**	**99,4**	**100,0**
Combinaisons de systèmes				
73p-tr + 50p-tr	79,4	97,2	99,7	100,0
73p + 50p	76,9	96,4	99,5	99,2
73p-tr + 73s-tr	77,2	95,9	99,0	100,0
50p-tr + 50s-tr	79,2	95,9	99,5	100,0
35p-tr + 35s-tr	77,3	94,2	99,4	100,0
73s-tr + 50s-tr + 35s-tr	79,1	94,8	99,6	100,0
73p-tr + 50p-tr + 35p-tr	79,6	97,3	99,8	100,0
73p-tr + 50p-tr + 35p-tr+73s-tr	77,9	96,7	99,6	100,0
73p-tr + 50p-tr + 35p-tr+35s-tr	**80,0**	**96,9**	**100,0**	**100,0**
73p-tr + 50p-tr + 35p-tr + 73s-tr+35s-tr	79,1	96,5	99,8	100,0

BIBLIOGRAPHIE

[1] C. Corredor-Ardoy., et al. Language Identification with Language-independent acoustic models. *In Proc. Eurospeech*, Grèce, 1994.

[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

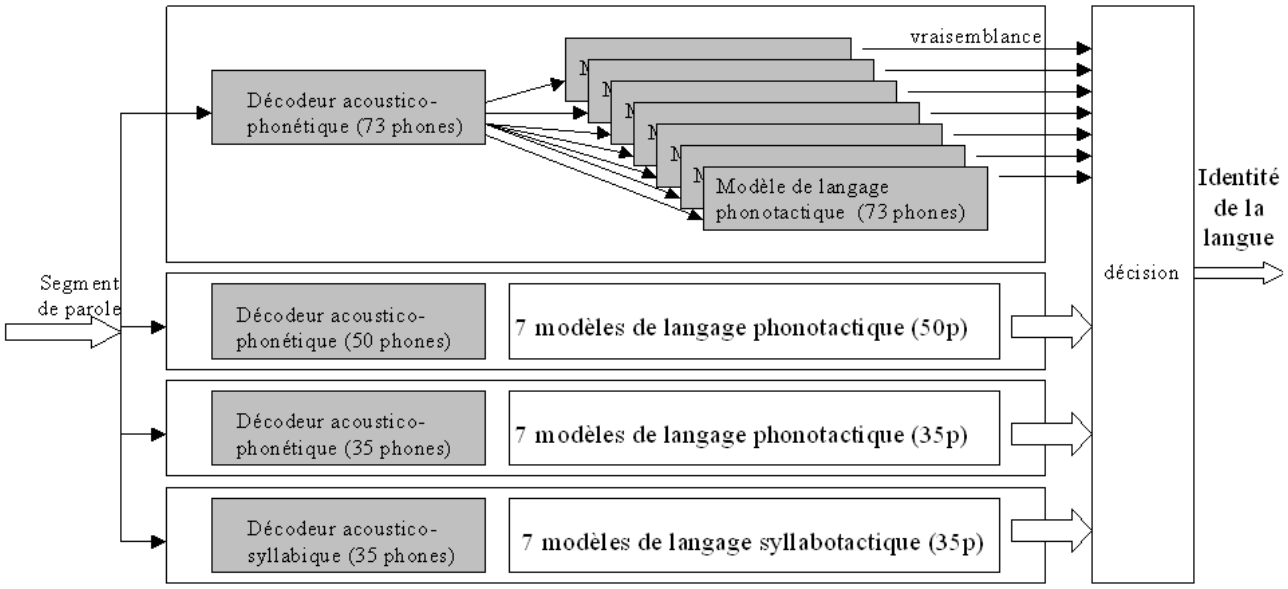

Figure 1 : Le système d'IAL (PPRLM) le plus performant combine l'approche phonotactique et l'approche syllabotactique utilisant les modèles acoustiques multilingues en contexte .