UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1

N^0 attribué par la bibliothèque
/__/__/__/__/__/__/__/__/__/__/

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER - GRENOBLE 1

Discipline : Informatique

présentée et soutenue publiquement

par

LÊ Việt Bắc

le 1er juin 2006

Titre :

Reconnaissance automatique de la parole pour des langues peu dotées

Directeur de thèse :	Jean Caelen
Codirecteurs de thèse :	Laurent Besacier
	Brigitte Bigi

JURY

M. Christian Boitet	Président
M. Renato De Mori	Rapporteur
M. Jean-Paul Haton	Rapporteur
M. Vincent Berment	Examinateur
M. Jean Caelen	Directeur de thèse
M. Laurent Besacier	Codirecteur de thèse
Mlle. Brigite Bigi	Codirectrice de thèse

Thèse préparée au sein du laboratoire de Communication Langagière et Interaction Personne-Système,
Fédération IMAG – Université Joseph Fourier – Grenoble I

Kính tặng ba mẹ tôi!

à mes parents …

Remerciements

Je voudrais tout d'abord remercier Jean CAELEN pour m'avoir accueilli au sein du laboratoire CLIPS/IMAG et pour avoir accepté d'être mon directeur de thèse.

Je tiens à remercier également Laurent BESACIER et Brigitte BIGI pour avoir accepté d'encadrer cette thèse. Un grand remerciement très chaleureux à Laurent BESACIER, qui m'a guidé tout au long de ces années de thèse, pour ses critiques, ses conseils très précis sur mes travaux de recherche et pour avoir relu, corrigé et commenté très soigneusement ce manuscrit. Je voudrais remercier Brigitte BIGI pour son aide dévouée sur mes travaux de thèse, pour ses conseils très utiles et sa relecture de tout mon manuscrit.

J'adresse mes remerciements à Renato DE MORI et Jean-Paul HATON pour avoir accepté d'être rapporteurs de ma thèse. Je voudrais remercier aussi Christian BOITET pour avoir accepté d'être le président du jury. Je remercie Vincent BERMENT pour son aide et sa participation au jury de cette thèse.

Un grand remerciement également à Jean-François SERIGNAT, responsable de l'équipe GEOD pour m'avoir accueilli dans l'équipe GEOD et pour son aide dévouée. Je tiens à remercier Eric CASTELLI, directeur adjoint du Centre MICA (Hanoi, Vietnam) pour m'avoir accueilli au sein des projets de collaboration CORUS et TALK, pour ses suggestions sur mon sujet et pour sa relecture de ce manuscrit. Un grand merci aux collaborateurs de mes travaux de recherche (Do Dat, Sethserey, Luis Villaseñor, …).

J'adresse mes remerciements à Tanja SCHULTZ pour m'avoir accueilli dans le laboratoire Interactive Systems Labs (Carnegie Mellon University) et pour l'intérêt porté à mes travaux de recherche.

Je tiens à remercier également tous les membres de l'équipe GEOD (Solange, Yannick, Tien-Ping, Pedro, Richard, Anas, …) pour leur accueil et leur sympathie. Je remercie Denis TUFFELLI pour sa relecture de ce manuscrit. Un grand remerciement à mes amis vietnamiens au CLIPS et à Grenoble (An Te, Quoc Cuong, Do Dat, Ngoc Hoa, Trung Hung, Bao Quoc, Hoang Nam, …) avec qui j'ai partagé de grands moments au cours de ma thèse.

Enfin, je voudrais exprimer mes plus profonds remerciements à mes parents, à mon petit frère, à toute ma grande famille et à Hà Trang, ma petite amie, pour leurs sentiments, leurs soutiens et leurs encouragements dans tout le temps où j'ai effectué cette thèse.

Un grand merci à tous !

LÊ Việt Bắc

Résumé

Dans la plupart des langues peu dotées, les services liés aux technologies du traitement de l'oral sont inexistants. L'originalité de mon travail de thèse vient de la volonté d'aborder ces langues pour lesquelles peu ou pas de ressources nécessaires pour la reconnaissance automatique de la parole sont disponibles. Ce manuscrit présente notre méthodologie qui vise à développer et adapter rapidement un système de reconnaissance automatique de la parole continue pour une nouvelle langue peu dotée.

La nature statistique des approches nécessite de disposer d'une grande quantité de ressources (vocabulaires, grands corpus de texte, grands corpus de parole, dictionnaires de prononciation) pour le développement d'un système de reconnaissance automatique de la parole continue à grand vocabulaire. Ces ressources ne sont cependant pas disponibles directement pour des langues peu dotées. Par conséquent, une première façon d'accélérer la portabilité des systèmes de reconnaissance vocale vers une nouvelle langue est de développer une méthodologie permettant une collecte rapide et facilitée de ressources écrites et orales. Dans ce travail, nous proposons tout d'abord des solutions pour résoudre les difficultés principales de récupération et de traitement des ressources textuelles spécifiques aux langues peu dotées : recueil d'un vocabulaire, collecte de documents à partir de l'Internet, normalisation de textes, segmentation de textes, filtrage. Une boîte à outils générique « open source » nommée *CLIPS-Text-Tk* a notamment été développée pour faciliter le portage des outils de traitement de corpus textuels vers une nouvelle langue.

Ensuite, la plus grande partie de notre travail de thèse concerne la construction rapide de modèles acoustiques pour une langue peu dotée. Nous proposons des concepts et des méthodes d'estimation de similarités entre unités phonémiques (phonème, polyphone, groupe de polyphones, …). Ces mesures de similarité sont ensuite utilisées pour la portabilité et l'adaptation rapide des modèles acoustiques multilingues indépendant et dépendant du contexte vers une nouvelle langue peu dotée. Pour les langues peu dotées qui ne disposent pas encore de dictionnaire phonétique, une modélisation acoustique à base de graphèmes est aussi proposée et évaluée.

Enfin, les ressources écrites et orales collectées pour le vietnamien et le khmer ainsi que les résultats expérimentaux obtenus par nos systèmes de reconnaissance automatique de la parole en vietnamien et en khmer sont présentés et valident le potentiel des méthodes que nous avons proposées.

Mots-clés: langues peu dotées, reconnaissance automatique de la parole, ressources écrites et orales, similarités entre des unités acoustique-phonémiques, modélisation acoustique crosslingue, adaptation de modèles acoustiques, modélisation acoustique graphémique, modélisation statistique du langage.

Abstract

Nowadays, computers are heavily used to communicate via text and speech. Text processing tools, electronic dictionaries, and even more advanced systems like text-to-speech or dictation are readily available for several languages. There are however more than 6900 languages in the world and only a small number possess the resources required for implementation of Human Language Technologies (HLT). Thus, HLT are mostly concerned by languages for which large resources are available or which have suddenly become of interest because of the economic or political scene. On the contrary, languages from developing countries or minorities have been less worked on in the past years. One way of improving this "language divide" is do more research on portability of HLT for multilingual applications.

Among HLT, we are particularly interested in Automatic Speech Recognition (ASR). Therefore, we are interested in new techniques and tools for rapid development of ASR systems for under-resourced languages or π-languages when only limited resources are available. These languages are typically spoken in developing countries, but can nevertheless have many speakers. In this work, we investigate Vietnamese and Khmer, which are respectively spoken by 67 million and 13 million people, but for which speech processing services do not exist at all.

Firstly, given the statistical nature of the methods used in ASR, a large amount of resources (vocabularies, text corpora, transcribed speech corpora, phonetic dictionaries) is crucial for building an ASR system for a new language. Concerning text resources, a new methodology for fast text corpora acquisition for π-languages is proposed and applied to Vietnamese and Khmer. Some specific problems in text acquisition and text processing for π-languages such as text normalization, text segmentation, text filtering are resolved. For fast developing of text processing tools for a new π-language, an open source generic toolkit named *CLIPS-Text-Tk* was developed during this thesis.

Secondly, for acoustic modeling, we address particularly the use of acoustic-phonetic unit similarities for multilingual acoustic models portability to new languages. Notably, an estimation method of the similarity between two phonemes is first proposed. Based on these phoneme similarities, some estimation methods for polyphone similarity and clustered polyphonic model similarity are investigated. For a new language, a source/target acoustic-phonetic unit mapping table can be constructed with these similarity measures. Then, clustered models in the target language are duplicated from the nearest clustered models in the source language and adapted with limited data to the target language. Results obtained for Vietnamese demonstrate the feasibility and efficiency of these methods. The proposal of grapheme-based acoustic modeling, which avoids building a pronunciation dictionary, is also investigated in our work. Finally, our whole methodology is applied to design a Khmer ASR system which leads to 70% word accuracy and which was developed in only five months.

Keywords: under-resourced languages, LVCSR, speech and language resources acquisition, statistical language modeling, acoustic-phonetic unit similarities, crosslingual acoustic modeling and adaptation, grapheme-based ASR.

Table des matières

Liste des figures

Liste des tableaux

Introduction

De nos jours, les ordinateurs sont largement utilisés pour communiquer par l'intermédiaire du texte et de la parole. Outils de traitement de texte, dictionnaires électroniques, services en lignes, voire des systèmes plus avancés comme la dictée ou la synthèse vocale, sont disponibles pour un petit nombre de langues parmi les 6000 parlées dans le monde. D'après V. Berment, « l'idée s'impose alors qu'aux moyens traditionnels doivent s'ajouter les outils informatiques appropriés sans lesquels les buts visés ne peuvent plus être atteints. L'informatisation d'une langue occupe ainsi une place essentielle dans ce vaste contexte. » [Berment 2004]

Dans le cadre de cette thèse, nous nous concentrons sur les langues qui ont peu de ressources informatiques utilisables pour l'implémentation de technologies en langage naturel**.** Certaines de ces langues sont cependant des langues majoritaires[1] des pays en voie de développement car ces langues peuvent être parlées par un grand nombre de locuteurs. Notre travail porte sur la langue vietnamienne et la langue khmère qui sont notamment parlées par 65,8 millions et 12,1 millions de personnes au Vietnam et au Cambodge, respectivement. Pour ces langues, très peu de ressources électroniques utilisables sont disponibles.

Les activités de recherche décrites dans ce manuscrit sont en phase avec l'évolution du domaine de la reconnaissance automatique de la parole de ces vingt dernières années où on est de plus en plus amené à traiter des documents de nature multilingue. Dans ce domaine qui commence à être abordé au niveau international, il subsiste un certain nombre de verrous, notamment en ce qui concerne la généricité des systèmes de reconnaissance automatique de la parole, et leur portabilité vers de nouvelles langues. Notre objectif est d'étudier et de proposer des solutions et des outils permettant d'accélérer la portabilité des systèmes de reconnaissance automatique de la parole continue à grand vocabulaire vers une nouvelle langue.

L'originalité de notre approche par rapport à l'existant vient de la volonté d'aborder des langues peu dotées, pour lesquelles peu ou pas de corpus sont disponibles, ce qui nécessite des méthodologies innovantes qui vont bien au-delà du simple réapprentissage ou de l'adaptation de modèles. Ce problème spécifique du multilinguisme et des langues peu dotées est notamment mis en avant dans un tutorial d'IBM [Gao 2005] sur les futurs défis en reconnaissance automatique de la parole.

Dans le premier chapitre de ce manuscrit, nous présenterons le contexte d'étude et les concepts de base utilisés dans ce travail. Nous présenterons une méthode d'évaluation du niveau d'informatisation d'une langue, nous permettant de définir ce qu'est une langue peu dotée dans notre contexte d'étude. Ensuite, les principes fondamentaux de la reconnaissance automatique de la parole seront présentés. Nous terminerons ce chapitre par la présentation des activités de recherche récentes sur la reconnaissance automatique de la parole multilingue : les ressources

[1] Par la suite, nous définirons une langue majoritaire comme une langue qui est parlée par la majorité d'une population d'un pays ou d'une région quelconques.

linguistiques multilingues, la modélisation acoustique multilingue et la portabilité de systèmes multilingues vers une nouvelle langue.

Le chapitre 2 présentera notre travail sur le recueil rapide de ressources textuelles dans le contexte de la reconnaissance automatique de la parole : vocabulaire, corpus de texte, modèles statistiques du langage. Les difficultés de récupération et de traitement des ressources textuelles pour des langues peu dotées seront présentées et des solutions seront proposées. Pour construire un corpus de texte, dans un but de généricité, la construction de composants réutilisables pour plusieurs langues et tâches spécifiques sera également présentée. Lors de la mise en pratique de notre méthodologie, une boîte à outils générique de type « open source » de récupération et de traitement d'un corpus de texte nommé *CLIPS-Text-Tk* a été développée.

Le chapitre 3, qui constitue la contribution principale de nos travaux, présentera le recueil de signaux et la construction rapide de modèles acoustiques pour des langues peu dotées. Une première façon d'accélérer la portabilité des systèmes de reconnaissance automatique de la parole continue grand vocabulaire en langue source vers une langue peu dotée est de développer une méthodologie permettant une collecte rapide et/ou facilitée de ressources textuelles et acoustiques, contenant à la fois des signaux de parole de taille limitée et un dictionnaire de prononciation. Ensuite, la plus grosse partie de ce chapitre sera consacrée à la proposition de méthodes de portabilité et d'adaptation rapide des modèles acoustiques indépendants et dépendants du contexte vers une nouvelle langue peu dotée. Nous présenterons également notre travail sur la modélisation acoustique à base de graphèmes pour les langues peu dotées qui ne disposent pas encore de dictionnaire phonétique.

Les premières applications de notre méthodologie au vietnamien et au khmer seront présentées dans les chapitres 4 et 5. Les résultats associés (ressources obtenues et expériences de reconnaissance automatique de la parole) seront également présentés.

Le document se terminera par les conclusions et les perspectives dans le chapitre 6.

Chapitre 1

Contexte d'étude et état de l'art

1. Contexte d'étude

1.1. Informatisation d'une langue

D'après la base de donnée « *Ethnologue : Languages of the World* »[1] [SIL 2005], il y a 6912 langues vivantes parlées dans plus de 200 pays au monde. Elles sont réparties en cinq régions : Amérique, Afrique, Europe, Asie et Pacifique. La distribution par région géographique des langues et des populations est résumée dans le tableau 1.1.

Région	Langues		Locuteurs	
	Nombre	Pourcentage	Nombre	Pourcentage
Afrique	2 092	30,3	675 887 158	11,8
Amérique	1 002	14,5	47 559 381	0,8
Asie	2 269	32,8	3 489 897 147	61,0
Europe	239	3,5	1 504 393 183	26,3
Pacifique	1 310	19,0	6 124 341	0,1
TOTAL	6 912	100,0	5 723 861 210	100,0

Tableau 1.1 : Répartition des langues par régions géographiques [SIL 2005]

Parmi les 6912 langues parlées dans le monde, seul un tout petit nombre d'entre-elles possède les ressources nécessaires pour implémenter des technologies issues du traitement du langage naturel (*Human Language Technologies*, HLT[2]). Ainsi, de telles technologies ont surtout été développées pour des langues qui ont des ressources disponibles en quantité importante ou qui sont soudainement apparues comme intéressantes en raison de la scène économique ou politique.

D'une manière générale, les langues minoritaires ou les langues venant de pays en voie de développement sont moins abordées par la communauté du traitement automatique de la langue naturelle. Pour réduire cette « fracture linguistique » qui existe au niveau de l'informatisation de ces langues, une approche consiste à conduire des recherches sur la portabilité des technologies

[1] http://www.ethnologue.com
[2] http://www.hltcentral.org

en langage naturel, vers des applications multilingues. Ce problème a été de plus en plus étudié au cours des dernières années, par exemple dans le groupe SALTMIL[1], pour lequel, une *langue minoritaire* signifie une langue parlée par une minorité d'une population d'un pays ou d'une région quelconque. Ce terme n'est cependant pas directement lié au niveau d'informatisation car, par exemple, il existe des langues minoritaires qui possèdent déjà des ressources informatiques importantes, comme le catalan et le basque.

Au cours des dernières années, les langues minoritaires et les langues peu dotées ont attiré une attention croissante dans la communauté du Traitement Automatique de la Langue Naturelle (TALN). Des projets qui visent à la revitalisation, la standardisation et à la normalisation linguistique ont été lancés pour favoriser l'usage de ces langues et pour contribuer à leur survie. Les locuteurs de langues minoritaires ont pris conscience du fait que leurs langues appartiennent à l'acquis culturel du monde, et ils sont de plus en plus enclins à utiliser leur langue maternelle à une échelle plus large. L'augmentation du nombre de pages sur l'Internet en langues minoritaires en est une illustration.

Concernant les recherches en traitement automatique de la langue pour des langues peu dotées et des langues minoritaires, on a vu récemment émerger des sessions spéciales ou des ateliers dans des conférences telles que LREC[2] ou TALN[3] :

- atelier LREC 1998 : *Language Resources for European Minority Languages*[4] *;*
- atelier LREC 2000 : *Developing language resources for minority languages: reusability and strategic priorities*[5] *;*
- atelier LREC 2002 : *Portability Issues in Human Language Technologies (HLT)*[6]*;*
- atelier TALN 2003 : *Traitement automatique des langues minoritaires et des petites langues*[7] ;
- atelier LREC 2004 : *1st Steps for Language Documentation of Minority Languages: Computational Linguistic Tools for Morphology, Lexicon & Corpus Compilation*[8] ;
- atelier TALN 2005 : *Traitement des langues peu dotées*[9].

Les langues peu dotées font l'objet de recherches de groupes ou d'organisations internationaux. Fondé en 1934, SIL International[10] est une organisation qui étudie, documente,

[1] Speech And Language Technology for MInority Languages : http://isl.ntf.uni-lj.si/SALTMIL/
[2] International Conference On Language Resources And Evaluation : http://www.lrec-conf.org
[3] Conférence sur le TALN : http://www.atala.org/rubrique.php3? id_rubrique=14
[4] http://www.lrec-conf.org/lrec98/ceres.ugr.es/_rubio/elra/minority.html
[5] http://www.lrec-conf.org/lrec2000/www.cstr.ed.ac.uk/SALTMIL/lrec00.html
[6] http://www.lrec-conf.org/lrec2002/lrec/wksh/Portability.html
[7] http://www.sciences.univ-nantes.fr/info/recherche/taln2003/page/acte_sommaire.html
[8] http://www.lrec-conf.org/lrec2004/doc/ws/prg_minority-languages.pdf
[9] http://taln.limsi.fr/site/talnRecital05/actes-articles.htm
[10] Summer Institute of Linguistics: http://www.sil.org/sil/

et aide à développer des langues moins connues du monde (les langues minoritaires). SIL se met notamment au service des milliers de communautés linguistiques à travers le monde dont les langues ne sont pas écrites. SALTMIL est un groupe d'intérêt spécial au sein de l'association ISCA[1] pour les langues minoritaires. L'objectif de ce groupe est de promouvoir la recherche et le développement dans le traitement des langues minoritaires.

Au niveau des publications scientifiques, il existe quelques recherches sur le traitement du langage naturel et la collecte de ressources pour des langues minoritaires et des langues peu dotées : méthodes pour informatiser des langues peu dotées [Berment 2004], construction automatique de corpus textuels pour langues minoritaires [Scannell 2003, Ghani 2005]. Dans le domaine du traitement automatique de la parole, on ne trouve cependant que quelques travaux ponctuels portant sur des langues asiatiques peu dotées comme par exemple le thaï [Suebvisai, 2005] ou l'indonésien [Martin, 2005].

1.2. Les projets en collaboration

Cette section décrit le contexte de travail de ma thèse, qui a conduit au choix des deux langues abordées.

Ce travail a pu être réalisé grâce à une collaboration entre le laboratoire CLIPS/IMAG[2] (Grenoble, France) et le centre MICA[3] (Hanoï, Vietnam). MICA a été créé en 2002 pour participer au développement des technologies de l'information au Vietnam et pour répondre aux préoccupations relatives à leur évolution. La collaboration avec le laboratoire CLIPS/IMAG porte sur le traitement de la langue vietnamienne depuis sa création, en 2002. Fruit de cette collaboration, le projet CORUS "Traitement de la parole en langue vietnamienne" est soutenu par le MAE français (Ministère des Affaires Etrangères).

Plus récemment, le Département de Génie Informatique et Communication de l'Institut de Technologie du Cambodge - ITC[4] (Phnom Penh, Cambodge) s'est associé à cette collaboration afin d'initier un groupe de recherche spécialisé en traitement de la parole en langue khmère. Le projet TALK "Traitement Automatique de la Langue Khmère" est un projet de collaboration internationale entre le centre MICA, l'institut ITC et le laboratoire CLIPS/IMAG. Ce projet est soutenu par l'AUF (Agence Universitaire pour la Francophonie). L'objectif du projet TALK est de mettre en place au sein de l'équipe de l'ITC, un groupe de recherche sur la parole en langue khmère, pour concevoir des applications d'interaction et de communication parlée et doter les systèmes d'une composante langagière fiable et performante [Talk 2005].

1.3. La langue vietnamienne

Le vietnamien est parlé par environ 65,8 millions de personnes au Vietnam et environ 67,4 millions dans le monde. On trouve notamment des communautés vietnamiennes en Australie, au

[1] International Speech Communication Association : http://www.isca-speech.org/
[2] http://clips.imag.fr
[3] http://www.mica.edu.vn
[4] http://www.itc.edu.kh/fr/

Cambodge, au Canada, en Chine, en France, au Laos, et aux Etats-Unis (*source : Ethnologue[1] 1999*). Son origine est toujours sujette à débat parmi les linguistes. Il est cependant généralement admis que la langue vietnamienne a des racines communes et fortes avec le môn-khmer qui fait partie de la branche austro asiatique[2] et qui comprend le môn parlé en Birmanie et le khmer, la langue cambodgienne, aussi bien que les khmu, bahnar et bru, d'autres langues parlées par les habitants des îles du nord du Vietnam. C'est une langue tonale qui possède six tons. L'orthographe est latine depuis le XVIIè siècle, avec des caractères accentués pour les tons.

1.4. La langue khmère

Le khmer est parlé par environ 12,1 millions de personnes au Cambodge et environ 13,3 millions de personnes dans le monde (*source : Ethnologue[3] 2004*). Il appartient également au groupe des langues môn-khmères. La langue khmère est une langue atonale contrairement aux langues chinoise, thaïe ou vietnamienne. Cependant, le khmer possède comme ses cousins austro-asiatiques plusieurs registres vocaliques : les voyelles peuvent être allongées (dites voyelles longues), raccourcies (dites voyelles brèves), diphtonguées, reposer sur des consonnes aspirées ou non aspirées, ce qui en modifie complètement le sens (par exemple *slap* signifie mourir ; *slaap* signifie aile d'un oiseau). Cette particularité fait du cambodgien un des plus riches systèmes vocaliques au monde. Au niveau de l'écriture, pour adapter les fontes informatiques, il faut gérer un ordonnancement, sur plusieurs niveaux, de 33 consonnes, 32 consonnes souscrites, 21 voyelles dépendantes et 14 voyelles indépendantes, sans compter les consonnes empruntées au thaï et au français, les chiffres, les ligatures, les diacritiques et la ponctuation.

2. Niveau d'informatisation d'une langue

2.1. Mesure du niveau d'informatisation : *l'indice-σ*

Pour évaluer de manière quantitative le degré d'informatisation d'une langue, V. Berment a proposé dans sa thèse [Berment 2004] le protocole suivant : pour chaque service ou ressource, un groupe d'utilisateurs représentatifs des locuteurs de la langue attribue un niveau de criticité C_k et une note N_k. La moyenne pondérée des notes, appelée *indice-σ*, reflète leur satisfaction globale (tableau 1.2). La criticité est une mesure de l'importance relative d'un service pour un groupe d'évaluation donné. Pour une langue, il y a 5 groupes de services ou ressources à évaluer tels que : le traitement de texte, le traitement de l'oral, la traduction automatique, la reconnaissance optique de caractère et les ressources linguistiques.

[1] http://www.ethnologue.com/show_language.asp?code=vie
[2] http://www.ethnologue.com/family_index.asp
[3] http://www.ethnologue.com/show_language.asp?code=khm

	Services / ressources	Criticité (/10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte				
	Saisie simple			
	Visualisation / impression			
	Recherche et remplacement			
	Sélection du texte			
	Tri lexicographique			
	Correction orthographique			
	Correction grammaticale			
	Correction stylistique			
Traitement de l'oral				
	Synthèse vocale			
	Reconnaissance de la parole			
Traduction				
	Traduction automatisée			
ROC				
	Reconnaissance optique de caractères			
Ressources				
	Dictionnaire bilingue			
	Dictionnaire d'usage			
Total		ΣC_k		$\Sigma C_k N_k$
Moyenne (/20)				$\Sigma C_k N_k$ / ΣC_k

Tableau 1.2 : Tableau d'évaluation du niveau d'informatisation d'une langue

À partir de l'indice-σ, V. Berment a défini 3 niveaux d'informatisation pour classer les langues en 3 groupes :

- **langues-π** ont une moyenne entre 0 et 9,99 (langues peu dotées) ;

- **langues-μ** ont une moyenne entre 10 et 13,99 (langues moyennement dotées) ;

- **langues-τ** ont une moyenne entre 14 et 20 (langues très bien dotées).

Comme illustré dans la figure 1.1, une langue-π est ainsi définie comme une langue dont l'indice-σ, n'atteint pas 10/20, et qui est encore insuffisante aux yeux de ses évaluateurs. Une langue-μ est définie comme une langue moyennement dotée qui obtient un indice-σ inférieur à 14/20.

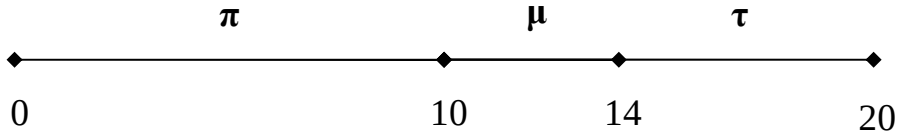


Figure 1.1 : Définition des langues peu, moyennement et très bien dotées

À titre d'exemple, le tableau 1.3 montre une évaluation du niveau d'informatisation présentée dans [Berment 2004] pour la langue khmère. Cette évaluation a été complétée par Michel Antelme, responsable de l'enseignement du khmer à l'INALCO, en fonction de leur connaissance des logiciels et ressources existants en 2004.

	Services / ressources	Criticité (/10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte				
	Saisie simple	10	16	160
	Visualisation / impression	10	14	140
	Recherche et remplacement	8	12	96
	Sélection du texte	6	12	72
	Tri lexicographique	5	0	0
	Correction orthographique	2	0	0
	Correction grammaticale	0	0	0
	Correction stylistique	0	0	0
Traitement de l'oral				
	Synthèse vocale	5	0	0
	Reconnaissance de la parole	5	0	0
Traduction				
	Traduction automatisée	8	4	32
ROC				
	Reconnaissance optique de caractères	9	0	0
Ressources				
	Dictionnaire bilingue	10	4	40
	Dictionnaire d'usage	10	0	0
Total		88		540
Moyenne				**6,14 / 20**

Tableau 1.3 : Tableau d'évaluation du niveau d'informatisation pour le khmer

De la même manière, nous avons demandé à E. Castelli et Q-C. Nguyen, tous deux chercheurs du centre MICA[1] à Hanoï (Vietnam), qui sont responsables des projets de traitement automatique de la langue vietnamienne, de compléter le tableau d'évaluation du niveau d'informatisation (tableau 1.4) pour le vietnamien, en fonction de leur connaissance des logiciels et ressources existants en 2005.

Comme nous pouvons le voir sur les tableaux 1.3 et 1.4, les services liés au traitement de l'oral sont inexistants pour la langue khmère (synthèse vocale et reconnaissance automatique de la parole) et pour la langue vietnamienne. C'est aussi le cas pour une majorité de langues dans le monde, dont certaines sont parlées par plusieurs dizaines de millions de locuteurs, comme le bengali (189 millions), le tamoul (63 millions), y compris au sein de l'Europe des 25 (lituanien, letton, polonais, etc.)[2].

[1] http://www.mica.edu.vn
[2] source : http://www.populationdata.net

	Services / ressources	Criticité (/10)	Note (/20)	Note pondérée (Criticité x Note)
Traitement du texte				
	Saisie simple	10	16	160
	Visualisation / impression	10	16	160
	Recherche et remplacement	8	17	136
	Sélection du texte	6	17	102
	Tri lexicographique	5	6	30
	Correction orthographique	2	6	12
	Correction grammaticale	0	0	0
	Correction stylistique	0	0	0
Traitement de l'oral				
	Synthèse vocale	5	0	0
	Reconnaissance de la parole	5	0	0
Traduction				
	Traduction automatisée	8	6	48
ROC				
	Reconnaissance optique de caractères	9	12	108
Ressources				
	Dictionnaire bilingue	10	13	130
	Dictionnaire d'usage	10	0	0
Total		88		886
Moyenne				**10 / 20**

Tableau 1.4 : Tableau d'évaluation du niveau d'informatisation pour le vietnamien

2.2. Catégoriser des langues dans notre contexte d'étude

Le calcul de l'indice σ nous permet de vérifier que les deux langues de notre étude (le vietnamien et le khmer) font effectivement partie de la classe des langues peu dotées (indice $\sigma < 10$; le vietnamien étant cependant à la limite). Le khmer, évalué dans [Berment, 2004] obtient un indice σ d'environ 6/20, tandis que notre évaluation du vietnamien donne un indice σ de 10/20 environ. Une partie de ces différences s'explique notamment par le fait que le vietnamien, contrairement au khmer, utilise une écriture latine accentuée, rendant plus « accessibles » les tâches de reconnaissance automatique de caractères et de tri par exemple.

Face à la critique qui pourrait être faite concernant notre choix d'aborder des technologies (la reconnaissance vocale) peut être moins importantes, en terme de développement, que d'autres liées au traitement de textes et aux dictionnaires, nous argumenterons que développer des systèmes de traitement de la parole dans une langue « peu dotée » peut permettre de collecter des ressources utiles pouvant être ensuite remises au « pot commun » d'une autre langue donnée (dictionnaire phonétique, corpus oral, transcriptions de conversations spontanées par exemple).

Dans le but de notre recherche, nous décidons de regrouper tout d'abord les langues selon le critère de la disponibilité de ressources linguistiques. Cependant, nous ne considérons désormais que les ressources qui sont nécessaires pour la construction d'un système de reconnaissance

automatique de la parole, soit :

- corpus textuels ;
- corpus de parole ;
- lexiques ;
- dictionnaires de prononciation.

Une **langue bien dotée** est ainsi définie comme une langue qui possède des ressources disponibles pour la reconnaissance automatique de la parole, notamment, la reconnaissance automatique de la parole continue grand vocabulaire. Les langues appartenant à cette catégorie sont souvent des langues très bien dotées informatiquement (langues-τ) ou des langues moyennement dotées informatiquement (langues-μ) qui atteignent un indice-σ supérieur ou égal à 10/20. Les langues bien dotées sont souvent des langues parlées par un nombre important de locuteurs (ou langues majoritaires) : anglais, français, espagnol, chinois, arabe, allemand, italien, russe, portugais, coréen, japonais, …

Une **langue peu dotée** est définie comme une langue qui ne possède pas encore ou pas beaucoup (en quantité et en qualité) de ressources linguistiques pour la construction d'un système de reconnaissance automatique de la parole, particulièrement dans un contexte d'apprentissage statistique où les données doivent être disponibles en grande quantité. Les langues appartenant à cette catégorie sont fréquemment des langues peu dotées informatiquement (langes-π) dont l'indice-σ, n'atteint pas 10/20. Les langues peu dotées sont soit des langues minoritaires, parlées par une minorité de la population[1], comme le basque, le breton, le castillan, le catalan, le galicien, l'occitan, etc. en Europe par exemple[2], mais aussi des langues parlées par plusieurs dizaines de millions de locuteurs, comme[2] le vietnamien (67 millions), le khmer (13 millions), le bengali (189 millions), le tamoul (63 millions), etc.

3. Reconnaissance automatique de la parole

Cette partie est dédiée à la présentation des principes de la reconnaissance automatique de la parole. Le schéma de base étant une chaîne de traitements, nous présenterons donc dans un premier temps le traitement dans sa globalité avant de détailler chaque sous tâche spécifique. Pour chacune d'elles, nous décrirons quelques approches possibles en insistant sur les méthodes et algorithmes couramment utilisés de nos jours.

3.1. De l'onde au langage, schéma général

Le principe général a beaucoup évolué car, en passant d'une reconnaissance fondée sur l'exemple à une reconnaissance fondée sur le modèle, la suite de traitements s'est allongée. De

[1] http://www.nationmaster.com/encyclopedia/minority-language
[2] http://www.ethnologue.com/country_index.asp

façon générale, les systèmes de reconnaissance automatique de la parole actuels, à base de modélisation statistique, suivent le schéma représenté par la figure 1.2 [Dutoit 2002].

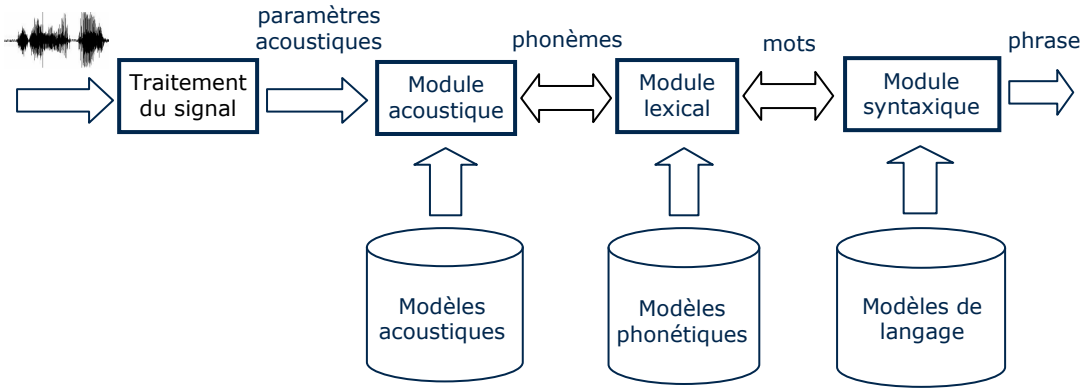


Figure 1.2 : Reconnaissance automatique de la parole par modélisation statistique

À partir d'un signal de parole, le premier traitement consiste à extraire les paramètres caractéristiques. Ces paramètres sont mis en entrée d'un module acoustique, ou un décodage acoustico-phonétique. Ce décodage acoustico-phonétique peut produire à son tour une ou plusieurs hypothèses phonétiques associées en général à une probabilité pour chaque segment (une fenêtre ou une trame) de signaux de la parole. Ce générateur d'hypothèses locales est souvent modélisé par des modèles statistiques d'unités élémentaires de parole, par exemple un phonème. Pour entraîner des modèles acoustiques, nous apprenons les modèles des unités acoustiques d'un corpus étiqueté.

Le générateur d'hypothèses interagit avec un module lexical pour forcer le décodage acoustico-phonétique à ne reconnaître que des mots représentés dans le module lexical. Les modèles phonétiques sont représentés par un dictionnaire de prononciation (dictionnaire phonétique) ou par des automates probabilistes qui sont capables d'associer une probabilité à chaque prononciation possible d'un mot.

Pour la reconnaissance automatique de la parole continue grand vocabulaire, le générateur interagit avec un module syntaxique pour forcer le reconnaisseur à intégrer des contraintes syntaxiques, voire sémantiques. Ces contraintes sont souvent formalisées par des modèles de langage. Pour reconnaître ce qui est dit, on commence par chercher, grâce aux modèles d'unités acoustiques, l'unité qui est supposée avoir été produite, puis construisons, à partir du treillis d'unités acoustiques et d'un modèle statistique du langage, la suite de mots la plus probable.

Avant de présenter ces modules séparément, nous donnons l'équation bayésienne appliquée au problème de la reconnaissance automatique de la parole. Soient x une séquence de vecteurs acoustiques inconnus et w_i (i=1..K) une de K classes possibles pour cette observation (ex. phonèmes, mots, etc.). La classe reconnue est :

$$w^* = \arg\max_i \frac{p(x/w_i).P(w_i)}{p(x)} = \arg\max_i p(x/w_i).P(w_i) \qquad (1.1)$$

Le mot reconnu w^* sera donc celui qui maximise cette quantité, parmi tous les mots candidats w_i. Ce que nous appelons signal est donc la séquence de vecteurs acoustiques (et non le signal brut). Ainsi, nous trouvons la séquence la plus probable pour ce signal. Le prétraitement rend le signal x exploitable.

Les probabilités $P(x \mid w_i)$ d'observer le signal x sachant la séquence w_i nécessitent un modèle acoustique pour être estimées. Les probabilités $P(w_i)$ *a priori* de la séquence, indépendamment du signal, nécessitent un modèle de langage pour être estimées.

$P(x)$ est la probabilité du signal. Elle est la même pour toutes les séquences possibles donc sa valeur n'est pas prise en compte.

Nous présentons maintenant les principes généraux des différents modules constituant un système de reconnaissance automatique de la parole : décodage acoustico-phonetique, dictionnaire phonétique, modélisation statistique du langage.

3.2. Analyse acoustique et paramétrisation du signal

Au niveau acoustique, la parole apparaît physiquement comme une variation de la pression de l'air causée et émise par le système articulatoire [Boite 2000, Calliope 1989]. Ce signal continu est converti numériquement en intervalles de temps discret (phase d'échantillonnage), puis quantifié en valeurs discrètes d'amplitude (phase de quantification ou conversion analogique / numérique). La figure 1.3 représente le signal de parole en français « *allô ! le service réservation ?* » après la phase d'échantillonnage et de quantification.

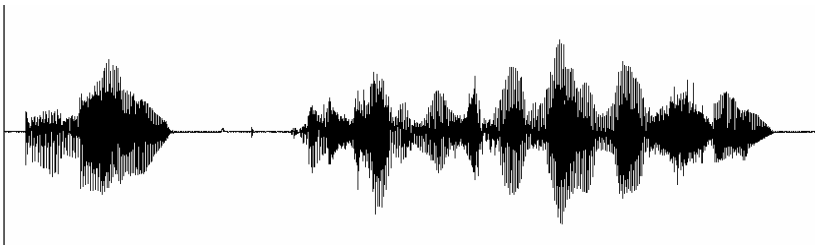


Figure 1.3 : Signal temporel de la phrase « Allô, le service réservation ? »

Pour la modélisation acoustique et pour le décodage acoustico-phonétique, le signal numérique n'est pas exploitable directement car la quantité de données est trop importante et beaucoup de ces données sont inutiles (telles que le bruit de fond) ou redondantes.

Pour réduire cette quantité de données et supprimer les informations inutiles, plusieurs algorithmes sont nécessaires. Le signal de parole est fortement non stationnaire [Boite 2000]. Généralement, une analyse spectrale, dont l'outil est la transformée de Fourier discrète, est effectuée environ toutes les 10ms sur une fenêtre (ou trame) de 10 à 20ms de signal vocal pondérées par une fenêtre de pondération (Hamming, Hanning, …). Dans la fenêtre de 10 à 20ms, on suppose que le signal vocal est suffisamment stable. On en extrait des coefficients, dits coefficients spectraux. Ensuite, d'autres coefficients (Mel-cepstrum par exemple [Davis 1980]) peuvent être dérivés de ces coefficients spectraux.

Suite à ces prétraitements, le signal est généralement transformé en une série de vecteurs comportant typiquement de 8 à 50 coefficients. Plusieurs types de coefficients, donc de représentations, sont couramment utilisés selon les systèmes et les besoins : ZCR (Taux de passage par zéro), Energie, MFCC (*Mel Filter Cepstral Coefficients*), LPC (*Linear Prédictive Coding*), PLP (*Perceptual Linear Predictive*), …Une présentation détaillée de ces coefficients se trouve dans [Rabiner 1978, Davis 1980, Boite 2000].

D'une façon générale, l'ensemble des paramètres extraits du signal fournit un vecteur de grande dimension. Il est alors courant de réduire la taille de ce vecteur en gardant le maximum d'informations utiles. Par exemple, proche de l'analyse en composantes principales, la LDA (*Linear Discriminant Analysis*) [Haeb-Umbach 1992, Balakrishnama 1999] est une méthode d'extraction de paramètres qui fournit une transformation linéaire de vecteurs-paramètres de dimension n dans un espace de dimension $m < n$ de telle sorte que les individus appartenant à une même classe soient proches les uns des autres, et que ceux appartenant à des classes différentes soient éloignés.

3.3. Décodage acoustico-phonétique à base de modélisation acoustique

D'après [Haton 1991], un décodage acoustico-phonétique (DAP) est défini généralement comme la transformation de l'onde vocale, en unités phonétiques - une sorte de transcodage qui fait passer d'un code acoustique à un code phonétique - ou plus exactement comme la mise en correspondance du signal et d'unités phonétiques prédéfinies (opération de couplage / identification) dans lequel le niveau de représentation passe du continu au discret.

Ce module est composé d'une première partie consistant à extraire les paramètres choisis pour représenter le signal, et d'une seconde partie qui, à partir de ces jeux de paramètres, apprend des modèles d'unités acoustiques ou décode le signal d'entrée, selon que l'on veuille apprendre ou reconnaître.

3.3.1. Modélisation acoustique

Les approches statistiques et les modèles probabilistes sont très utilisés, de nos jours, dans les systèmes de reconnaissance automatique de la parole. Ces approches, notamment celles basés sur les Modèles de Markov Cachés (HMM), ont atteint des performances remarquables avec des vocabulaires de plus en plus importants et une robustesse au bruit et à la variabilité des locuteurs de plus en plus grande [Rabiner 1993]

Dans les années 70, l'approche consistait en un paradigme de reconnaissance de mots « par l'exemple ». Ces premiers systèmes fonctionnaient à base de patrons de vecteurs acoustiques ou « *template-based systems* » en anglais [Huang 2001]. Le principe consistait à faire répéter plusieurs exemples des mots à reconnaître et à les analyser sous forme de vecteurs acoustiques dans un patron. Ensuite, pour reconnaître un mot inconnu, il « suffisait » de comparer le jeu de vecteurs acoustiques extraits du signal avec les suites d'exemples appris (ou enregistrés) précédemment. Ce principe de base n'est cependant pas implémentable directement parce qu'un même mot peut être prononcé de nombreuses de façons différentes, en changeant le rythme de

l'élocution. La superposition du signal inconnu aux signaux de base doit dès lors se faire en acceptant une certaine « élasticité » temporelle, formalisée mathématiquement par l'algorithme *Dynamic Time Warping* (DTW) [Silverman 1990].

Cette approche pionnière s'est rapidement confrontée aux grands problèmes de la reconnaissance automatique de la parole. Comment faire face à la variabilité due aux locuteurs et au contexte d'enregistrement, comment élaborer une construction sémantique et non simplement lexicale et donc comment gérer de très grands dictionnaires ?

De tous ces obstacles sont apparus des unités acoustiques plus petites (en termes de temps) et les modèles probabilistes. Les unités acoustiques ne sont plus des mots mais des phonèmes. Le principe consiste alors à déduire des modèles de phonèmes plutôt que des exemples de mots. Ainsi, ces modèles sont beaucoup plus souples, dans le sens où ils couvrent beaucoup plus de variations et permettent la gestion de gros vocabulaires sans modifier le nombre d'unités acoustiques représentées. Les modèles peuvent être applicables pour n'importe quelle voix. Il est même possible de découper encore ces petites unités acoustiques au sein du modèle lui-même. Enfin, ces unités acoustiques peuvent être non plus des phonèmes, mais des combinaisons de phonèmes, c'est-à-dire un phonème en fonction de son contexte. Par exemple, on modélisera le phonème [ɛ] suivi du phonème [ʤ], ou le phonème [ɛ] suivi du phonème [ʧ], etc. au lieu du phonème [ɛ]. Nous parlons alors de polyphones : diphones, triphones ou même quintphones selon le nombre de contextes pris en compte.

Pour la modélisation statistique acoustique, les modèles de Markov cachés (HMM) sont aujourd'hui utilisés dans un très grand nombre des systèmes de reconnaissance automatique de la parole. Chaque unité de parole est modélisée par un HMM. Dans le cas de petits lexiques, ces unités de parole peuvent être les mots. Dans le cas de grands lexiques, on préférera souvent utiliser des modèles de phonèmes (ou polyphones), ce qui limitera le nombre de paramètres à estimer. Dans ce dernier cas, lors de la reconnaissance, les mots seront construits (dynamiquement) en termes de séquences de phonèmes et les phrases en termes de séquences de mots.

Les HMMs supposent que le phénomène modélisé est un processus aléatoire et inobservable qui se manifeste par des émissions elles-mêmes aléatoires. Cette approche markovienne offre une flexibilité séduisante de modélisation pour un phénomène aussi complexe que la parole.

3.3.2. Modèles de Markov Cachés

Les modèles de Markov cachés sont apparus dans la problématique de la reconnaissance automatique de la parole dans les années 70 [Baker 1975, Jelinek 1976]. L'idée sous-jacente est que la parole peut être caractérisée par un processus aléatoire dont les paramètres peuvent être estimés d'une manière appropriée. Les modèles HMM ont prouvé leur efficacité dans de nombreux domaines de la reconnaissance automatique de la parole. Cependant, les modèles ont été améliorés au fil des recherches pour repousser leurs limites intrinsèques, particulièrement en intégrant des notions de corrélation entre trames et de modélisation de trajectoires.

Un modèle de Markov discret est un automate stochastique à nombre d'états fini N [Rabiner 1993]. Un processus aléatoire se déplace d'état en état à chaque instant et on note q_t l'état atteint par le processus à l'instant t.

Dans le cas des modèles de Markov discrets cachés, l'état réel q_t n'est pas directement observable mais le processus émet un symbole discret après chaque changement d'état. Les observations ne sont plus univoquement liées à une seule classe bien déterminée mais sont donc des fonctions statistiques de ces classes qui ne sont plus observées directement. Ou encore, les états du modèle ne sont plus observés directement à partir des observations qui sont supposées être produites par ces états mais à travers une fonction statistique différente pour chaque classe. On a donc un processus doublement stochastique : modèle stochastique relatif au modèle de Markov sous-jacent et celui décrivant la relation entre les classes (états) et les observations. La figure 1.4 présente un modèle de Markov caché ergodique, où toutes les transitions entre tous les états sont autorisées.

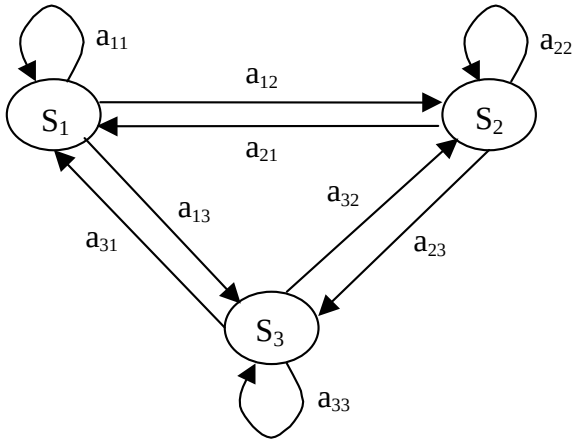


Figure 1.4 : Exemple de modèle de Markov caché ergodique

Concrètement, un modèle de Markov caché HMM est représenté par $\Phi = (A, B, \pi)$ qui est caractérisé par les éléments suivants :

- $S = \{s_1, s_2, \ldots, s_N\}$: un ensemble des états du modèle avec N le nombre d'états. On note q_t l'état à l'instant t ;
- $O = \{o_1, o_2, \ldots, o_M\}$: un alphabet des observations avec M nombre fini de symboles d'observation par état. Les symboles d'observation correspondent à chaque sortie physique du système réel qu'on modélise. On note x_t l'observation à l'instant t ;
- $A = \{a_{ij}\}$: une matrice des probabilités de transition entre états, dont a_{ij} est la probabilité de transition de l'état i à l'état j. On a :

$$a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i),\ 1 \leq i, j \leq N \tag{1.2}$$

$$\sum_{j=1}^{N} a_{ij} = 1, 1 \le i \le N \tag{1.3}$$

- $B = \{b_i(k)\}$: une matrice des probabilités d'émission des observations dans chaque état, dont $b_i(k)$ est la probabilité d'émission de l'observation o_k dans l'état s_i. On a :

$$b_i(k) = P(x_t = o_k \mid q_t = s_i),\ 1 \le i, j \le N \tag{1.4}$$

- $\pi = \{\pi_i\}$: une matrice de distribution de l'état initial. On a :

$$\pi_i = P(q_0 = s_i),\ 1 \le i \le N \tag{1.5}$$

$$\sum_{i=1}^{N} \pi_i = 1 \tag{1.6}$$

Pour la modélisation acoustique, le modèle souvent utilisé est donc le modèle HMM gauche-droit (ou de Bakis), illustré par la figure 1.5, dans lequel on ne peut pas revenir à un état précédent.

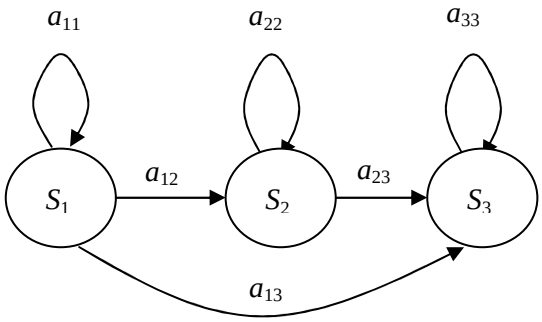


Figure 1.5 : Exemple de HMM à 3 états gauche-droit

Soit un modèle de Markov caché, il existe trois problèmes fondamentaux à résoudre :

- *Évaluation* : Soient un modèle Φ et une séquence d'observations $X = \{x_1, x_2, \ldots, x_T\}$. Comment calculer $P(X \mid \Phi)$, la probabilité que la séquence des observations ait été émise par le modèle Φ ?

- *Décodage* : Soient un modèle Φ et une séquence d'observations $X = \{x_1, x_2, \ldots, x_T\}$. Comment déterminer la séquence d'états cachés $Q = \{q_0, q_1, \ldots, q_T\}$ qui a la plus forte probabilité d'avoir généré la séquence des observations ?

- *Apprentissage :* Soient un modèle Φ et un ensemble d'observations. Comment ajuster les paramètres du modèle Φ pour maximiser la probabilité $P(X \mid \Phi)$?

Le problème de l'*évaluation* est résolu par l'algorithme *Forward.* Le problème de *décodage* peut être résolu en utilisant l'algorithme de *Viterbi*. Enfin, le problème d'apprentissage du modèle peut être résolu par l'algorithme *Baum-Welch* (ou *Forward-Backward*). Le lecteur trouvera de plus amples informations sur les modèles de Markov cachés et ces algorithmes dans [Rabiner 1993].

3.3.3. Modélisation acoustique indépendante ou dépendante du contexte

Pour la modélisation acoustique à base des modèles de Markov cachés, les séquences de mots sont divisées en unités de base, fréquemment les phonèmes. Cependant, ces unités ne sont pas utilisées directement en modélisation dépendant du contexte. Par exemple, un [a] précédé d'un [ɲ] n'est pas identique à un [a] précédé d'un [m] par exemple. C'est ce que l'on appelle le phénomène de coarticulation. D'où l'utilisation de polyphones, où chaque phonème est caractérisé par son contexte précédent et suivant. Par exemple, pour le phonème [u], on peut avoir les diphones, triphones (un contexte gauche et un droit) et quintphones suivants (deux contextes gauches et deux droits) :

monophone :	u
diphones :	u(b|), u(ɲ|), u(|t), ...
triphones :	u(b|n), u(ɲ|t), u(sil|a),…
quintphones :	u(sil,b|n,o), u(a,ɲ|t,e), u(a,m|c, sil),…

A titre d'exemple, pour le système de reconnaissance du français au CLIPS, les unités principalement utilisées sont les triphones car elles offrent une bonne catégorisation entre les « sons ». Les monophones catégorisent trop et les unités *n*-phones pour *n*>3 peuvent s'avérer trop coûteuses si le corpus d'apprentissage est limité [Vaufreydaz 2002].

Pour l'utilisation de monophones, on parle de modélisation acoustique indépendante du contexte[1]. Pour l'utilisation de polyphones, on parle de modélisation acoustique dépendante du contexte[2], dans le sens où la prononciation d'un phonème est caractérisée par ses phonèmes précédents et suivants. Bien entendu, cela engendre une très importante quantité d'unités.

Dans les conditions réelles où le nombre de représentants de polyphones dans le corpus d'apprentissage est insuffisant, nous ne pouvons pas modéliser tous les contextes possibles des phonèmes. En conséquence, nous devons regrouper les polyphones similaires (en même contexte droit par exemple) dans un ensemble de polyphones en utilisant une procédure nommée *clustering* sur un arbre de décision [Huang 2001]. Cet arbre de décision, illustré par la figure 1.6, est construit en appliquant une question de contexte du type par exemple : Est-ce que le contexte gauche est une VOYELLE ? Est-ce que le contexte droit est le phonème [ɲ] ? etc.

[1] En anglais : Context Independent acoustic modeling - CI
[2] En anglais : Context Dependent acoustique modeling - CD

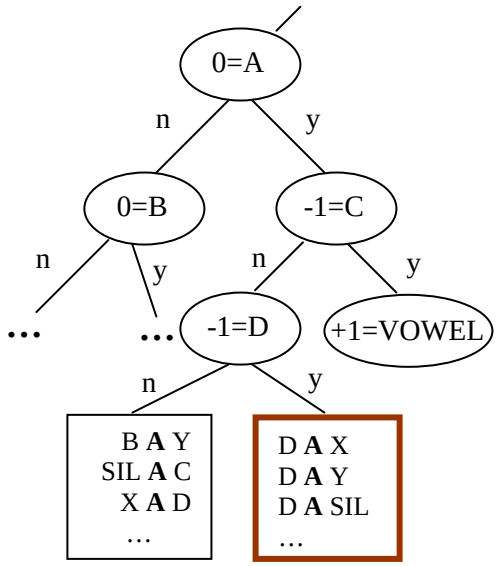


Figure 1.6 : Exemple d'arbre de décision

3.4. Modélisation de la prononciation

Le lien entre le module acoustique et le module syntaxique est fait par le module lexical qui force le décodage acoustico-phonétique à ne reconnaître que des mots qui sont présents dans les modèles phonétiques. Les modèles phonétiques sont souvent représentés par un dictionnaire de prononciation (dictionnaire phonétique). Celui-ci est simplement composé, pour chaque entrée lexicale, d'une description phonétique en terme d'unités acoustiques. Par exemple :

comme	[k ɔ m]
regardé	[ʁ ə g a ʁ d e]
voyageur	[v w a j a ʒ œ ʁ]

Dans ce dictionnaire, nous pouvons intégrer des variantes phonétiques pour assouplir la prononciation et tenir compte des variations d'élocution. Par exemple :

finalement(1)	[f i n a l m ɑ̃]
finalement(2)	[f i n a l ø m ɑ̃]

Nous pouvons, par le même principe, ajouter les liaisons entre mots, comme nous pouvons le constater dans l'exemple ci-dessous, grâce auquel nous modélisons autant « ils ont » que « ils iront ».

ils_ont	[i l z ɔ̃]
ils_iront	[i l z i r ɔ̃]

En plus, dans le dictionnaire de prononciation, il est parfois intéressant de modéliser des suites de mots courts ou longs qui sont considérés comme des mots composés[1] dans le vocabulaire et dans le dictionnaire de prononciation :

[1] En linguistique, un mot composé est un ensemble de mots formant une unité syntaxique et sémantique et il est considéré comme un mot unique.

il y a	[i l i j a]
pomme de terre	[p ɔ m d ə t ɛ ʁ]
au fur et à mesure	[o f y ʁ e a m ə z y ʁ]

3.5. Modélisation statistique du langage

Les systèmes de contrôle vocal, dont la taille du vocabulaire est petite, utilisent souvent la reconnaissance des mots isolés, ou des formes phonétiques simples. Dans ce cas, les systèmes de reconnaissance automatique de la parole dépendent peu du modèle de langage. Un décodeur acoustique-phonétique peut alors atteindre un taux de reconnaissance très élevé (> 95% dans [Rabiner 1993]).

Un système de reconnaissance automatique de la parole continu grand vocabulaire dépend généralement fortement de la connaissance linguistique de la parole. Les meilleurs systèmes de décodage acoustico-phonétique qui n'utilisent aucun modèle de langage n'atteignent qu'un taux d'exactitude en phonèmes de l'ordre de 50% environ. La modélisation du langage est donc une réelle nécessité pour la reconnaissance automatique de la parole continue grand vocabulaire. Un module linguistique est nécessaire dans le système pour déterminer la forme lexicale correspondante, c'est-à-dire la séquence de mots la plus probable, au sens langagier.

Dans l'équation bayésienne appliquée à la reconnaissance automatique de la parole (équation I.1 précédente), une probabilité *a priori* de la séquence est utilisée. Celle-ci se calcule à partir d'une modélisation du langage. Ainsi, la suite *« je suis ici »* est plus probable, en terme de langage, que *« jeu suis ici »*, ou encore « *jeux suit y si* », bien que l'acoustique soit quasi-similaire. Pour une même suite de phonèmes, il peut exister plusieurs centaines de phrases possibles. Le rôle principal du modèle de langage est de les classer selon leur plausibilité linguistique.

La modélisation statistique du langage est par conséquent un module important dans un système de reconnaissance automatique de la parole à grand vocabulaire. Différents types de modèles de langage ont été proposés dans le but de restreindre les séquences de mots envisagées au cours de la reconnaissance. Les modèles statistiques du langage décrivant statistiquement les contraintes sur l'ordre des mots, sont traditionnellement employés. Les modèles que nous présentons ici sont les modèles statistiques les plus utilisés dans le domaine de la reconnaissance automatique de la parole. Une méthode d'évaluation des modèles de langage sera aussi présentée dans cette section.

3.5.1. Les modèles n-grammes

Un modèle statistique du langage consiste, pour une séquence de mots $W = w_1, w_2, \ldots, w_N$, à calculer la probabilité $P(W)$:

$$P(W) = \prod_{i=1}^{N} P(w_i \mid w_1,\ldots,w_{i-1}) = \prod_{i=1}^{N} P(w_i \mid h_i) \qquad (1.7)$$

où $h_i = w_1, \ldots, w_{i-1}$ est considéré comme l'histoire du mot w_i et $P(w_i \mid h_i)$ est la probabilité du mot w_i, sachant tous les mots précédents.

En pratique, au fur et à mesure que la séquence de mots h_i s'enrichit, une estimation des valeurs des probabilités conditionnelles $P(w_i \mid h_i)$ devient de plus en plus difficile car aucun corpus de texte d'apprentissage ne peut permettre d'observer toutes les combinaisons possibles de $h_i = w_1, \ldots, w_{i-1}$.

Afin de réduire la complexité du modèle de langage, et par conséquent de son apprentissage, l'approche *n*-grammes peut être utilisée. Le principe est donc le même, seul l'historique est limité aux *n*-1 mots précédents. La probabilité $P(W)$ est donc approximée par :

$$P(W) = \prod_{i=1}^{N} P(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) \tag{1.8}$$

En pratique, en fonction de la taille du corpus d'apprentissage, on peut choisir des tailles d'historique différentes. On parle alors de modèle unigramme si $n = 1$ (sans historique), bigramme si $n = 2$ ou trigramme si $n = 3$.

L'estimation de probabilité du mot w_i sachant son histoire réduite consiste à compter le nombre d'occurrences $C(w_i)$ des *n*-grammes dans un corpus d'apprentissage. Cette estimation est évaluée selon un vocabulaire de référence *V*. On a :

$$P(w_i \mid w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_{i-1}, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1})} \tag{1.9}$$

Cette estimation devient difficile lorsque la taille du vocabulaire *V* croît car même un modèle *n*-gramme ($n>3$) requiert une très grande quantité de données textuelles d'apprentissage pour estimer efficacement le modèle. Le nombre maximum de trigrammes pour un vocabulaire *V* est $|V|^3$. Ce problème est par exemple important dans notre contexte de travail sur les langues peu dotées car un grand corpus de textes est souvent difficile à acquérir.

Pour résoudre le problème du manque de données d'apprentissage, quelques techniques peuvent être appliquées telles que : les modèles *n*-classes, les modèles cache [Kuhn 1990], la méthode de Good-Turing pour le lissage avec le repli de Katz [Katz 1987].

3.5.2. Apprentissage de modèles statistiques du langage

Générer un modèle statistique consiste à apprendre, à partir d'un corpus, les probabilités de tous les unigrammes, bigrammes, etc existants. Cependant, la modélisation statistique du langage nécessite de disposer de corpus textuels de qualité et de taille conséquentes pour calculer des probabilités correctes. Ceci est le principal problème de ces approches statistiques, source de nombreuses recherches.

Plusieurs recherches se concentrent sur la construction de corpus de textes en grande quantité en collectant des pages Web. D. Vaufreydaz dans le cadre de sa thèse a proposé une

approche exploitant les ressources du Web pour résoudre ce problème [Vaufreydaz 2002]. Un robot parcours des pages web et collecte les informations exploitables. Cette collecte consiste en un jeu complexe d'heuristiques et de filtres. Ces travaux permettent de répondre aux besoins de très grand corpus, et ce de façon quasi-automatique.

Dans le domaine de la recherche d'information, il existe des approches à base de requêtes lancées sur un moteur de recherche pour collecter des documents pertinents dans une langue considérée [Nishimura 2001, Scannell 2003, Monroe 2002].

Enfin, l'approche décrite dans [Zhu 2001] consiste non pas à collecter des documents, comme nous le faisons traditionnellement, mais à essayer d'interroger des moteurs de recherche sur le Web avec des unigrammes, bigrammes et trigrammes consécutivement. Ensuite, le nombre de pages contenant ces *n*-grammes, répondu par le moteur, est utilisé dans le calcul des probabilités associées à ces *n*-grammes. Cette approche pionnière montre un avantage important : elle peut mieux couvrir des *n*-grammes que les méthodes traditionnelles à base de corpus. Mais l'approche rencontre immédiatement des difficultés car l'occurrence de chaque *n*-grammes (1-, 2-, 3-grammes) est générée directement par une requête d'interrogation des moteurs de recherche. Avec plusieurs milliards de *n*-grammes pour une langue, nous ne pouvons pas calculer en temps réel des modèles *n*-grammes. Par conséquent, c'est une méthode théorique qui montre l'intérêt potentiel des ressources de l'Internet.

3.5.3. Évaluation des modèles statistiques du langage par la perplexité

En général, la performance des modèles de langage est évaluée par le taux de reconnaissance de mots du système de reconnaissance automatique de la parole. Néanmoins, les modèles de langage peuvent être évalués séparément par une mesure de qualité propre. La mesure utilisée le plus couramment est la perplexité [Bahl 1977]. Cette mesure se calcule sur un texte non vu au cours de l'apprentissage (*test-set perplexity*). Une autre mesure s'appuie sur le jeu de Shannon [Shannon 1951].

Dans la théorie de l'information, une langue L est considérée comme une source qui émet une suite de mots w_i à partir d'un ensemble fini d'éléments V (vocabulaire) [Abramson 1963]. A chaque émission d'un nouvel élément w_i, la source L apporte une quantité d'information. La quantité d'information contenue dans une suite de mots $W = w_1, w_2, \ldots w_n$ émise par une langue, vaut $- \log P(w_1, w_2, \ldots w_n)$ où P est la probabilité de la séquence de mots.

La valeur moyenne de la quantité d'information d'une source L est souvent appelée entropie est notée $H(W)$. Si la génération des mots est indépendante de leur histoire, l'entropie d'une source qui émet des symboles indépendants est calculée comme suit :

$$H(W) = -\sum_{w_i} P(w_i) \log P(w_i) \tag{1.10}$$

Si la source génère des séquences de symboles de longueurs n, la valeur de l'entropie d'une langue L est alors calculée par la formule suivante :

$$H(W) = -\lim_{n\to\infty}\frac{1}{n}\left\{\sum_{w_i} P(w_1,w_2,...w_n)\log P(w_1,w_2,...w_n)\right\} \quad (1.11)$$

Quand l'entropie d'une langue *L* est estimée en utilisant un corpus de *n* mots, elle est par conséquent approximée par :

$$H(W) = -\frac{1}{n}\log P(w_1,w_2,...w_n) \quad (1.12)$$

La perplexité *PP*(*W*), qui est utilisée souvent pour évaluer la performance d'un modèle de langage, est définie comme suit :

$$PP(W) = 2^{H(W)} \quad (1.13)$$

Dans le cadre de la modélisation statistique du langage, l'entropie d'un modèle de langage *M* se calcule au moyen d'une quantité nommée *logprob LP*(*W*). Par exemple, dans le cas d'un modèle bigramme de mots, celle-ci est définie comme suit :

$$LP(W) = -\frac{1}{n}\log\left\{P(w_1)\prod_{i=2}^{n} P(w_i \mid w_{i-1})\right\} \quad (1.14)$$

et la perplexité *PP*(*W*) du modèle de langage bigramme dérivée d'un corpus est calculée comme suit :

$$PP(W) = 2^{LP(W)} = \left\{P(w_1)\prod_{i=2}^{n} P(w_i \mid w_{i-1})\right\}^{-\frac{1}{n}} \quad (1.15)$$

Le modèle de langage est appris à partir d'un corpus d'apprentissage et on évalue la perplexité avec ce modèle sur un corpus de texte différent du corpus d'apprentissage. Même si la perplexité est un bon indicateur de la qualité des modèles de langage, sa corrélation avec le taux de reconnaissance du système de reconnaissance automatique de la parole n'est pas certaine.

Le lecteur trouvera une explication détaillée sur la théorie de l'information telles que la quantité d'information, l'entropie dans [Hamming 1986, Cover 1991].

3.6. Reconnaissance de la parole pour une langue peu dotée

Comme l'illustre la figure 1.7, le développement d'un système de reconnaissance automatique de la parole continue à grand vocabulaire dans une nouvelle langue nécessite souvent de rassembler une grande quantité de corpus, contenant d'une part des signaux étiquetés de parole, pour l'apprentissage des modèles acoustiques du système, d'autre part des données textuelles, pour l'apprentissage des modèles statistiques du langage du système. Par ailleurs, la construction d'un système de reconnaissance automatique de la parole nécessite d'autres

ressources linguistiques telles que les vocabulaires et les dictionnaires de prononciation de bonne qualité.

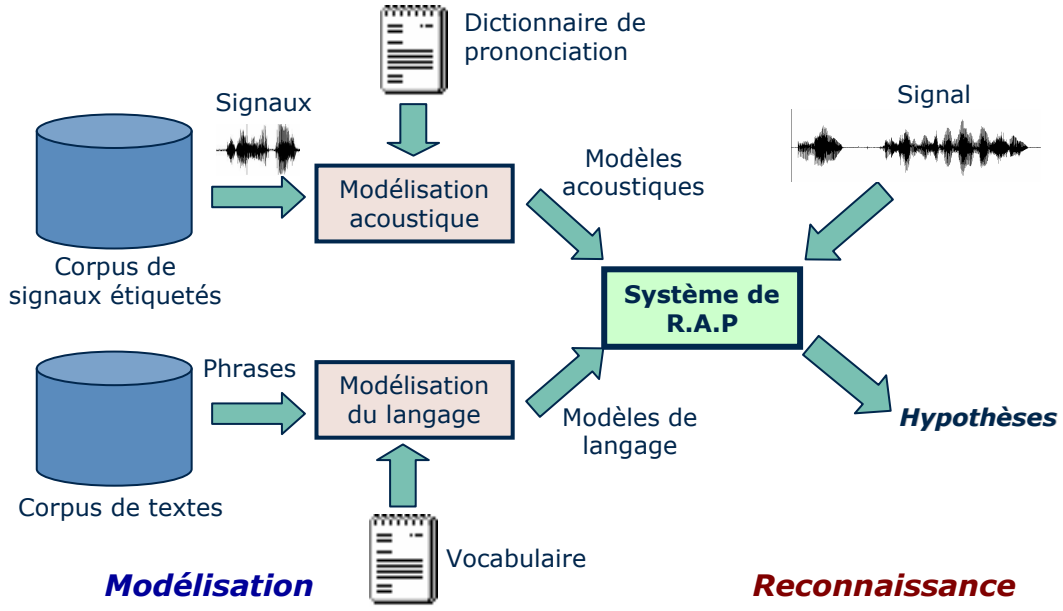


Figure 1.7 : Ressources nécessaires pour construire un système de reconnaissance vocale

De telles ressources acoustiques et linguistiques sont désormais disponibles pour la plupart des langues occidentales comme l'anglais, le français, l'espagnol, et pour quelques langues asiatiques comme le chinois, le japonais, le coréen. Ainsi, porter un système de reconnaissance vers une nouvelle langue est une tâche très fastidieuse si aucun corpus de grande envergure n'existe dans la langue cible, puisqu'il faut collecter toutes les ressources nécessaires. Précisons aussi qu'étant donné la nature statistique des modèles généralement utilisés en reconnaissance automatique de la parole, ces ressources doivent être disponibles en quantité importante.

4. Reconnaissance automatique de la parole multilingue

Le multilinguisme est au cœur des enjeux actuels concernant les échanges culturels et économiques, désormais mondialisés. Ainsi, le développement de services et d'interfaces adaptés à ce contexte donne lieu à de nouvelles problématiques dans le domaine du traitement automatique du langage naturel.

En ce qui concerne la communication homme-homme médiatisée par la machine, les recherches en traduction automatique de parole sont évidemment centrales [Waibel 2004] ; pour illustrer cela, on peut notamment citer les projets de traduction automatique CSTAR[1] et NESPOLE![2] dans lesquels le laboratoire CLIPS s'est impliqué. Concernant la transcription automatique, on a vu récemment émerger le thème de la *reconnaissance automatique de la parole multilingue*, qui fait désormais l'objet de sessions spéciales dans des conférences telles

[1] http://www.c-star.org/
[2] http://nespole.itc.it/

que ICSLP ou ICASSP [Special 2004]. En dehors des travaux pionniers de CMU sur ce thème [Schultz 2001, Waibel 2004], on peut citer les récentes études suivantes :

- transcription d'un flux audio bilingue (journaux télévisés avec présentateur galicien et reportages en espagnol standard [Dieguez-Tirado 2005]) ;
- traitement du langage naturel et ressources pour les langues minoritaires (action SALTMIL[1] de l'ISCA et le concept d'*E-Inclusion*[2]) ;
- travaux sur des langues asiatiques peu dotées (laotien [Berment 2004], thaï [Suebvisai 2005], indonésien [Martin 2005]) ;
- applications dans des environnements bilingues (travaux d'IBM sur la reconnaissance de noms propres anglais sur un flux audio en français canadien [Lejeune 2005]) ;
- reconnaissance bilingue de noms propres anglais - chinois [Ren 2005].

En résumé, il s'agit d'être capable non seulement de traiter rapidement de nouvelles langues, mais aussi de développer des systèmes capables de commuter d'une langue à l'autre (ce qui nécessite éventuellement l'utilisation d'une phase d'identification des langues [Fugen, 2003]). Tout cela relève d'un même domaine de recherche que l'on peut identifier par le terme générique de *reconnaissance automatique de la parole multilingue*.

4.1. Ressources linguistiques multilingues

D'après J. Godfrey et A. Zampolli [Cole 1997], le terme « ressources linguistiques » réfère à un (large) ensemble de données linguistiques, qui sont décrites sous forme exploitable par la machine, et utilisées pour construire, améliorer et évaluer des systèmes ou des techniques de traitement automatique des langues naturelles écrites ou orales.

Aujourd'hui, la plupart des techniques de reconnaissance automatique de la parole, notamment les systèmes de reconnaissance automatique de la parole continue grand vocabulaire, utilisent des approches statistiques. Cependant, la nature statistique des approches nécessite de disposer d'un grand nombre de données (textuelles et signaux) pour entraîner les modèles sous-jacents et tester les performances des systèmes. Par conséquent, un grand corpus de parole qui contient des dizaines d'heures de signaux enregistrés par une centaine de locuteurs (pour la modélisation acoustique) et un corpus de texte propre avec des millions de mots écrits (pour la modélisation statistique du langage) sont nécessaires pour le développement d'un système de reconnaissance automatique de la parole continue grand vocabulaire. Ces ressources ne sont, bien sûr, pas disponibles directement pour des langues peu dotées.

Avec l'émergence de la reconnaissance automatique de la parole multilingue, les techniques de portabilité des systèmes de reconnaissance automatique de la parole vers de nouvelles langues ont commencé à être étudiées dans la communauté scientifique [Godfrey 1994, Hovy 1999]. Ces domaines de la reconnaissance automatique de la parole multilingue nécessitent des

[1] Speech And Language Technology for MInority Languages : http://193.2.100.60/SALTMIL/
[2] voir session special Eurospeech 2005 : "*E-Inclusion for Spoken Language Processing*"

corpus de texte et de signaux multilingues. Ces corpus, qui couvrent un grand nombre de langues, doivent être « uniformes » à travers les langues. Cette uniformité se réfère non seulement à la quantité totale de données par langue mais aussi à la qualité des données telles que les conditions d'enregistrement (bruit, microphone, etc.), le scénario de collecte (tâche, type de locuteur, etc.) et le protocole de transcription de signaux [Schultz 2002].

Dans cette section, nous présentons les activités récentes de collection et distribution de ressources linguistiques (notamment les ressources textuelles et signaux multilingues) dans la communauté scientifique. Tout d'abord, un aperçu sur les organismes de création, de regroupement et de distribution des ressources linguistiques sera abordé. Et puis, nous présentons quelques corpus et bases de données de texte et de parole multilingues utilisés pour construire des systèmes de reconnaissance automatique de la parole multilingue.

4.1.1. Organismes de distribution des ressources linguistiques

Le consortium de données linguistiques - LDC[1], fondé en 1992 par l'ARPA[2] et géré par l'Université de Pennsylvanie (Etats-Unis), est un consortium composé de laboratoires de recherche universitaire, industrielle et gouvernementale. Il crée, collecte et distribue des corpus textuels et oraux, des lexiques et d'autres ressources pouvant être utilisées dans le cadre de recherches et développements sur le traitement automatique de la langue et l'identification de la langue parlée (corpus monolingue : TIDIGITS, TIMIT[3], SWITCHBOARD [Godfrey 1992] ; corpus téléphonique multilingue : POLYPHONE, OGI[4], CALLHOME[5], CALLFRIEND, …). LDC distribue des centaines de corpus vocaux différents[6] qui representent au total des milliers d'heures de signaux.

L'association européenne pour les ressources linguistiques - ELRA[7], une organisation à but non lucratif, a été fondée par la Commission Européenne en 1995 dans le but de distribuer des bases de données aussi bien pour la recherche que pour l'industrie. ELRA joue le rôle d'une organisation chargée de soutenir les activités en matière de création, de vérification et de distribution de ressources linguistiques en Europe. Jusqu'à maintenant, ELRA offre des ressources dans la plupart des langues majoritaires européennes (anglais, français, espagnol, allemand, italien, suédois, grec, néerlandais, portugais etc.) et d'autres langues du monde en quantité plus limitée (turc, japonais, russe, chinois, malais, irlandais, arabe, coréen).

De nos jours, sous la gestion et la distribution d'ELRA et de LDC, il existe des corpus textuels et des bases de données vocales avec la transcription orthographique et phonétique disponibles dans plus de 20 langues ; une dizaine de langues possèdent quant à elles des dictionnaires de prononciation distribués.

---

[1] The Linguistic Data Consortium: http://www.ldc.upenn.edu/
[2] The Advanced Research Projects Agency (maitenant: DARPA): http://www.darpa.mil/
[3] TIMIT - Acoustic-Phonetic Continuous Speech Corpora
[4] The OGI Multi-language Telephone Speech Corpus: http://cslu.cse.ogi.edu/corpora/mlts/
[5] CALLHOME est un corpus de la parole multilingue qui servira à construire des systèmes de reconnaissance de la conversation téléphonique grand vocabulaire (*Large Vocabulary Conversational Speech Recognition*).
[6] http://www.ldc.upenn.edu/Catalog/byType.jsp#speech
[7] The European Language Resources Association: http://www.elra.info/

4.1.2. Corpus de parole multilingues

Pour le développement des systèmes de reconnaissance automatique de la parole multilingues, il est nécessaire de construire de grandes bases de données (corpus) multilingues pour entraîner des modèles acoustiques et tester les systèmes. Dans cette section, nous présentons quelques corpus de signaux vocaux multilingues bien connus dans la communauté qui sont enregistrés dans des environnements différents.

a) GlobalPhone

La base de données multilingue GlobalPhone [Schultz 2002] a été collectée dans le cadre du Projet GlobalPhone[1]. En 2002, elle rassemble 15 langues : arabe, chinois (mandarin et changhaï), croate, tchèque, français, allemand, japonais, coréen, portugais, russe, espagnol, suédois, tamil et turc.

Pour chaque langue, 100 phrases de texte, collectées à partir des journaux économiques ou politiques nationaux et internationaux, sont lues par environ 100 locuteurs natifs. Cela correspond à 20 heures de signaux ou environ 100 000 mots parlés par langue. L'enregistrement des locuteurs a été fait dans les conditions environnementales de type bureau. Le projet GlobalPhone rassemble plus de 300 heures de signaux enregistrés par plus de 1500 locuteurs. La partie française a été enregistrée au CLIPS [Vaufreydaz 2000].

b) SpeechDat

Dans le cadre du projet *SpeechDat(II)* [2] [Hoge 1999], un total de 28 bases de données de parole a été collecté en couvrant la plupart des langues officielles de l'Union Européenne et aussi quelques variantes dialectales majoritaires et langues minoritaires. Chaque base de données de SpeechDat possède une transcription orthographique pour tous les fichiers de signaux et un dictionnaire phonétique qui contient tous les mots dans les transcriptions orthographiques. 20 bases de données ont été collectées à travers le réseau téléphonique fixe (FDB), 5 bases de données à travers le réseau mobile (MDB), et 3 bases de données ont été recueillies pour la vérification de locuteurs par téléphone (SDB). Les bases de données FDB et MDB sont enregistrées à partir de 500 à 5000 appels de locuteurs différents en mode session simple (sauf deux MDBs utilisent le mode multisessions). La durée de chaque session d'enregistrement est de 4 à 8 minutes. Ces bases de données sont prévues pour le développement d'un certain nombre d'applications telles que des services d'information (par exemple l'information d'horaires), des services de transaction (par exemple achats en ligne, opérations bancaires par téléphone) et d'autres services de traitement d'appel.

Des bases de données supplémentaires ont été enregistrées dans différents cadres : SpeechDat(E) (FDBs pour cinq langues européennes orientales et centrales), SALA (SpeechDat des pays Amérique latine), SpeechDat-Car (bases de données de SpeechDat enregistrées dans

[1] The GlobalPhone Project: http://www.cs.cmu.edu/~tanja/GlobalPhone/
[2] The SpeechDat Projects : http://www.speechdat.org/

une voiture) et une en anglais australien. Une partie des bases de données SpeechDat(II) est distribuée par l'ELRA.

c) OGI

Le corpus de parole téléphonique multilingue OGI contient des données de parole téléphonique enregistrées par 2052 locuteurs en 11 langues : anglais, farsi, français, allemand, hindi, japonais, coréen, mandarin, espagnol, tamil et vietnamien pour un total de 38,5 heures de signaux environ [Muthusamy 1992].

Le corpus OGI a été enregistré dans le but de recherches sur l'identification automatique de la langue parlée et la reconnaissance automatique de la parole multilingue. Le corpus téléphonique multilingue OGI est distribué par le LDC [Godfrey 1994]

4.2. Modélisation acoustique multilingue

Comme nous l'avons mentionné dans la section II de ce chapitre, un des composants importants d'un système de reconnaissance automatique de la parole est le dictionnaire de prononciation. Il indique les unités qui composent chaque entrée (ou mot) du dictionnaire. Les unités souvent utilisées sont les phonèmes. La qualité d'un système de reconnaissance automatique de la parole dépend fortement de l'exactitude du dictionnaire de prononciation.

En utilisant les unités acoustiques phonétiques, les similitudes entre les langues peuvent être exprimées à l'aide de l'Alphabet Phonétique International (API) [IPA 1999] et d'autres symboles phonétiques comme SAMPA[1], Worldbet [Hieronymus 1993] qui classifient des sons basés sur la connaissance phonétique. La figure 1.8 présente l'API pour les consonnes et les voyelles : les consonnes sont classifiées selon la position d'articulation, le type d'articulation, le voisement,… et les voyelles sont classifiées selon la position de la langue et l'arrondissement.

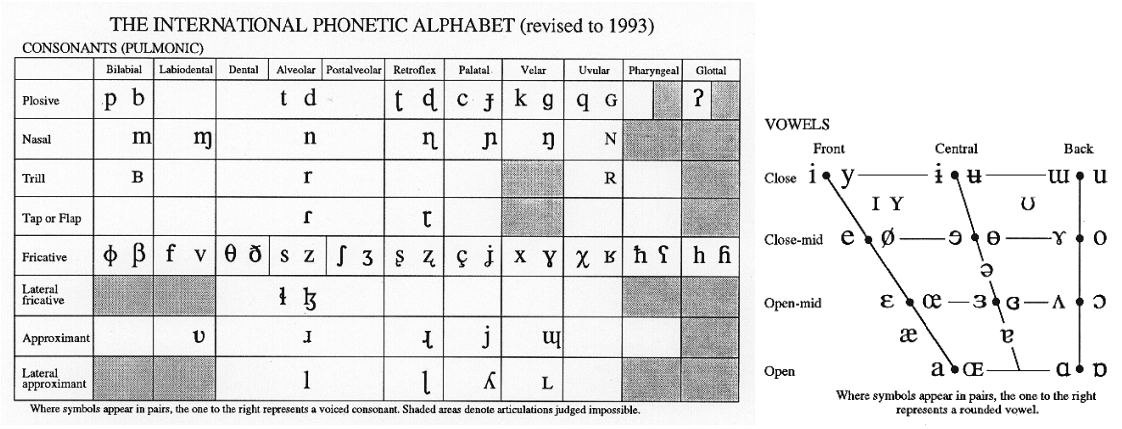


Figure 1.8 : API pour les consonnes et les voyelles [IPA 1999]

[1] Speech Assessment Methods Phonetic Alphabet: http://www.phon.ucl.ac.uk/home/sampa/

Les travaux de CMU dans le cadre du Projet GlobalPhone [Schultz 1999, Schultz 2001, Schultz 2002] consistent à concevoir un système de reconnaissance automatique de la parole multilingue basé sur les phonèmes. À partir de l'inventaire de l'alphabet phonétique international, ils ont défini un ensemble de phonèmes globaux nommé *Global Phoneme Set*. Le tableau 1.5 est un exemple de l'ensemble des phonèmes globaux obtenus avec 5 langues (croate, japonais, coréen, espagnol et turc) en notation Worldbet. Il regroupe au total 78 phonèmes ainsi qu'un silence et deux modèles de bruit pour modéliser les effets spontanés de la parole. Il y a 14 phonèmes qui sont partagés entre chacune des 5 langues, et la moitié de cet ensemble n'appartient qu'à une seule langue. Il est intéressant de constater qu'en utilisant l'ensemble des phonèmes globaux, les phonèmes multilingues sont groupés et partagés entre les langues, ce qui réduit le nombre total d'unités acoustiques à modéliser (78 phonèmes multilingues contre 170 phonèmes monolingues dans le cas des 5 langues de GlobalPhone). La figure 1.9 illustre les fréquences relatives des consonnes pour 5 langues dans la base de données d'apprentissage de GlobalPhone.

Phonèmes [WorldBet]	KO	SP	CR	TU	JA	∑
n, m, s, l, tS, p, b, t, d, g, k, i, e, o	x	x	x	x	x	14
f, j, z		x	x	x	x	
r, u	x	x	x	x	x	6
dZ	x		x	x		
a	x	x	x			
S			x	x	x	4
h	x			x	x	
4	x	x			x	
n̄, x, L		x	x			
A				x	x	
N	x	x				10
V, Z			x	x		
y, 7	x			x		
ts			x		x	
p', t', k', dZ', s', oE, oa, 4i, uE, E, /\, i/\, u/\, iu, ie,	x					17
io, ia	x					
D, G, T, V, r(, ai, au, ei, eu, oi, a+, e+, i+, o+, u+		x				15
palatal c, palatal d			x			2
ix, soft				x		2
?, Nq, V[, A:, e:, i:, o:, 4:					x	8
Monolingue ∑ = 170	40	40	30	29	31	
Multilingue						78

Tableau 1.5 : Ensemble de phonèmes globaux (Global Phoneme Set) [Schultz 1999]

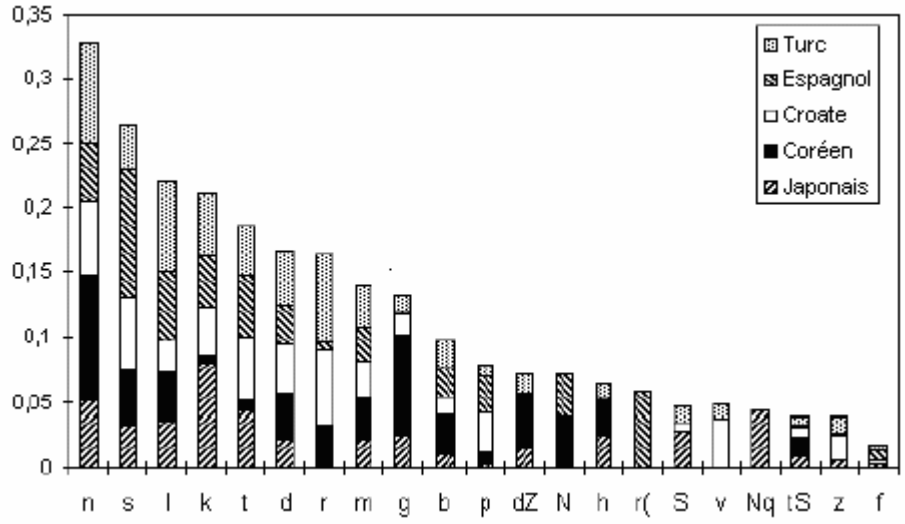


Figure 1.9 : Fréquences relatives des consonnes pour cinq langues de GlobalPhone [Schultz 1999]

En ce qui concerne la modélisation acoustique multilingue, T. Schultz introduit 3 méthodes différentes pour combiner des modèles acoustiques [Schultz 2001] : ML-sep (séparation des langues), ML-mix (mélange des langues), ML-tag (étiquetage des langues). Pour la méthode ML-sep, chaque modèle du phonème dans une langue est appris uniquement sur les données correspondantes de cette langue ; il n'y a pas de données échangées entre les langues. Pour la méthode ML-mix, les données sont partagées entre les langues pour entraîner des modèles communs. La méthode ML-tag est une méthode hybride où chaque modèle du phonème appartenant à une langue est associé par une étiquette de langue afin de préserver l'information de cette langue. Pour la méthode ML-tag, les composants gaussiens *(codebook weights)* des modèles du phonème sont partagés entre les langues (comme la méthode ML-mix) tandis que les distributions multigaussiennes *(distribution weights)* des modèles sont apprises séparément (comme la méthode ML-sep). Cette modélisation acoustique multilingue est une modélisation relativement indépendante de la langue[1] car le système de reconnaissance automatique de la parole multilingue obtenu peut, en plus de reconnaître la parole des langues concernées, reconnaître de la parole d'une nouvelle langue en adaptant les modèles acoustiques avec une quantité limitée de données en langue cible[2].

Récemment, des travaux sur la modélisation acoustique à base de graphèmes ont vu le jour [Killer 2003b, Kanthak 2003]. La représentation orthographique (caractère ou graphème) y est utilisée comme l'unité de modélisation acoustique à la place du phonème. Les résultats d'expérimentations présentés montrent que, si la modélisation indépendante du contexte à base de graphèmes n'est pas efficace, la modélisation dépendante du contexte obtient en revanche des résultats comparables avec les approches à base de phonèmes pour des langues telles que l'anglais et surtout l'allemand (qui présente une faible différence graphèmes / phonèmes). En

[1] Language Independent Acoustic Modeling
2 Language Adaptive Acoustic Modeling

conséquence, avec les langues peu dotées qui ne possèdent pas encore un dictionnaire phonétique, cette approche de modélisation acoustique à base de graphèmes présente un potentiel intéressant pour construire rapidement un système de reconnaissance automatique de la parole (ceci sera abordé aux chapitres 4 et 5 concernant notre travail sur la langue vietnamienne et khmère).

Pour la reconnaissance automatique de la parole multilingue à base de graphèmes, on peut également citer les travaux de S. Kanthak [Kanthak 2003]. Le tableau 1.6 présente son utilisation des graphèmes à travers quatre langues : hollandais, allemand, français et italien. La plupart des 26 caractères de l'alphabet latin est partagée par toutes les langues (seul le graphème *'q'* est absent dans le vocabulaire hollandais). Par contre, le français et l'allemand sont des langues qui possèdent des caractères privées qui n'existent pas dans les 3 autres langues. En comparaison du nombre de phonèmes dans le tableau 1.5, les graphèmes semblent donner une meilleure couverture à travers ces quatre langues.

Graphèmes	DU	FR	IT	GE	Total
a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, r, s, t, u, v, w, x, y, z	x	x	x	x	25
q		x	x	x	1
à, â, ç, è, é, ê, ë, ï, ô, û		x			10
ä, ö, ü, β				x	4
Monolingue Σ = 117	25	36	26	30	
Multilingue					40

Tableau 1.6 : Utilisation des graphèmes à travers des langues [Kanthak 2003]

4.3. Portabilité des modèles acoustiques vers une nouvelle langue

En reconnaissance automatique de la parole, si les ressources orales et écrites sont en grande quantité (des dizaines à centaines d'heures de signaux vocaux par exemple), et si un dictionnaire de prononciation est disponible pour une langue cible, l'adaptation du système de reconnaissance peut correspondre alors à un simple réapprentissage des modèles sur ces nouvelles données. Dans le contexte des langues peu dotées, la quantité de données collectées reste bien souvent inférieure à ce qu'elle est pour les langues bien dotées. La construction d'un système de reconnaissance automatique de la parole nécessite donc également des techniques d'adaptation rapide au niveau des modèles acoustiques comme cela est proposé dans [Kunzmann 2001] et [Schultz 2001] par exemple.

Une approche possible consiste à obtenir un tableau de correspondances phonémiques (*phone mapping*) entre une ou plusieurs langues sources, et la langue cible. Ensuite, les modèles acoustiques des phonèmes en langue source peuvent être dupliqués pour obtenir des modèles acoustiques en langue cible. L'avantage d'une telle approche est qu'elle ne nécessite pas ou peu

de signaux d'apprentissage en langue cible puisque les modèles acoustiques du système de reconnaissance en langue cible sont en fait ceux d'une autre langue. Cependant, on retrouve dans cette approche le problème de la couverture phonémique (i.e. reconnaître du vietnamien avec des modèles acoustiques appris sur du français). L'intérêt d'une telle approche est donc surtout pour l'étape de 'bootstraping' des systèmes, ceux-ci pouvant ensuite être améliorés en adaptant, par exemple, les modèles acoustiques avec une quantité réduite de signaux en langue cible.

Un problème important consiste donc à obtenir le tableau de correspondances phonémiques entre langue cible et langue source. Pour cela, on distingue les méthodes manuelles à base de connaissances (*knowledge-based*), des méthodes automatiques (*data-driven*). Les méthodes manuelles consistent à chercher les couples de phonèmes source/cible les plus proches dans le tableau d'API et nécessitent des connaissances acoustiques et phonétiques des deux langues (source et cible). Une approche automatique consiste plutôt à disposer d'un corpus vocal en quantité limitée en langue cible et étiqueté, puis à utiliser un décodeur phonémique (utilisant une ou plusieurs langues sources) et calculer la matrice de confusion entre les phonèmes reconnus en langue source et les phonèmes de référence en langue cible.

Le tableau de correspondances phonémiques entre langue cible et source peut être obtenu en calculant la distance des modèles acoustiques source/cible ou la distance entre phonèmes. J-J. Sooful a comparé quelques méthodes d'évaluation de la distance entre deux phonèmes : distance de Kullback-Leibler, distance métrique Bhattacharyya, distance euclidienne, etc [Sooful 2001]. Pour implémenter ces méthodes, on doit déjà disposer de modèles acoustiques de bonne qualité en langue source et cible. Cela devient très difficile dans le contexte de notre travail puisque nous disposons seulement d'une faible quantité de données vocales en langue cible.

A l'université John Hopkins, Baltimore, USA [Beyerlein 1999], un travail consistant à améliorer les systèmes de reconnaissance vocale pour des langues peu dotées est en cours. Ils utilisent la méthode à base de connaissances et la méthode automatique pour construire les tableaux de correspondances des unités phonétiques (phonème ou sous-phonème) entre une (cas monolingue) ou plusieurs (cas multilingue) langues sources et une langue cible. Quelques techniques d'adaptation des modèles de Markov caché : MLLR[1] [Leggetter 1995], MAP[2] [Gauvain 1994] sont employées en utilisant des signaux de la parole en quantité limité en langue cible. Une nouvelle méthode nommée *Discriminative Model Combination* (DMC) [Beyerlein 1998] est développée pour combiner les modèles acoustiques (au niveau phonème et sous-phonème) dans plusieurs langues sources différentes. Ils ont montré que la faible performance des modèles acoustiques en langue cible (langue tchèque) peut être améliorée en combinant, par la méthode DMC, différents modèles acoustiques en langues sources (anglais, espagnol, russe, et mandarin).

[1] Maximum Likelihood Linear Regression
[2] Maximum A Posteriori

Dans le cadre du projet GlobalPhone, plusieurs techniques de portabilité et d'adaptation des modèles acoustiques vers une nouvelle langue sont réalisées [Schultz 2001]. Une méthode efficace nommée PDTS (*Polyphone Decision Tree Specialization*) est proposée par T. Schultz et appliquée sur l'arbre de décision multilingue pour résoudre le problème de la disparité des contextes (*context mismatch*) à travers les langues.

En résumé, la modélisation acoustique multilingue et les approches de portabilité et d'adaptation des modèles acoustiques multilingues présentent un potentiel intéressant pour notre travail sur les langues peu dotées. Nous disposons, en effet, de seulement une faible quantité de données vocales en langue cible, et les modèles acoustiques multilingues nous permettent d'une part de construire rapidement des modèles acoustiques initiaux en langue peu dotée (bootstraping) et d'autre part d'améliorer la performance des modèles acoustiques en utilisant indirectement les données d'apprentissage « crosslingues ». Par conséquent, un de nos objectifs de travail est d'étudier et de proposer des solutions efficaces et des outils permettant d'accélérer la portabilité et l'adaptation des modèles acoustiques vers une nouvelle langue peu dotée.

Chapitre 2

Recueil rapide de ressources textuelles et modélisation statistique du langage

D'après J. Godfrey et A. Zampolli [Cole 1997], le terme « ressources linguistiques » réfère à un (large) ensemble de données linguistiques, qui sont décrites sous forme exploitable par la machine, et utilisées pour construire, améliorer et évaluer des systèmes ou des techniques de traitement automatique des langues naturelles écrites ou orales. Des exemples de ressources linguistiques sont les corpus textuels et vocaux, les bases de données lexicales, les grammaires, etc. Ce chapitre présente le recueil rapide des ressources textuelles dans le contexte de reconnaissance automatique de la parole. Tout d'abord, nous présentons le recueil d'un vocabulaire, composante indispensable pour la construction d'un système de reconnaissance automatique de la parole. Ensuite, des solutions sur la récupération et le traitement d'un corpus de texte pour des langues peu dotées (notamment pour les langues vietnamienne et khmère) sont proposées et validées. Ce chapitre se termine par la modélisation statistique du langage pour des langues peu dotées.

1. Recueil d'un vocabulaire

Dans le domaine du traitement automatique de la parole, les systèmes de transcription ont évolué depuis des tâches de reconnaissance à vocabulaires limités, vers des tâches à grands et très grands vocabulaires dans un contexte de dialogue interactif [Church 2003]. Un vocabulaire, dans notre contexte de travail, est défini comme une liste close d'unités lexicales qui peuvent être reconnues par un système de reconnaissance automatique de la parole. La taille du vocabulaire et la sélection des unités lexicales dans le vocabulaire influencent fortement les performances du système de transcription automatique (la perplexité des modèles de langages, l'espace de recherche, le taux de reconnaissance, …) puisque tous les mots hors-vocabulaire ne peuvent pas être reconnus par le système.

Nous abordons dans cette section deux méthodes de recueil d'un vocabulaire pour le traitement automatique de la parole :

1. récupération d'un vocabulaire à partir de ressources lexicales existantes ;
2. génération automatique d'un vocabulaire à partir de données textuelles.

En fait, pour les applications du traitement automatique de la parole, la première méthode de récupération (décrite dans la section 1.1 suivante) peut être utilisée pour une langue ou une

tâche spécifique qui possède déjà un vocabulaire de bonne qualité, c'est-à-dire avec une bonne couverture lexicale. La couverture lexicale indique le taux des mots du vocabulaire présents dans un corpus de texte. L'avantage de cette méthode réside dans l'assurance de la qualité du vocabulaire car il est construit par des linguistes. Par ailleurs, pour des langues ou des tâches spécifiques qui ne possèdent pas de vocabulaire ayant une bonne couverture lexicale, nous pouvons appliquer la seconde méthode (décrite dans la section 1.2) pour générer automatiquement un vocabulaire. Cependant la qualité du vocabulaire dépend de la qualité de la source de données textuelles utilisée.

1.1. Récupération d'un vocabulaire à partir de ressources lexicales existantes

Un dictionnaire est une liste de mots (morphèmes libres, mots composés, expressions lexicalisées) classés sous leur lemme (ou entrée), accompagnés de leur définition ou leur correspondance dans une autre langue. Depuis longtemps, les dictionnaires deviennent un outil indispensable dans la vie quotidienne : apprendre et maîtriser une langue, consulter des mots scientifiques, etc. En plus, le dictionnaire joue un rôle indispensable dans tout le domaine du traitement de la langue naturelle et de la communication langagière : dialogue homme-machine, traduction automatique assistée par ordinateur, traitement de la langue parlée (synthèse et reconnaissance), etc.

Les prototypes utilisant de petits dictionnaires deviennent de moins en moins convaincants, les besoins s'orientent vers de grandes quantités d'informations lexicales. Cependant, la réalisation d'une base lexicale de qualité est une tâche lourde dans le processus d'informatisation d'une nouvelle langue. Nous pouvons utiliser des approches et des outils de récupération et de production d'un nouveau dictionnaire, à partir de ressources dictionnairiques informatisées multilingues hétérogènes [Doan-Nguyen 1998], ou par un travail coopératif sur l'Internet [Boitet 2001, Berment 2004]. Le travail coopératif est, par exemple, le principe du Projet Wikipédia[1].

Avec le développement des technologies de traitement des langues naturelles, il y a de plus en plus de ressources dictionnairiques disponibles sur support informatique. Elles peuvent être de différentes formes telles que fichiers de données dans un logiciel de dictionnaire sur l'ordinateur, bases de données lexicales dans un système de traitement des langues naturelles, dictionnaire gratuit publié sur l'Internet,… Par exemple, dans le contexte du projet Papillon[2], une base lexicale multilingue comprenant entre autre l'allemand, l'anglais, le français, le japonais, le malais, le lao, le thaï, le vietnamien et le chinois est en cours de construction. Ce projet est totalement ouvert pour permettre l'élargissement arbitraire du nombre de langues, et mettre en œuvre un schéma de construction coopérative et d'utilisation mutualisée. Un serveur a été développé pour permettre à tout le monde de consulter librement la base, et d'y contribuer éventuellement, tout en garantissant la protection de la base (qualité, éthique) [Boitet 2001].

Grâce à ces communautés en ligne « ouvertes », nous pouvons obtenir des ressources dictionnairiques très variées en format, en quantité, en qualité, en information linguistique, etc.

[1] Wikipédia : http://www.wikipedia.org/ et Wiktionary : http://wiktionary.org/
[2] http://www.papillon-dictionary.org

qui répondent à nos besoins. Pour développer un système de reconnaissance automatique de la parole, ces dictionnaires peuvent être choisis et filtrés pour obtenir une liste de mots (un vocabulaire) dans la langue considérée.

Dans notre travail, nous avons récupéré un dictionnaire bilingue franco-vietnamien à partir du projet Papillon. Après avoir filtré le dictionnaire bilingue, le vocabulaire se compose d'environ 40 000 mots en vietnamien répartis en mots isolés, mots composés et mots empruntés[1]. Pour générer un vocabulaire de syllabes, nous avons filtré ce vocabulaire de 40 000 mots. Le vocabulaire de syllabes obtenu contient finalement 6 686 syllabes. Par ailleurs, à partir du projet KhmerOS[2] (*Khmer Software Initiative*), un vocabulaire traditionnel « Chuon Nat » a été obtenu pour la langue khmère. Ce vocabulaire se compose de 16 000 mots khmers.

1.2. Génération d'un vocabulaire à partir de données textuelles

Un vocabulaire dans une langue considérée peut être généré automatiquement par un outil qui estime la densité des unités lexicales dans un ou plusieurs grands corpus de texte. La densité représente l'occurrence d'une unité lexicale dans le corpus de texte. Le problème de la construction automatique d'un vocabulaire consiste à obtenir la meilleure couverture sur un ou plusieurs corpus de texte. Pour construire un vocabulaire à partir d'un corpus de texte, le nombre d'entrées et les entrés lexicales du vocabulaire sont choisies empiriquement, selon la couverture désirée. Par exemple, on sélectionne simplement 20 000 entrées les plus fréquentes observées sur un corpus de texte. Si on a plusieurs corpus de texte, une méthode de détermination de combinaison optimale des vocabulaires produits par les différents corpus a été proposée dans [Allauzen 2004].

Cependant, ces techniques de construction automatique s'avèrent difficiles pour les langues peu dotées pour les raisons suivantes :

1. la qualité de la source de données textuelles est très variable ainsi que celle des outils de traitement du corpus de texte ;

2. pour des langues non-segmentées en mots ou syllabes (langue khmère par exemple), la plupart des algorithmes de segmentation en mots ou syllabes nécessitent justement un vocabulaire prédéfini !

La qualité du vocabulaire généré automatiquement dépend ainsi de la qualité de la source de données textuelles utilisée. Par ailleurs, la technique n'est applicable que sur des textes segmentés en mots (langues occidentales par exemple) ou sur des textes segmentés en syllabes (chinois, coréen, japonais, vietnamien).

Dans notre travail, nous ne pouvons appliquer cette technique d'obtention automatique d'un vocabulaire qu'en vietnamien. Pour le vietnamien, à cause de la faible qualité du corpus de texte utilisé, la technique de génération automatique d'un vocabulaire pour la reconnaissance automatique de la parole en langue vietnamienne consiste en deux étapes : enrichissement d'un

[1] les mots empruntés franco-vietnamiens par exemple.
[2] http://www.khmeros.info/

vocabulaire existant puis limitation de la taille du vocabulaire. Nous rappelons que la langue vietnamienne est une langue segmentée en syllabes. Le vocabulaire initial utilisé dans cette section contient 6 686 syllabes en vietnamien (appelé V_0).

1.2.1. Enrichissement d'un vocabulaire existant

Tout d'abord, à partir d'un grand corpus de texte, nous pouvons choisir empiriquement les formes les plus fréquentes selon une couverture lexicale désirée. Par exemple, sur le corpus de texte vietnamien, nous avons choisi une liste de 2000 formes lexicales les plus fréquentes, appelée V_1. Ainsi, les formes lexicales de V_1 qui existent déjà dans le vocabulaire V_0 sont exclues. Le reste de la liste V_1, appelé V_2, contient 180 formes (figure 2.1) telle que : $V_2 = V_1 \setminus V_0$.

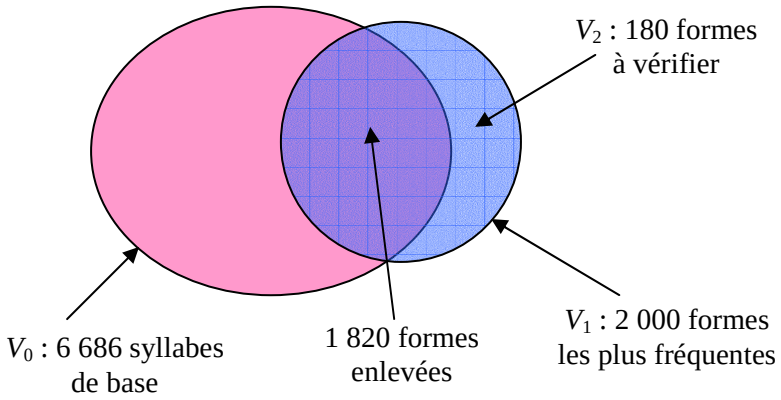


Figure 2.1 : Vérification des formes lexicales

Nous notons qu'à l'aide d'un analyseur phonétique vietnamien, nous pouvons vérifier si une forme lexicale est considérée comme une syllabe vietnamienne, même si celle-ci n'est pas dans le vocabulaire de syllabes V_0. Alors, la liste de formes exclues V_2 peut être divisée en deux parties :

1. *les formes lexicales « vietnamiennes »* ;
2. *les formes lexicales « étrangères » et les fautes orthographiques*. Les formes lexicales « étrangères » ne sont pas des mots vietnamiens mais elles sont souvent utilisées dans la vie quotidienne, comme par exemple les mots 'Internet', 'e-mail', 'mobile', 'fax', 'Euro', 'USD', etc.

Ces deux types de formes lexicales peuvent être vérifiées manuellement. Une liste de 114 nouvelles entrées lexicales, appelée V_3, est ajoutée au vocabulaire V_0. On obtient ainsi un vocabulaire appelé V_4 de 6 686 + 114 = 6 800 entrés lexicales.

1.2.2. Limitation de la taille du vocabulaire

La taille du vocabulaire influence fortement les performances du système de reconnaissance automatique de la parole. En effet, si on augmente la taille du vocabulaire, la taille du modèle de langage et l'espace de recherche du système de reconnaissance croient proportionnellement. Par

exemple, quand nous augmentons la taille du vocabulaire du vietnamien de 2 000 à 20 000, la taille des modèles de langage passe de 119 Mo à 312 Mo (voir le tableau 2.1). De plus, nous avons besoin de plus de données textuelles pour entraîner les modèles de langage. Cependant, si la taille du vocabulaire est petite, le taux de mots hors-vocabulaire du système augmente.

Nombre de mots du vocabulaire	Taux de couverture lexicale	Taille des modèles de langage	Nombre de 2-grammes	Nombre de 3-grammes
2 000	82,4 %	119 Mo	1 363 640	2 835 460
20 000	99,5 %	312 Mo	5 517 800	5 039 753

Tableau 2.1 : Influence de la taille du vocabulaire

Nous avons effectué une analyse de la couverture lexicale, selon la taille du vocabulaire de 40 000 mots et du vocabulaire de 6 800 syllabes enrichies, sur un grand corpus de textes en vietnamien collecté à partir du Web. Nous constatons que les 20 000 mots ou 3 500 syllabes (~50% des mots et syllabes) les plus fréquents du corpus de texte couvrent déjà 99,5% des mots du corpus (figures 2.2 et 2.3).

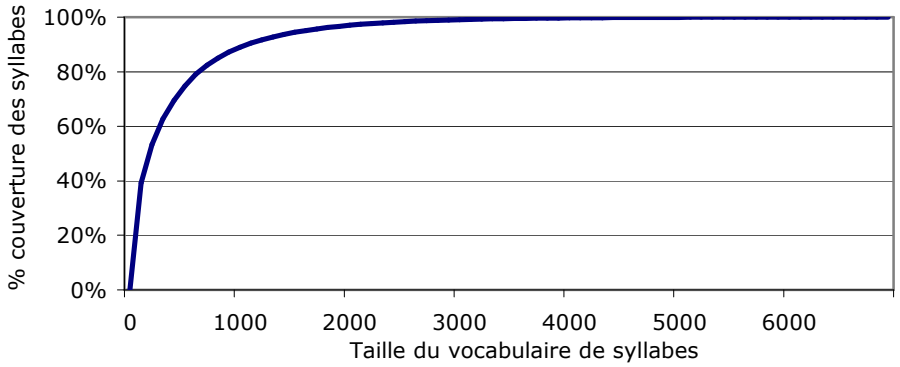


Figure 2.2 : Couverture lexicale sur le corpus aspiré du Web selon la taille du vocabulaire de syllabes du vietnamien

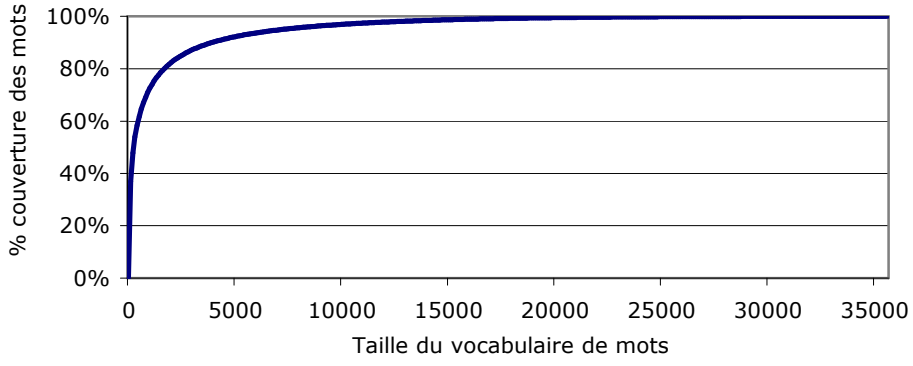


Figure 2.3 : Couverture lexicale sur le corpus aspiré du Web selon la taille du vocabulaire de mots du vietnamien

En conclusion, la sélection des entrées lexicales et la taille appropriée du vocabulaire sont très importantes pour le système de reconnaissance automatique de la parole. La taille optimale du vocabulaire dépend de la tâche pour laquelle le système est dédié et du système lui-même [Rosenfeld 1995]

2. Récupération d'un corpus de textes pour des langues peu dotées

Concernant le recueil de données textuelles en grande quantité pour la construction des systèmes de reconnaissance automatique de la parole, une approche intéressante consiste à « télécharger » un grand nombre de sites Web dans la langue donnée et à filtrer les données récupérées pour les rendre exploitables. Ces données textuelles peuvent servir d'une part à calculer des modèles statistiques du langage, et d'autre part à obtenir un corpus pouvant ensuite être prononcé par des locuteurs en vue de la constitution d'une base de signaux conséquente (figure 2.4).

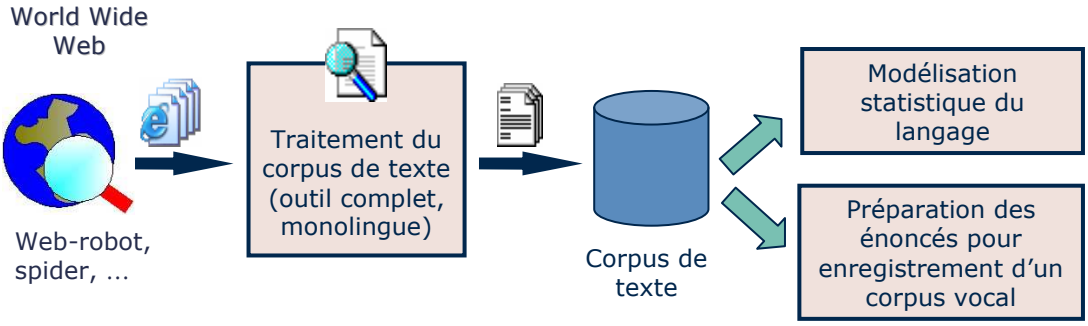


Figure 2.4 : Procédé de récupération de données textuelles sur le Web pour la construction d'un système de reconnaissance automatique de la parole

Une telle approche a déjà été relativement bien validée pour une langue bien dotée telle que le français [Vaufreydaz 2002]. Cependant, en appliquant cette méthode sur les langues peu dotées, les problèmes spécifiques concernent le nombre de sites Web peu important, la faible vitesse de transmission, la qualité variable des documents qui nécessitera alors plus d'outils de traitements. Par exemple, la séparation en mots s'avère extrêmement difficile pour des systèmes d'écriture comme le khmer, surtout si l'on ne dispose pas d'un vocabulaire. Dans cette section, nous présenterons une méthode de collecte de documents pour les langues peu dotées. Les traitements spécifiques pour des langues peu dotées et une approche générique « multilingue » de récupération et de traitement rapide d'un corpus de texte seront notamment présentés.

2.1. Collecte de documents à partir de l'Internet

Comme illustré par la figure 2.5, de nos jours, la plupart des pays du monde sont connectés à l'Internet qui devient ainsi un élément indispensable dans notre vie et travail quotidien.

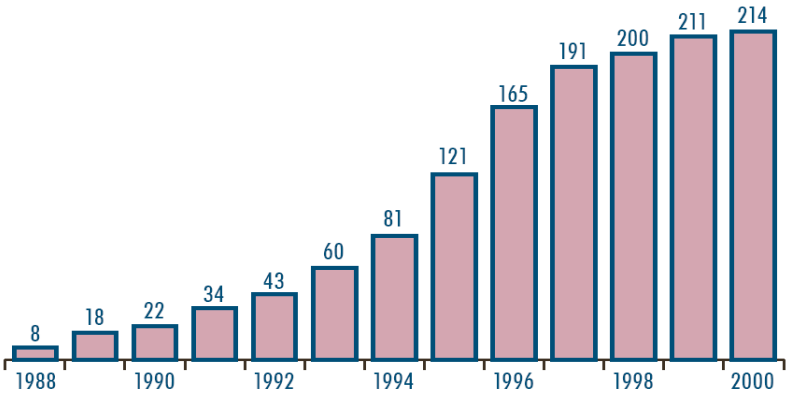


Figure 2.5 : Nombre de pays connectés à l'Internet [UIT 2001]

D'après des statistiques de l'Union Internationale des Télécommunications [UIT 2001], le nombre de serveurs Web au monde a augmenté de 0,1 millions d'hôtes en 1996 à 26 millions d'hôtes en 2000 (figure 2.6).

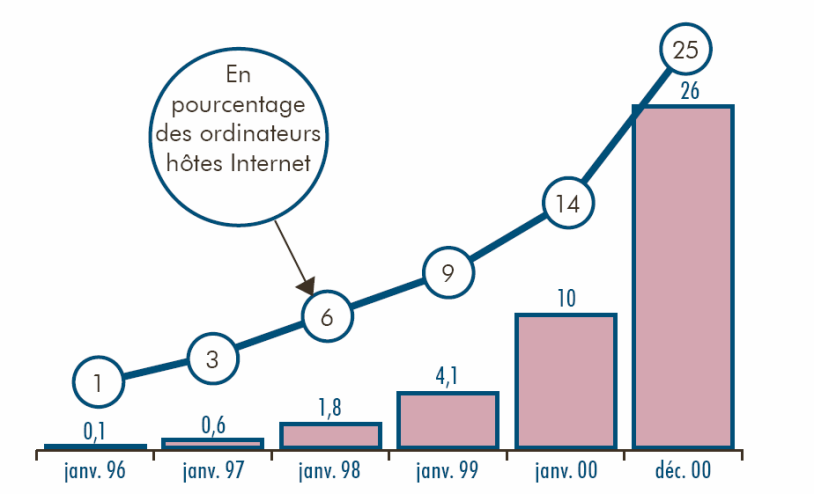


Figure 2.6 : Nombre de serveurs Web, en million [UIT 2001]

Le World Wide Web est devenu une des plus importantes sources d'information disponible de manière électronique, particulièrement pour des langues peu dotées. Le Web constitue ainsi un corpus gratuit, riche, important et accessible pour de nombreuses langues.

Plusieurs recherches se concentrent actuellement sur la construction de corpus de textes en grande quantité en collectant des pages Web (documents pertinents dans une langue quelconque). Pour la construction d'un corpus de texte à partir de l'Internet, les documents *html* sont récupérés. Nous allons aborder des méthodes de récupération suivantes :

- utilisation d'un robot du Web ;
- récupération de documents pertinents utilisant des moteurs de recherche ;
- détermination des sites pertinents et récupération régulière des documents html.

2.1.1. Utilisation d'un robot Web

D. Vaufreydaz dans la cadre de sa thèse [Vaufreydaz 2002] a proposé une approche exploitant les ressources du Web pour construire un grand corpus de texte en français. Par un jeu complexe d'heuristiques et de filtres, un robot parcours des pages web et collecte les informations exploitables. Ces travaux permettent de répondre au besoin de très grand corpus, obtenus de façon quasi-automatique. En effet, en partant d'un document situé n'importe où sur le Web, il est possible d'accéder à des millions de pages html car les robots peuvent atteindre et trouver tous les documents de texte et aussi les pages Web qui ont un lien direct ou indirect avec ce point de départ. Cette approche a été validée pour le français [Vaufreydaz 2001] qui possède déjà un grand nombre de sites Web.

2.1.2. Récupération de documents pertinents utilisant des moteurs de recherche

Dans le domaine de la recherche d'information, il existe des approches à base de requêtes lancées sur un moteur de recherche pour collecter des documents pertinents dans une langue considérée [Nishimura 2001, Scannell 2003, Monroe 2002]. La figure 2.7 illustre une méthode de recueil automatique d'un corpus de texte par des requêtes entraînables proposé et développé dans l'outil *CorpusBuilder*[1] [Ghani 2005].

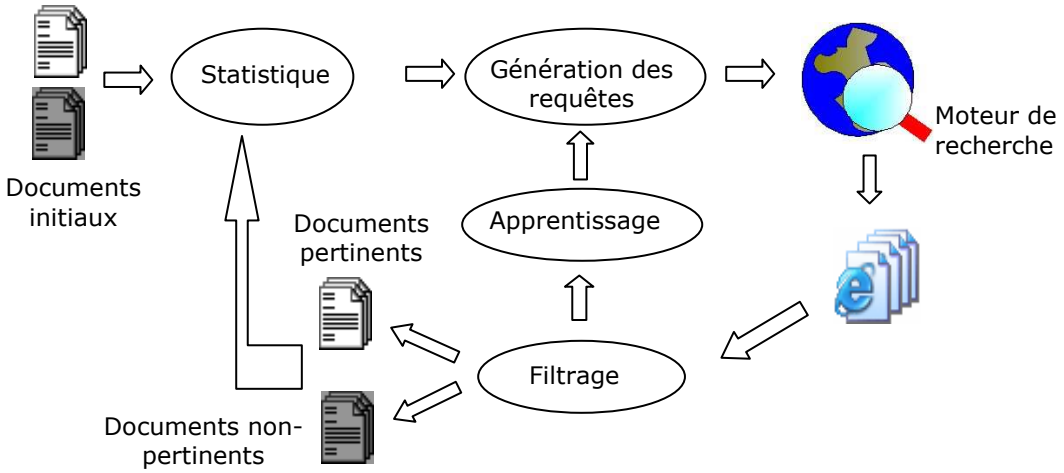


Figure 2.7 : Récupération de documents utilisant des moteurs de recherche [Ghani 2005]

D'abord, la collecte manuelle de quelques documents initiaux en langue cible permet d'extraire des requêtes initiales. Ensuite, ces requêtes sont lancées sur un des moteurs de recherche connus comme Google[2], Yahoo! Search[3], AltaVista[4], MSN Search[5], etc. Les documents renvoyés par les moteurs de recherche sont collectés et sauvegardés sur la machine locale. Puis, ces documents sont séparés en un groupe de documents dans la langue cible (documents pertinents ou documents positifs) et un groupe contenant des documents dans d'autres langues (documents non-pertinents ou documents négatifs). Cette catégorisation est effectuée à l'aide d'un outil

[1] http://www.cs.cmu.edu/~TextLearning/corpusbuilder/
[2] http://www.google.fr
[3] http://search.yahoo.com
[4] http://www.altavista.com/
[5] http://search.msn.com/

d'identification de la langue écrite et du système d'encodage de caractères (par exemple *TextCat*[1] permet de catégoriser des textes) ; elle est basée sur des modèles *n*-grammes de caractères. Les documents récupérés sont analysés, puis de nouvelles requêtes sont extraites et on met à jour les modèles *n*-grammes de caractères.

Pour les langues qui possèdent déjà un grand nombre de documents Web indexés par les moteurs de recherche, les approches à base de requêtes peuvent être utilisées pour récupérer un grand corpus dans une langue spécifique ou dans une tâche spécifique. La qualité des résultats de la recherche effectuée par les moteurs de recherche dépend fortement de la qualité de la formulation de la requête obtenue à partir de documents. Ainsi, formuler une requête qui représente correctement notre besoin est un challenge.

2.1.3. Récupération de documents pour des langues peu dotées

Les deux méthodes précédentes ne sont pas très efficaces dans le contexte particulier des langues peu dotées. Comme illustré dans le tableau 2.2, le nombre d'hôtes de l'Internet en langues peu dotées (par exemple khmer ou laotien) et les bandes passantes dans ces pays sont très faibles par rapport aux langues bien dotées.

Langue	Nombre d'hôtes de l'Internet (cf. INTERNIC[2] 2001)	Utilisateurs de l'Internet (cf. CIA[3], 2003)	Bande passante de l'Internet [UIT-ICS 2001]
français	325 103	21 900 000	191 898 Mbps
vietnamien	9 037	3 500 000	34 Mbps
khmer	818	30 000	6 Mbps
laotien	937	15 000	2 Mbps

Tableau 2.2 : Difficultés de la récupération des documents à partir de l'Internet pour des langues peu dotées

Il est par conséquent nécessaire de développer une méthode de récupération d'un corpus de documents html à partir de l'Internet plus efficace dans le contexte spécifique des langues peu dotées.

Le point de départ de la méthode consiste à déterminer manuellement des sites Web riches en ressources et présentant un haut débit de transmission. On préfèrera par exemple des sites de nouvelles, au fort contenu rédactionnel tels que *VnExpress*[4] pour le vietnamien et *Cambodiacic*[5] pour le khmer. Si aucune information sur le site n'existe dans la langue cible, nous pouvons utiliser la technique de récupération des documents pertinents à base de moteurs de recherche. Bien que cette technique ne soit pas très efficace pour collecter un grand nombre de documents du Web, elle nous permet de trouver plus facilement des adresses de sites pertinents en langue cible, car les sites pertinents disposent souvent de bons poids d'indexation sur les moteurs de

[1] Text categorization tool: http://odur.let.rug.nl/~vannoord/TextCat/
[2] International Network International Center http://www.internic.net
[3] CIA The World Factbook: http://www.cia.gov/cia/publications/factbook/index.html
[4] http://www.vnexpress.net
[5] http://cambodiacic.org

recherche.

À partir des adresses « url » des sites Web pertinents en langue cible, nous continuons à collecter exhaustivement ces sites Web en utilisant un robot (par exemple avec *wget*[1], *CLIPS-Index*[2], …).

Le tableau 2.3 présente une comparaison des méthodes de récupération de documents en langue khmère. En utilisant la méthode à base de moteurs de recherche (utilisation de l'outil *CorpusBuilder* qui est adapté et amélioré pour notre expérimentation), nous obtenons seulement 113 documents pertinents en khmer en 15 jours ! Ces documents ne sont pas suffisants pour notre problème.

Méthodes de récupération	Temps de collecte total	Nombre de sites Web récupérés	Nombre de documents pertinents	Taille du corpus
Moteur de recherche	15 jours	17	113	2,3 Mo
Collecte exhaustive sur les sites Web riches en ressources	2 h	3	13 131	174 Mo

Tableau 2.3 : Récupération des documents à partir de l'Internet pour le khmer

À partir des 17 sites Web détectés par la méthode à base de moteurs de recherche, nous choisissons 3 sites les plus riches en ressources. Puis, nous utilisons l'outil *wget* pour collecter tous les documents de ces sites. Un total de 13 131 pages html peut alors être obtenu très rapidement, soit 174 Mo de données bruitées. Le temps de collecte est d'environ 2 heures seulement. Cette méthode est intéressante car nous pouvons contrôler la performance et la vitesse de récupération. Par contre, les documents proviennent d'une même source de site Web. Ils ont l'avantage (ou l'inconvénient, suivant ce que l'on veut en faire) d'être homogènes en format et en qualité. Ainsi le traitement de ces documents devient plus facile dans les étapes suivantes.

2.2. Du document html au corpus de texte : principaux problèmes

2.2.1. Conversion des encodages

Pour une langue quelconque, particulièrement pour une langue peu dotée, il existe souvent beaucoup de systèmes d'encodage différents pour coder un caractère[3] et ces systèmes d'encodage sont souvent incompatibles entre eux. Pour développer une boîte à outils multilingue, nous devons alors choisir une représentation interne unique pour les données textuelles.

Depuis quelques années, le consortium Unicode[4] a pour ambition de proposer une norme de codage de caractères la plus universelle possible. Unicode spécifie notamment un numéro

[1] http://www.gnu.org/software/wget/wget.html
[2] http://slmg-index.imag.fr/
[3] Par exemple pour le vietnamien: VNI, TCVN3, VIQR, VISCII, VPS, CP1258, … ; pour le khmer: Limon, ABC Zero Space, SEAsite, …
[4] http://www.unicode.org

unique pour chaque caractère, quelle que soit la plate-forme, quel que soit le logiciel et quelle que soit la langue. Unicode est utilisé dans de nombreux systèmes d'exploitation, dans tous les navigateurs récents, et dans de nombreux autres produits et applications. L'apparition du standard Unicode, ainsi que la disponibilité d'outils le gérant, sont parmi les faits les plus marquants de la globalisation récente du développement logiciel. L'UTF-8[1] est le format de transformation des caractères Unicode en ASCII le plus commun pour les applications liées à l'Internet. Il assure aussi une compatibilité avec les manipulations simples de chaînes en ASCII dans les langages de programmation.

Nous avons, par conséquent, choisi la norme Unicode et son format de transformation UTF-8 comme format d'encodage unique de nos données textuelles. Ainsi, une procédure de conversion de tous les systèmes d'encodage vers UTF-8 est nécessaire. Cette procédure est dépendante de chaque langue et de chaque système d'encodage considéré.

2.2.2. Segmentation en syllabes

La syllabe est considérée comme unité structurante de la langue. Généralement, la structure d'une syllabe se décompose souvent en 3 parties : l'attaque (une ou plusieurs consonnes - *facultatif*), le noyau (une voyelle ou une diphtongue - *obligatoire*) et la coda (une ou plusieurs consonnes - *facultatif*). A cause de la caractéristique facultative des consonnes sur l'attaque et sur la coda, il y a parfois des ambiguïtés de segmentation d'une phrase en syllabes.

En effet, la syllabe est reconnue comme unité fondamentale dans plusieurs applications dans le domaine du traitement automatique de la parole : reconnaissance automatique de la parole à base de syllabes [Jones 1997, Yang 1998, Nguyen 2002], identification automatique de la langue [Antoine 2004], synthèse de la parole à base de syllabes [Chen 2003], etc. L'avantage de l'utilisation de syllabes à la place des mots dans les applications de traitement automatique de la parole est que la syllabe a une meilleure couverture de la langue car elle est l'unité structurante de la langue. Par exemple, avec plus de 6500 syllabes vietnamiennes, nous pouvons couvrir tous les mots possibles en vietnamien. Ainsi, le problème de mots hors-vocabulaires est bien résolu en décomposant le mot inconnu en une série de syllabes.

La syllabe est une unité linguistique indispensable dans notre travail de traitement des langues peu dotées. Similairement au chinois, japonais et coréen, la syllabe joue un rôle morphologique fondamental pour le vietnamien. Au niveau phonétique, les syllabes vietnamiennes sont souvent prononcées séparément dans une phrase. Ainsi, dans nos expérimentations, nous utilisons les *syllabes* plutôt que les *mots* comme unités de reconnaissance automatique de la parole en vietnamien. De plus, le processus de segmentation syllabique est trivial pour le vietnamien car chaque syllabe vietnamienne est séparée par un espace.

Par contre, pour la plupart des langues, la segmentation d'une phrase en syllabes est plus difficile, car une phrase est souvent écrite en chaînes de mots, pour les langues occidentales, ou en chaînes de caractères sans espace entre les mots et les syllabes, pour des langues non

[1] Unicode Transformation Format: http://www.ietf.org/rfc/rfc2279.txt

segmentées. Des travaux sur la segmentation en syllabes ont vu le jour, comme des méthodes de syllabation pour l'anglais [Kahn 1976] et pour le français [Boula-de-Mareüil 1997].

V. Berment, dans le cadre de sa thèse [Berment 2004], a construit un outil nommé « Sylla » permettant de mettre au point rapidement des « modèles syllabiques » pour une langue peu dotée. Il a appliqué cet outil pour construire des modèles grammaticaux des syllabes des langues d'Asie du Sud-est : laotien, birman, thaï et khmer. L'outil et la méthode de construction d'un modèle syllabique permet de créer rapidement un « reconnaisseur syllabique » : pour une chaîne de caractères en entrée, le reconnaisseur teste si la chaîne peut constituer une syllabe dans la langue considérée.

Pour la segmentation en syllabes, un segmenteur syllabique sera construit en employant un algorithme de programmation dynamique, à l'aide d'un modèle syllabique, qui segmente une phase de texte en optimisant le critère de « plus longue chaîne d'abord » (*Longest Matching*), ou le critère de « plus petit nombre de syllabes » (*Maximal Matching*).

Avec l'aide de V. Berment, nous avons généré un modèle syllabique pour le khmer. Ensuite, à partir de ce modèle syllabique, nous avons construit un outil de segmentation syllabique en utilisant la méthode de segmentation « plus petit nombre de syllabes ». La figure 2.8 illustre un exemple de segmentation d'une phrase khmère en syllabes effectuée par l'outil de segmentation.

Phrase khmère originale :	នេះជាចំនួនខ្ពស់បំផុត។
Phrase segmentée syllabique :	នេះ ជា ចំ នួន ខ្ពស់ បំ ផុត ។
Prononciation :	/ nih ciɜ cam nuɜn kʰpas bam pʰot /

Figure 2.8 : Exemple de segmentation automatique en syllabes d'une phrase khmère

Nous avons évalué la performance du segmenteur syllabique sur un corpus de 47 phrases khmères non-segmentées, soit 621 syllabes. Pour la référence, le corpus de texte est segmenté manuellement en syllabes par un étudiant cambodgien. Le taux de syllabes segmentés correctement est de **81,5%**.

2.2.3. Segmentation en mots

Il existe beaucoup de systèmes d'écritures non segmentées en mots comme le chinois, le japonais, le khmer, le laotien, le thaï, le vietnamien, etc. Pour ces langues, une phrase est écrite en chaînes de caractères sans espace entre les mots, ou alors l'espace n'est pas utilisé pour déterminer la frontière de mots (par exemple le vietnamien pour lequel l'espace délimite la frontière de syllabes). Ainsi, la procédure de segmentation d'une phrase en mots peut être très difficile. Le traitement d'un corpus de texte nécessite de résoudre ce problème, notamment pour les langues peu dotées que nous avons traité, qui sont des langues non segmentées.

Les méthodes de segmentation en mots à base d'un vocabulaire sont utilisées largement pour des langues asiatiques non-segmentées [Promchan 1998, Nie 2000]. Ces méthodes

recherchent dans un vocabulaire les mots correspondant à ceux du texte et, en cas d'ambiguïté, sélectionnent les mots qui optimisent un paramètre dépendant de la stratégie choisie [Berment 2004]. Une autre méthode de segmentation à base d'automates d'états finis est appliquée sur un corpus de texte vietnamien [Nguyen 2004].

Par ailleurs, il existe des solutions plus efficaces qui utilisent des méthodes statistiques et/ou passent par une phase d'apprentissage. Pour une phrase chinoise à segmenter, A. Wu construit un treillis de mots possibles en fonction d'un vocabulaire [Wu 2003]. Ensuite, il applique des méthodes statistiques pour décoder le chemin le plus probable sur le treillis. Une méthode statistique et linguistique de segmentation en mots est aussi proposée et implémentée sur la langue thaïe [Meknavin 1997]. Dans cette méthode, l'environnement des mots est analysé linguistiquement pour déterminer la segmentation la plus probable. Une autre méthode appliquée au vietnamien consiste à utiliser un algorithme d'apprentissage à base d'un réseau de neurones [Dinh 2001]. Cependant, les méthodes statistiques nécessitent de disposer d'un grand corpus de texte segmenté au préalable.

Les méthodes statistiques et les méthodes d'apprentissage complexes ne sont pas appropriées dans le cadre de notre travail car les ressources nécessaires pour implémenter ces méthodes n'existent pas. Pour une langue peu dotée considérée, nous cherchons des méthodes de segmentation plus performantes, plus rapides et plus faciles à implémenter qui tirent, au mieux, bénéfice des ressources limitées existantes pour la langue. Pour illustrer les performances de nos méthodes, nous les appliquons sur la langue khmère qui possède, comme le thaï, une écriture non-segmentée.

a) Méthodes à base d'un vocabulaire

Dans le but d'implémenter un outil simple de segmentation en mots, utilisable facilement dans plusieurs de langues, nous avons appliqué la méthode à base d'un vocabulaire de mots. Cette méthode utilise un algorithme de programmation dynamique qui recherche dans un vocabulaire les mots du texte et qui le segmente en optimisant un critère quelconque. Ce critère est choisi en fonction de la stratégie de segmentation « plus longue chaîne d'abord » (*Longest Matching*) ou « plus petit nombre de mots » (*Maximal Matching*).

Pour évaluer la performance de segmentation en mots, nous testons sur 2 corpus de textes en langue khmère, qui se distinguent par le pourcentage de mots hors-vocabulaire (MHV). Le tableau 2.4 présente la performance des méthodes de segmentation selon le taux de mots hors-vocabulaire. Nous constatons que la méthode « plus petit nombre de mots » (ou *Maximal Matching*) est meilleure. Quand le taux de mots hors-vocabulaires est petit, ces méthodes marchent très bien. Cependant, elles ont des performances médiocres en présence de cas d'ambiguïté et de mots hors-vocabulaires. Pour résoudre le problème d'ambiguïté, nous proposons d'améliorer le segmenteur à base d'un vocabulaire en intégrant les fréquences d'apparition des mots dans le vocabulaire (ou un modèle de langage unigramme). En cas de présence de mots hors-vocabulaire, nous pouvons combiner cette méthode de segmentation en mots avec un segmenteur syllabique présenté dans la section 3.2.2 pour détecter les formes lexicales (comme cela est fait sur le thaï [Kosawat 2003]).

Corpus de test	Nombre de phrases/ Nombre de mots	Taux de mots hors-vocab.	Méthode de segmentation	
			Longest Matching WCR (SCR)	*Maximal Matching* WCR (SCR)
KH-1	450 / 5735	0%	98,9% (94,7%)	**99,2% (96,0%)**
KH-2	520 / 6660	1,6%	98,1% (88,7%)	**98,4% (89,8%)**

Tableau 2.4 : Performance du segmenteur selon le taux de mots hors-vocabulaire (WCR : taux de mots segmentés corrects ; SCR : taux de phrases correctes)

b) Utilisation d'un modèle de langage unigramme

Les méthodes à base d'un vocabulaire utilisent un algorithme de programmation dynamique qui recherche dans un vocabulaire les mots du texte et qui le segmente en optimisant une fonction de coût quelconque. Si l'optimisation de la fonction de coût est la minimisation du nombre de mots, nous avons la méthode « plus petit nombre de mots ».

Dans cette section, nous essayons d'améliorer la fonction de coût en ajoutant la probabilité d'apparition de la chaîne de mots que nous venons de segmenter. C'est-à-dire, la segmentation de syllabes repose sur un modèle statistique du langage *n*-grammes. Pour créer un modèle de langage *n*-grammes de bonne qualité, nous avons besoin d'un grand corpus de texte segmenté. Puisqu'un tel corpus n'existe pas dans notre travail, nous décidons d'utiliser un petit corpus de texte segmenté pour créer simplement un modèle de langage unigramme.

Nous rappelons qu'en utilisant un modèle statistique du langage unigramme, pour une chaîne W segmentée en N mots w_1, w_2, …, w_N, la probabilité d'apparition de la chaîne est calculée par l'expression suivante :

$$P(W) = P(w_1, w_2, \ldots, w_N) = \prod_{i=1}^{N} P(w_i) \tag{2.1}$$

avec $P(w_i)$ la probabilité d'apparition du mots w_i dans le corpus d'apprentissage.

Soit $C(w_i)$ le nombre de fois où le mot w_i a été observé dans le corpus d'apprentissage et C le nombre total des mots dans le corpus, nous avons :

$$P(w_i) = \frac{C(w_i)}{C} \tag{2.2}$$

Nous constatons que, dans le cas de présence de mots inconnus dans la chaîne W, nous obtiendrons une probabilité nulle. Alors, la fonction de coût à base de modèle de langage devient inutile. Pour éviter cela, il est nécessaire de trouver un calcul permettant l'obtention d'une probabilité pour un mot inconnu. Nous pouvons, par exemple, utiliser la méthode de Good-Turing pour le lissage avec le repli de Katz [Katz 1987]. Dans notre expérimentation, nous corrigeons le $C(w_i)$ comme suit :

$$C(w_i) = \begin{cases} C(w_i) & \text{si } C(w_i) > 0 \\ \varepsilon & \text{ailleurs} \end{cases} \quad (2.3)$$

avec ε un nombre petit quelconque et $0 < \varepsilon < 1$.

Alors, une meilleure segmentation est celle qui maximise la probabilité d'apparition de la chaîne W de l'expression (2.3).

Nous avons comparé des méthodes de segmentation utilisées dans notre travail sur la langue khmère en calculant le taux de mots segmentés corrects (la figure 2.9) et le taux de phrases segmentées correctes (le tableau 2.5).

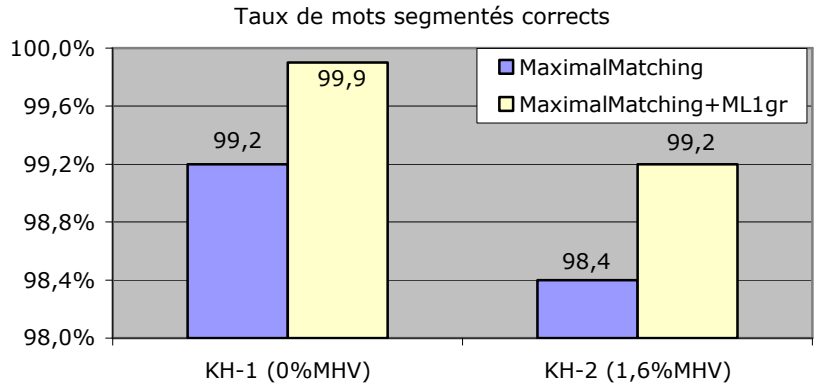


Figure 2.9 : Taux de mots segmentés corrects selon la méthode de segmentation

Corpus de test	**Taux de phrases segmentées correctes**		
	Maximal Matching	*Maximal Matching + Modèle unigramme*	*Gain absolu*
KH-1	96,0%	**99,8%**	3,96%
KH-2	89,8%	**94,2%**	4,90%

Tableau 2.5 : Taux de phrases segmentées correctes selon la méthode de segmentation

Nous constatons que la méthode à base d'un modèle de langage est meilleure que la méthode à base d'un vocabulaire. Par exemple, sur le corpus KH-1 qui ne contient pas de mots hors-vocabulaire, la méthode à base d'un modèle de langage unigramme améliore de 0,71% de mots segmentés corrects et de 3,96% le taux de phrases segmentées correctes.

3. *CLIPS-Text-Tk* – Boîte à outils générique

Le but de notre travail est de construire une méthodologie de construction rapide d'outils de récupération et de traitement d'un corpus de texte. Ce corpus sert non seulement pour la construction d'un système de reconnaissance automatique de la parole, mais éventuellement pour plusieurs autres applications dans le domaine du traitement de la langue (statistique

linguistique, recherche d'information, traduction automatique, …).

Afin de rendre les données recueillies sur le Web exploitables, un certain nombre de traitements sont nécessaires. Nous les avons répartis comme suit :

1. transformation html vers texte ;
2. normalisation des tags et restructuration des documents ;
3. conversion des encodages ;
4. séparation en phrases ;
5. séparation en mots ;
6. transcription des caractères spéciaux ;
7. transcription des nombres ;
8. conversion de la casse du caractère ;
9. suppression de la ponctuation ;
10. filtrage en fonction d'un vocabulaire donné.

Ces outils de récupération et de traitement d'un corpus de texte peuvent être construits spécifiquement pour chaque langue ou chaque tâche mais ils sont coûteux en temps de développement.

En effet, lors du traitement multilingue de corpus de textes, nous avons pu remarquer que certains traitements peuvent être considérés comme relativement indépendants de la langue (1-2-6-8-9-10), bien qu'il reste encore des traitements spécifiques dépendant de la langue cible (3-4-5-7) qui doivent être repensés pour chaque nouvelle langue cible.

Dans un but de généricité, la construction de composants réutilisables pour plusieurs langues et tâches spécifiques est très importante. Pour que les outils puissent être utilisés dans plusieurs langues et tâches différentes, nous construisons une boîte à outils générique « multilingue » qui contient des outils de traitement recyclables et adaptables.

La figure 2.10 présente l'architecture de la boîte à outils générique multilingue « *CLIPS-Text-Tk* » développée dans notre travail. Après avoir déterminé tous les problèmes et traitements nécessaires pour l'obtention et le traitement de données textuelles, nous avons décidé de décomposer ces traitements en un ensemble de petits modules. Ensuite, ces petits modules sont répartis en deux groupes :

- **les modules fixes**, qui travaillent indépendamment de la langue ;
- **les modules variables**, qui sont dépendants de la langue.

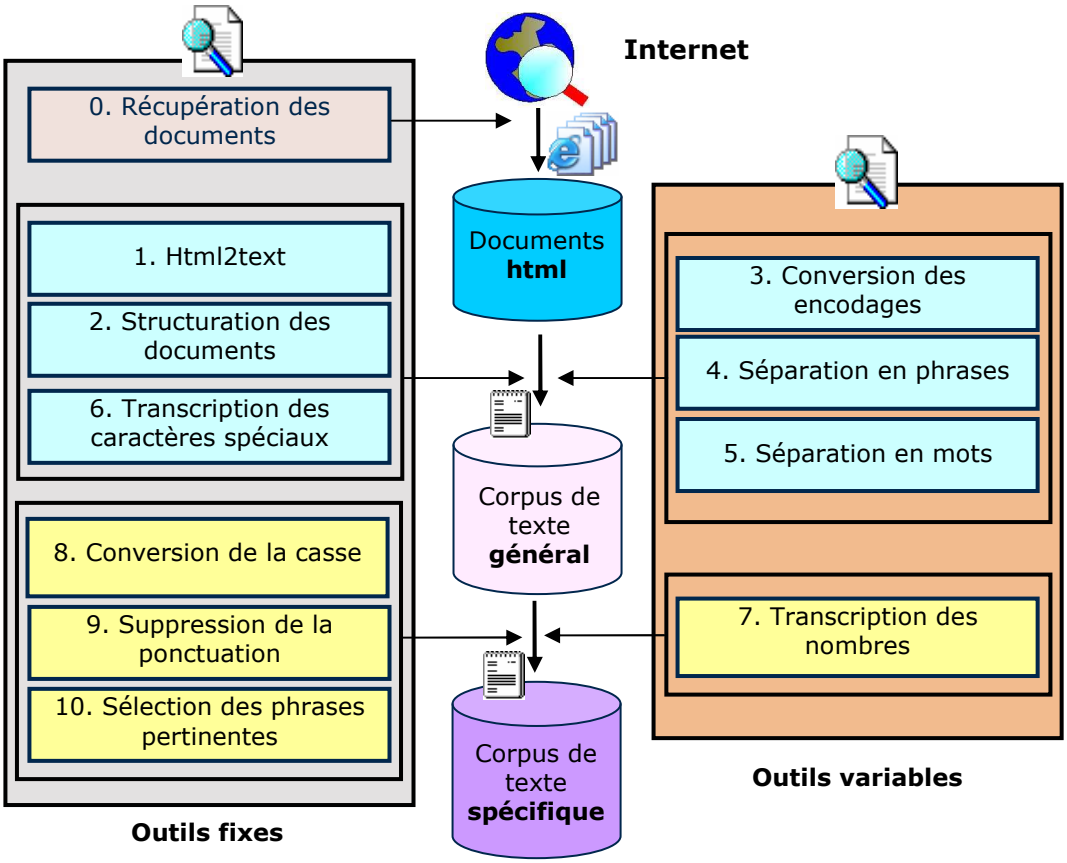


Figure 2.10 : Architecture de la boîte à outils générique « multilingue »

Ainsi, pour une nouvelle langue, nous hériterons de tous les **modules fixes** et adapterons rapidement les **modules variables** à cette langue. Cela économisera le temps de construction d'une boîte à outils complète.

De la même façon, pour une tâche spécifique (modélisation statistique du langage, recherche d'information, traduction automatique, …), nous pouvons d'une part hériter des outils de traitement, et d'autre part adapter rapidement les autres outils spécifiquement à la tâche. Pour choisir quels outils sont appropriés pour une langue ou une tâche spécifique, nous pouvons essayer chaque outil ou combiner consécutivement des outils dans la boîte à outils pour obtenir un meilleur traitement, comme cela est fait dans le travail de G. Adda sur la normalisation de textes en français pour la tâche de modélisation statistique du langage [Adda 1997].

En résumé, notre méthodologie de recueil rapide d'un corpus de texte pour une langue peu dotée ou une tâche spécifique consiste à construire une boîte à outils générique « modulaire » et « multilingue ». Cette boîte est facilement portable d'une langue à l'autre et facilement portable d'une tâche ou d'un domaine à l'autre. Pour faire cela, les outils de traitements dans la boîte à outils doivent respecter les critères suivants :

- ils sont standardisés ;

- ils sont « open source », compréhensibles et légers ;

- ils sont exécutables rapidement et stables.

Ces outils sont développés sous Linux sous forme de langages scripts.

4. Modélisation statistique du langage

4.1. Modélisation statistique du langage à partir de l'Internet

Le traitement automatique des langues fait de plus en plus appel à de volumineux corpus pour l'acquisition des connaissances. Un corpus est une collection de données textuelles qui a été collectée pour servir d'échantillon représentatif de la langue. Un corpus de référence est conçu pour fournir une information générale de la langue. Un corpus spécialisé se restreint à une situation particulière, comme dans le cas d'un domaine scientifique ou technique. Dans tous les cas, le corpus doit avoir une taille suffisamment grande pour représenter les variétés de la langue, ou du domaine.

L'apprentissage des modèles statistiques du langage de la plupart des systèmes de reconnaissance automatique de la parole est réalisé principalement avec deux types de corpus :

- des données journalistiques ;

- des transcriptions manuelles de données audio directement liées à l'application visée.

Cependant, comme nous l'avons déjà précisé, de telles données ne sont pas disponibles et/ou sont très coûteuses à collecter pour des langues peu dotées. Récemment, une autre source de données s'est ajoutée aux deux précédentes. Le World Wide Web est, en effet, devenu une des plus importantes sources d'information disponible de manière électronique. Le Web constitue ainsi un corpus gratuit, riche, énorme et accessible pour de nombreuses langues.

4.2. Construction d'un corpus de texte pour la modélisation du langage

Tout d'abord, nous utilisons la méthodologie et la boîte à outils générique *CLIPS-Text-Tk* proposée dans la section précédente pour construire un corpus de texte à partir de l'Internet.

A titre d'exemple, la quantité de pages Web que nous avons collectées en 2003 à partir d'un site de courrier vietnamien était de 2,5 Go. Après avoir traité le corpus de texte, la quantité de données textuelles pouvant servir à l'apprentissage d'un modèle de langage statistique était d'environ 400 Mo (5 millions phrases) ce qui commence à être suffisant pour apprendre un modèle de langage statistique en vietnamien. A titre de comparaison, une année complète du journal *Le Monde* en français correspond à 120 Mo en moyenne. La dernière étape est celle du filtrage, en fonction du vocabulaire, des phrases et aussi de blocs de mots. Les modèles de langage sont appris à partir de ce corpus de texte. Il est donc important de disposer de bonnes méthodes de filtrage. Pour améliorer la qualité du corpus de texte d'apprentissage, nous proposons dans la section suivante quelques méthodes de filtrage de documents et de phrases et testons leur efficacité dans la modélisation statistique du langage.

4.2.1. Filtrage de phrases du corpus de texte

Pour construire des modèles de langage à partir d'un corpus de texte, nous essayons dans

notre travail des méthodes de filtrage de phrases à base d'un vocabulaire avec des paramètres d'utilisateur différents :

- **filtrage 1 :** prendre toutes les phrases (ne pas appliquer de filtrage) ;
- **filtrage 2 :** prendre les phrases ayant au moins *N* mots et dont tous les mots appartiennent au vocabulaire ;
- **filtrage 3 :** prendre une séquence consécutive (un bloc) d'au moins *M* mots appartenant au vocabulaire. C'est la méthode des blocs minimaux qui a été proposée dans [Vaufreydaz 2002] ;
- **filtrage 4 :** utiliser une méthode hybride qui consiste à prendre les phrases entières ayant au moins *N* mots appartenant au vocabulaire (filtrage 2) et appliquer le filtrage par blocs minimaux de taille *M* (filtrage 3) sur les phrases rejetées.

Pour illustrer l'efficacité de chaque filtrage dans la tâche de modélisation statistique du langage, nous appliquons les filtrages sur les corpus de texte d'apprentissage du français (FR), vietnamien (VN) et khmer (KH). Pour estimer la perplexité (PPL) des modèles de langage, nous testons les modèles de langage du français et vietnamien sur un corpus de 216 phrases de dialogues courts extraites à partir du projet NESPOLE[1]. Le corpus de test du vietnamien est la traduction du corpus de test français. Le corpus de test du khmer contient 200 phrases de texte. Les taux de mots hors-vocabulaire des corpus de test sont 2% pour le français, 0,6% pour le vietnamien et 0,2% pour le khmer. Pour apprendre les modèles de langage, nous utilisons la boîte à outils SRILM [Stolcke 2002] en utilisant la méthode de Good-Turing pour le lissage avec le repli de Katz [Katz 1987].

Le tableau 2.6 présente la comparaison de la taille des corpus de texte obtenus et la perplexité des modèles de langage trigrammes selon la méthode de filtrage de phrase pour le français, vietnamien et khmer (corpus provenant de l'Internet). Les résultats montrent que le filtrage 2 est le meilleur dans le cas du vietnamien et khmer mais le filtrage 4 est le meilleur dans le cas du français.

Filtrage de phrase	FR : Web (Taille vocab. 20 000)		VN : Web (Taille vocab. 20 000)		KH : Web (Taille vocab. 7 000)	
	Taille (Mo)	PPL	Taille (Mo)	PPL	Taille (Mo)	PPL
Filtrage 1	686	539	868	260	95	88
Filtrage 2 (*N*=1)	156	580	370	**252**	43	**84**
Filtrage 3 (*M*=3)	366	637	667	359	81	88
Filtrage 4 (*M*=3, *N*=1)	411	**509**	729	259	85	87

Tableau 2.6 : Perplexité des modèles de langage selon la méthode de filtrage de phrase

[1] http://nespole.itc.it/

4.2.2. *Filtrage des informations redondantes*

Nous rappelons que pour contrôler la vitesse de récupération et la qualité des documents du Web pour des langues peu dotées, les documents sont récupérés exhaustivement à partir d'un ou quelques sites Web. Les documents qui proviennent d'une même source ont l'avantage d'être homogènes en format et en qualité. On préfèrera dans notre travail des sites de nouvelles, au fort contenu rédactionnel.

Cependant, nous constatons que les documents Web peuvent contenir beaucoup d'informations telles que menus, références, annonces, publicités, etc. qui sont répétées dans plusieurs pages. À titre d'exemple, la figure 2.11 illustre une page Web récupérée à partir d'un site de nouvelles vietnamiennes. Si nous appliquons directement l'outil de conversion html vers texte sur cette page, le corps du document (l'information pertinente) est extrait, accompagnant des informations textuelles redondantes comme menu, publicités, liens,… Ces informations sont répétées dans plusieurs pages de ce site et elles ont une influence sur la qualité du corpus de texte collecté.

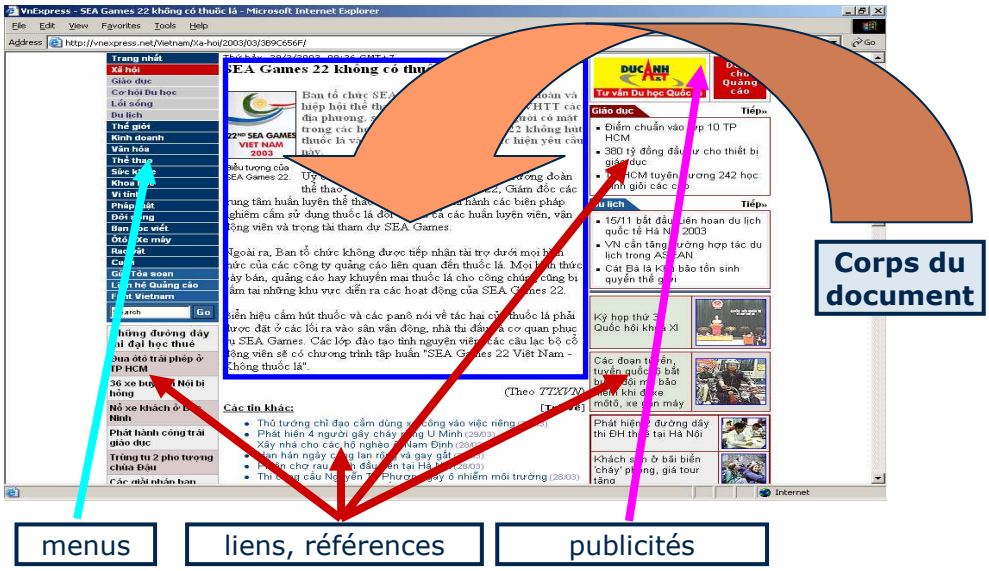


Figure 2.11 : Filtrage des informations redondantes dans les documents html

Nous avons estimé l'influence de ces informations redondantes sur des documents html collectés d'un site Web de nouvelles en vietnamien. Deux expérimentations sont comparées :

1. appliquer directement l'outil de conversion html vers texte sur tous les documents du site Web ;
2. utiliser un filtrage de l'information redondante avant l'application de l'outil de conversion html vers texte.

Le tableau 2.7 présente une comparaison de la taille des corpus de texte obtenus et la perplexité des modèles de langage trigrammes appris sur ces corpus dans les deux expérimentations. Nous constatons qu'en filtrant les informations redondantes, la taille du

corpus de texte est réduite de 50% environ dans notre expérimentation. Par ailleurs, la perplexité des modèles de langage construits sur ce corpus de texte est améliorée d'environ 22% en supprimant les informations redondantes [Le 2003b].

Filtrage de phrase	Conversion html vers texte		Conversion html vers texte avec un filtrage de redondance		Gain de perplexité
	Taille (Mo)	PPL	Taille (Mo)	PPL	
Filtrage 1	868	260	402	201	22,7%
Filtrage 2 (*N*=1)	370	**252**	226	**195**	22,6%
Filtrage 3 (*M*=3)	667	359	357	282	21,4%
Filtrage 4 (*M*=3, *N*=1)	729	259	373	199	23,2%

Tableau 2.7 : Influence des informations redondantes sur les modèles de langage

Par conséquent, pour obtenir un corpus de texte et des modèles de langage de meilleure qualité, les informations redondantes contenues dans les documents html doivent être filtrées et enlevées.

5. Conclusions du chapitre

Ce chapitre a présenté le recueil d'un vocabulaire de mots ou syllabes pour des langues peu dotées. A cause de la faible qualité de la source de données textuelles et la difficulté de segmentation en mots pour des langues non-segmentées, la technique de construction automatique d'un vocabulaire à partir d'un corpus de texte s'avère difficile. Notre approche consiste tout d'abord à récupérer un vocabulaire à partir de ressources lexicales existantes. A son tour, ce vocabulaire sert à segmenter le corpus de texte. Enfin, le vocabulaire peut être enrichi et limité en fonction du nombre d'occurrences de mots dans le corpus de texte segmenté.

En plus des traitements spécifiques sur des langues peu dotées comme les méthodes de récupération rapide des documents à partir de l'Internet, les méthodes de segmentation en syllabes et mots, nous proposons une méthode générique de traitement du corpus de textes. Une boîte à outils générique « multilingue » contenant des outils de traitement recyclables et adaptables est développée dans notre travail. Pour une nouvelle langue, nous hériterons de tous les outils indépendants de la langue et adapterons rapidement les outils dépendants de la langue. Cela économisera le temps de développement.

Pour la modélisation statistique du langage, avec la suppression des informations redondantes contenues dans les documents html, une réduction d'environ 50% sur la taille du corpus d'apprentissage et d'environ 22% sur la perplexité des modèles de langage est obtenue. Par ailleurs, en appliquant quelques méthodes de filtrage de phrases sur le corpus d'apprentissage, les résultats expérimentaux laissent supposer que, la méthode 2 (prendre les phrases entières des mots inclus dans un vocabulaire prédéfini) et la méthode 4 (prendre les phrases entières ou les blocs de mots inclus dans un vocabulaire prédéfini) sont plus adaptées

que la méthode 1 (ne pas appliquer de filtrages) et la méthode 3 (prendre seulement les blocs minimaux). De plus, ces méthodes de filtrage de phrases sont aussi intégrées dans la boîte à outils *CLIPS-Text-Tk* afin qu'elles soient utilisées dans d'autres tâches telles que la génération des énoncés pour l'enregistrement d'un corpus vocal ou l'obtention d'un corpus indépendant du vocabulaire.

Les résultats expérimentaux de la méthodologie présentée dans ce chapitre, pour la reconnaissance automatique du vietnamien et du khmer, seront présentés aux chapitres 4 et 5 de ce manuscrit.

Chapitre 3

Construction rapide de modèles acoustiques

1. Méthodologie

Le développement d'un système de reconnaissance automatique de la parole continue à grand vocabulaire dans une nouvelle langue nécessite de rassembler une grande quantité de corpus de parole, contenant des signaux de parole pour l'apprentissage des modèles acoustiques du système. De tels corpus et systèmes sont désormais disponibles pour la plupart des langues occidentales comme l'anglais, le français, l'espagnol, et pour quelques langues asiatiques comme le chinois, le japonais, le coréen ainsi que pour l'arabe.

Porter des modèles acoustiques vers une nouvelle langue est cependant une tâche très fastidieuse si aucun corpus de grande envergure n'existe dans la langue cible, puisqu'il faut alors collecter soi-même les ressources nécessaires : signal de parole, dictionnaire de prononciation, etc ou trouver des techniques permettant de se passer de ces grandes quantités de corpus. Précisons aussi qu'étant donné la nature statistique des modèles généralement utilisés en reconnaissance automatique de la parole (modèles acoustiques de phonèmes correspondant à des chaînes de Markov où chaque état est une distribution multigaussienne) les corpus collectés doivent être conséquents.

Dans le domaine de la portabilité de modèles acoustiques existants vers une nouvelle langue, nous définissons et distinguons ici les concepts de *langue source* et *langue cible* qui seront souvent utilisés dans ce chapitre. Pour la portabilité de modèles acoustiques, nous utilisons toujours un système de reconnaissance automatique de la parole existant dans une langue (monolingue) ou plusieurs langues source (multilingues). Ces modèles acoustiques initiaux sont construits et appris à partir d'une grande quantité de corpus de parole (d'une vingtaine à une centaine d'heures de signaux). Un tel système est appelé système source, la langue du système est appelée *langue source*. Par contre, dans notre contexte des langues peu dotées, la *langue cible* est une langue qui ne possède pas ou peu de signaux vocaux d'apprentissage. Alors, pour construire un système de reconnaissance automatique de la parole, nous devons appliquer des méthodes de portabilité et d'adaptation des modèles acoustiques existants en langue source vers cette langue cible.

Une première façon d'accélérer la portabilité des systèmes de reconnaissance automatique de la parole continue grand vocabulaire en langue source vers une langue peu dotée (langue

cible), est de développer une méthodologie permettant une collecte rapide et/ou facilitée de ressources textuelles (abordée sur la chapitre 2) et acoustiques, contenant à la fois des signaux de parole (voir section 2) de taille limitée et une dictionnaire de prononciation (voir section 3).

Dans la plus grosse partie de ce chapitre, nous proposons quelques méthodes de portabilité et d'adaptation rapide des modèles acoustiques *indépendant* et *dépendant* du contexte vers des langues peu dotées (voir sections 4 et 5). Une telle méthode est intéressante notamment lorsqu'on passe d'une langue source à une langue cible qui possèdent un système phonologique proche. Pour la portabilité, nous réutilisons des modèles des unités phonétiques existants en langue source qui sont les plus proches d'unités phonétiques correspondantes en langue cible. Pour cela, nous proposons notamment dans la section 4, des concepts et des méthodes d'estimation de similarités des unités phonémiques (phonème, polyphone, groupe de polyphones, …) à base de distances, en particulier des similarités phonétiques entre langue source et langue cible.

D'autre part, pour les langues peu dotées qui ne possèdent pas encore un dictionnaire phonétique parce que leurs systèmes phonologiques sont parfois méconnus, l'approche de modélisation acoustique à base de graphèmes présente un potentiel intéressant pour construire rapidement un système de reconnaissance automatique de la parole. Dans la section 6, nous présentons notre travail sur la modélisation acoustique à base de graphèmes pour des langues peu dotées. Nous proposons aussi une méthode générique permettant d'initialiser plus rapidement et plus efficacement des modèles acoustiques graphémiques.

2. Prototype d'acquisition d'un corpus de parole

Pour le recueil de signaux de parole, le CLIPS a développé un outil logiciel ne mettant en œuvre que du matériel standard : EMACOP (Environnement Multimédia pour l'Acquisition et la gestion de Corpus Parole) [Vaufreydaz 1998]. La plupart du temps, les campagnes d'enregistrement mobilisent d'importantes ressources humaines pour guider ou assister les locuteurs dans leur tâche de diction, pour organiser l'enregistrement, pour préparer les scénarios et les données, etc. Il faut pouvoir contrôler les différents scénarios pour varier les conditions de capture : la lecture d'un texte ou d'une suite de mots ou de mots isolés, la répétition après écoute d'une phrase, le dialogue en réponse à des questions, etc. Les méthodes d'acquisition rigoureusement contrôlées sont donc lourdes et les difficultés sont amplifiées dans le cas des langues peu dotées où les locuteurs ne sont pas forcément habitués à l'utilisation de moyens informatiques par exemple. C'est pourquoi, le développement d'un utilitaire portable de gestion et d'acquisition de grands corpus sur un matériel standard, nous est d'un grand bénéfice. Le logiciel respecte le format SAM de définition de bases de signaux et il permet l'acquisition, en mode client-serveur, de plusieurs locuteurs en même temps.

Cependant, la plateforme du logiciel EMACOP était monolingue au début de cette thèse. Ainsi l'interface EMACOP a dû être adaptée pour manipuler respectivement les caractères vietnamiens et khmers.

Pour cela nous avons utilisé le système d'encodage des caractères Unicode UTF-8 et

développé ainsi une nouvelle version d'EMACOP : EMACOP-Unicode. Cette version est actuellement utilisée au centre MICA (Vietnam) pour le vietnamien et à l'Institut ITC (Cambodge) pour le khmer. La figure 3.1 illustre l'interface du logiciel EMACOP-Unicode, installée sur la machine dédiée de l'ITC.

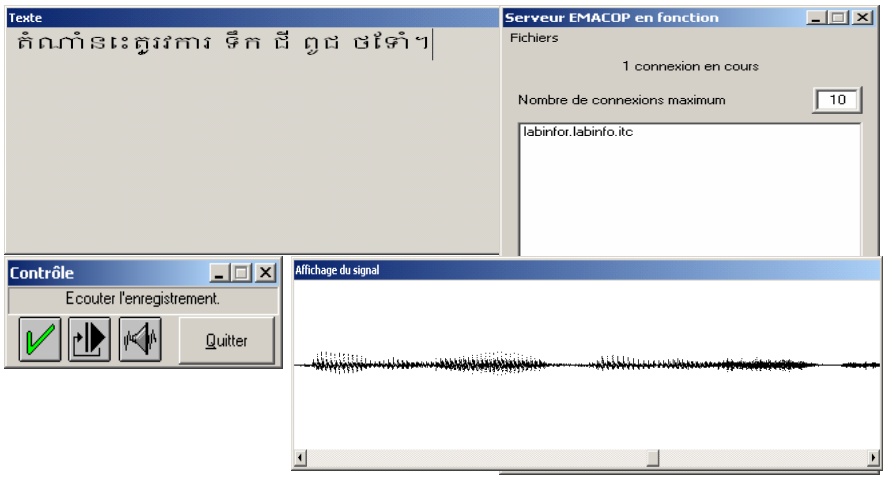


Figure 3.1 : Interface EMACOP utilisant l'encodage Unicode UTF-8

3. Construction du dictionnaire de prononciation

Un dictionnaire de prononciation est une ressource essentielle aux tâches de synthèse et de reconnaissance automatique de la parole, ou tout simplement pour enrichir un dictionnaire bilingue, permettant au locuteur étranger de connaître la prononciation du mot en langue cible. Cette tâche est cependant difficile pour des langues peu dotées dont le système phonologique est parfois méconnu, ou sujet à débats (langues peu ou mal décrites). Si nous mettons de côté les méthodes manuelles de phonétisation qui, bien que donnant les dictionnaires de prononciation de meilleure qualité, ne nous semblent pas entrer dans le cadre de notre méthodologie, on peut distinguer trois types d'approches automatiques pour constituer un dictionnaire phonétique dans une nouvelle langue.

3.1. Approches à base de règles

Habituellement, pour la modélisation acoustique, nous choisissons le phonème comme unité de modélisation [Calliope 1989, Boite 2000, Huang 2001]. Nous devons construire un dictionnaire phonétique contenant une représentation de chaque entrée lexicale d'un vocabulaire par une suite de phonèmes. Cette construction nécessite une bonne connaissance de la langue et de ses règles de phonétisation, qui par ailleurs ne doivent pas contenir trop d'exceptions. Cependant, ce type d'approche est assez coûteux en temps (écriture d'un analyseur phonétique), mais donnera des dictionnaires de prononciation de qualité très correcte pouvant ensuite être révisés manuellement relativement rapidement.

3.2. Approches utilisant un décodeur acoustico-phonétique

Certaines approches utilisent un système de reconnaissance phonémique appliqué sur des enregistrements des mots à phonétiser, permettant un premier étiquetage automatique en phonèmes d'une liste de mots, qui peut être alors révisé par un opérateur humain.

T. Sloboda utilise cette approche pour améliorer un dictionnaire phonétique existant pour une langue bien dotée [Sloboda 1995, 1996]. En utilisant un décodeur acoustico-phonétique sur des enregistrements de mots du vocabulaire à phonétiser, les *N*-meilleures hypothèses phonémiques sont extraites. Les nouvelles variantes de prononciation générées automatiquement sont alors ajoutées au dictionnaire phonétique. Avec cette méthode, le taux d'erreur de reconnaissance du système est réduit jusqu'à 8% en absolu.

Une telle méthode est potentiellement intéressante dans notre cas, surtout si l'on passe d'une langue source à une langue cible qui possèdent un système phonologique proche. L'avantage d'une telle approche est bien sûr sa rapidité. Ses inconvénients sont qu'elle nécessite l'emploi d'un système de reconnaissance automatique de phonèmes qui sera généralement celui d'une langue source bien dotée ; par ailleurs, l'autre défaut est que les unités phonémiques décrivant les mots en langue cible seront seulement celles pouvant être reconnues par le décodeur en langue source, d'où la nécessité d'employer si possible des décodeurs phonémiques multilingues pour augmenter au maximum la couverture phonémique dans l'alphabet phonétique international (API) [IPA 1999].

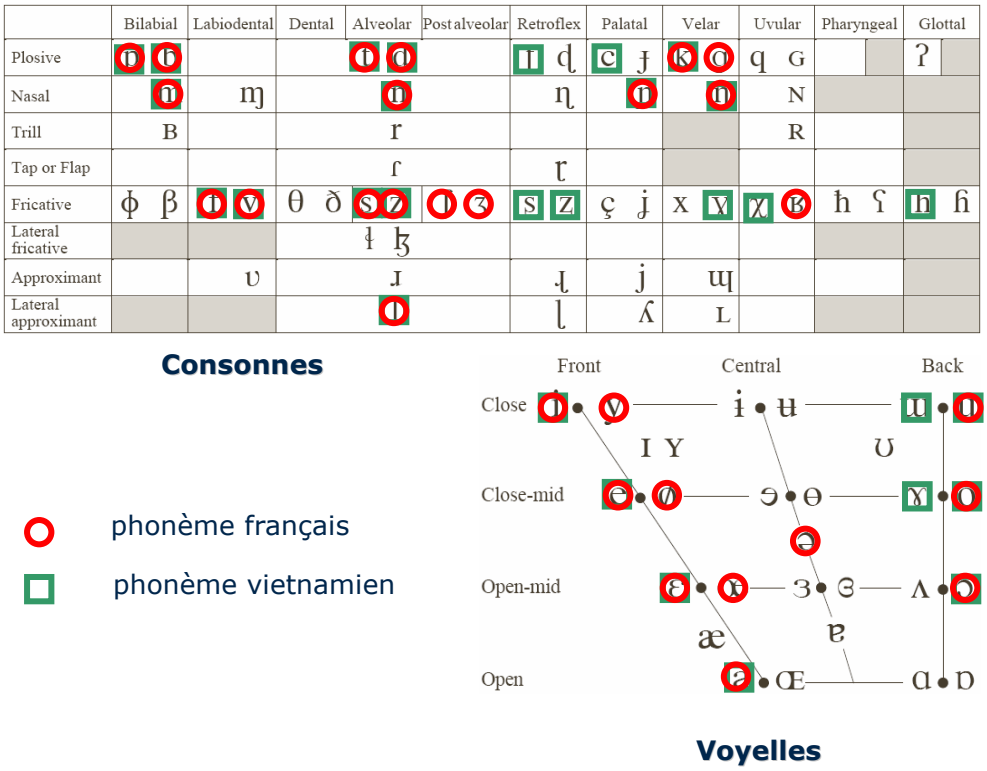


Figure 3.2 : Couverture phonémique du français et du vietnamien sur l'API

La figure 3.2 illustre ce problème avec pour la langue source le français et langue cible le vietnamien. La couverture du vietnamien par le français étant de 63%, il est évident qu'elle n'est pas optimale.

S. Stüker a proposé une méthode de génération d'un dictionnaire phonétique pour la langue suédoise et créole en votant parmi les décodeurs acoustico-phonétique monolingue dans 9 langues (8 monolingues et 1 multilingue - MM7) [Stuker 2002]. Pour chaque enregistrement de mot du vocabulaire à phonétiser, les hypothèses phonémiques sont décodées séparément. En votant parmi ces hypothèses, la prononciation du mot est choisie. Enfin, un système de reconnaissance automatique de la parole est construit et validé à partir du dictionnaire phonétique automatiquement généré. Cependant, les résultats de reconnaissance (taux d'erreur très élevés) montrent qu'un dictionnaire phonétique obtenu de la sorte ne peut être utilisé directement sans révision par un opérateur humain.

En conclusion, la méthode de génération automatique d'un dictionnaire phonétique pour une langue peu dotée à base d'un décodeur acoustico-phonétique n'est applicable qu'avec un décodeur de bonne qualité, une bonne couverture phonétique source/cible et une révision manuelle *a posteriori*.

3.3. Dictionnaire de prononciation à base de graphèmes

Comme nous l'avons déjà précisé, les systèmes récents de reconnaissance automatique de la parole continue à grand vocabulaire utilisent des unités de « sous-mots » *(sub-word unit)* comme unités de modélisation acoustique [Singh 2002]. Un type d'unité choisi fréquemment est le *phonème* (monophones ou polyphones). La performance de la modélisation acoustique à base de phonèmes dépend de la qualité du dictionnaire de prononciation (dictionnaire phonétique) qui est construit manuellement ou automatiquement.

Cependant, pour une langue peu dotée, il n'est pas facile de construire un tel dictionnaire de prononciation car on ne dispose parfois que de peu d'informations sur la langue cible. Une question se pose alors : *« Est-ce que la construction rapide d'un système de reconnaissance automatique de la parole est impossible pour des langues qui ne possèdent pas de dictionnaire phonétique ? »*

Depuis quelques années, des travaux utilisent la représentation orthographique (graphème) d'un mot à la place du phonème comme unité de modélisation acoustique [Kanthak 2002, Killer 2003b, Abdou 2004, Stuker 2004]. Pour le système de reconnaissance automatique de la parole à base de graphèmes, la représentation d'une entrée lexicale (un mot) dans le vocabulaire est alors une suite de *graphèmes*. Par conséquent, la construction automatique d'un dictionnaire de prononciation devient très simple. Le tableau 3.1 présente l'exemple d'un dictionnaire de prononciation en français à base de graphèmes.

Ainsi, pour les langues peu dotées pour lesquelles nous n'avons aucune connaissance linguistique, nous pouvons envisager d'utiliser l'approche de modélisation acoustique/graphémique (modélisation acoustique à base de graphèmes ou *grapheme based modelization*). La construction automatique d'un dictionnaire de prononciation devient alors très simple pour les systèmes utilisant une écriture latine.

Mot français	Prononciation à base de graphèmes
SIL	sil
abonnements	A B O N N E M E N T S
absolu	A B S O L U
bureautique	B U R E A U T I Q U E
chercheurs	C H E R C H E U R S
chocolat	C H O C O L A T
définitivement	D É F I N I T I V E M E N T
existence	E X I S T E N C E
...	…

Tableau 3.1 : Exemple du dictionnaire de prononciation en français à base de graphèmes

Cependant, pour les langues dont les systèmes d'écriture sont des caractères non-latins tels qu'alphasyllabiques[1] (laotien, thaïlandais, khmer, …), la construction d'un dictionnaire à base de graphèmes est un peu plus difficile. Elle consiste d'abord à convertir les caractères en une forme visible par l'ordinateur : la romanisation ou la latinisation[2]. Pour faire cela, nous pouvons utiliser les méthodes proposées de romanisation des caractères non-latins : noms de caractères Unicodes[3], tableau de romanisation du Khmer[4] (pour la langue khmère),… A titre d'exemple, V. Berment, dans le but de construire des outils de traitement de la langue laotienne, a transcrit des caractères laotiens d'un mot ou d'une phrase en utilisant des caractères latins [Berment 2004].

On remarquera aussi que dans un dictionnaire de prononciation à base de graphèmes, il n'existe qu'une variante de prononciation pour chaque entrée lexicale du vocabulaire.

4. Estimation de similarités entre unités phonémiques source / cible à base de distances

Les recherches sur la modélisation acoustique multilingue et la portabilité des modèles acoustiques font généralement l'hypothèse suivante : les représentations articulatoires des phonèmes sont similaires à travers les langues, les phonèmes sont donc considérés comme des unités **indépendantes de la langue** [Schultz 2001]. À partir de cette supposition, nous proposons dans cette section des mesures de similarité entre unités phonémiques (phonème, polyphone, groupe de polyphone, modèle polyphone, …) à base de distances, en particulier des similarités entre langue source et langue cible. Ces mesures de similarité seront ensuite utilisées pour la portabilité de modèles acoustiques multilingues dans la section suivante.

[1] http://fr.wikipedia.org/wiki/Alphasyllabaire
[2] La romanisation (ou latinisation) est la translittération ou la transcription d'une écriture non latine vers une écriture latine: http://fr.wikipedia.org/wiki/Romanisation
[3] http://www.unicode.org/charts/PDF/U1780.pdf
[4] http://www.eki.ee/wgrs/rom1_km.pdf

4.1. Distance entre deux phonèmes source/cible

Dans cette partie, nous présentons une mesure de *distance entre deux phonèmes*, en particulier la distance entre un phonème en langue source et un autre en langue cible. Cette distance peut être utilisée pour obtenir un tableau de correspondances entre phonèmes (*phone mapping table*) entre une ou plusieurs (multilingue) langues sources, et la langue cible.

Définition 1 : Soit M le nombre de phonèmes en langue source et N le nombre de phonèmes en langue cible. Soit $S = (s_1, s_2, \ldots, s_M)$ l'ensemble des phonèmes en langue source et $T = (t_1, t_2, \ldots, t_N)$ l'ensemble de phonèmes en langue cible. La similarité entre deux phonèmes source/cible est la distance acoustique/phonétique $d(s_i, t_j)$ entre les phonèmes s_i en langue source et t_j en langue cible (i=1..M, j=1..N).

En fait, il existe deux méthodes d'estimation de distance entre deux phonèmes quelconques de la langue source et de la langue cible. Les méthodes manuelles à base de connaissance phonétique (*knowledge-based*) consistent à estimer la distance des phonèmes source/cible suivant leur position dans le tableau de l'API. Les méthodes automatiques (*data-driven*) sont à base d'une méthode d'estimation de distance entre deux modèles acoustiques en langue source et en langue cible ou à base de matrices de confusion de phonèmes. La section suivante (4.1.1) présente une méthode automatique de construction d'une matrice de confusion, permettant d'estimer une distance entre phonèmes. La section 4.1.2, présentera quant à elle une nouvelle méthode d'estimation automatique de distance à base de connaissances phonétiques.

4.1.1. Méthodes automatiques (data-driven methods)

Il y a deux types de méthodes d'estimation de distance des phonèmes source/cible :

- méthodes à base de *distance entre deux modèles acoustiques* ;
- méthodes à base de *matrices de confusion entre phonèmes*.

Avec le premier type de méthodes à base de *distance de modèles acoustiques HMM*, on suppose qu'on dispose déjà de modèles acoustiques en langue source et en langue cible. Pour estimer la distance entre deux phonèmes source/cible, nous calculons la distance entre deux modèles acoustiques source/cible correspondants. Différentes méthodes d'estimation de distance de modèles HMM peuvent être proposées : distance euclidienne entre probabilités d'observation [Levinson 1983], méthode probabiliste [Juang 1985], … J. Köhler a été le premier à utiliser la distance entre modèles HMM pour estimer la similarité phonétique entre plusieurs langues dans un système multilingue [Kohler 1996]. J-J. Sooful a de son coté comparé quelques méthodes d'évaluation de la distance entre deux modèles HMM multilingues à base de leur distributions gaussiennes : *distance Kullback-Leibler, distance métrique Bhattacharyya, distance euclidienne, etc.* [Sooful 2001]. Pour implémenter ces méthodes, on doit déjà disposer de modèles acoustiques de bonne qualité en langue source et cible. Cela devient très difficile dans le contexte de notre travail puisque nous disposons seulement d'une faible quantité de données vocales en langue cible.

Le deuxième type de méthodes *à base de matrices de confusion de phonèmes* est celle utilisée dans notre travail. Elle nécessite de disposer d'un corpus vocal étiqueté phonétiquement, en quantité très limitée, en langue cible. Nous utilisons un décodeur phonémique en langue source pour décoder tous les signaux du corpus vocal en langue cible [Andersen 1994, Beyerlein 1999, Le 2005]. On obtient, pour chaque fichier de signal, une suite de phonèmes reconnus par le décodeur en langue source ; nous comparons alors cette suite de phonèmes avec la suite de phonèmes de référence étiquetés en langue cible sur un axe temporel pour obtenir la correspondance phonétique trame par trame (figure 3.3). On note qu'il n'est pas nécessaire que cet étiquetage de référence soit en très grande quantité.

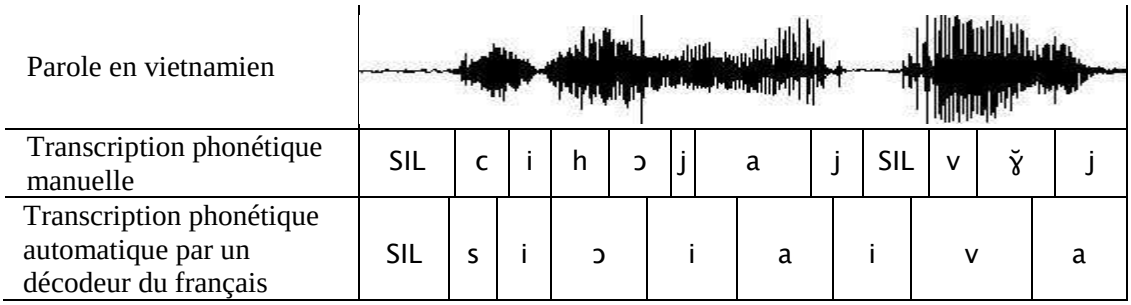


Figure 3.3 : Alignement temporel de phonèmes en langues source/cible

En décodant phonétiquement tous les signaux en langue cible, nous obtenons une correspondance phonétique globale qui nous permet de construire la matrice de confusion entre les phonèmes reconnus en langue source et les phonèmes de référence en langue cible. Nous avons publié cette approche et des expérimentations associées pour l'adaptation rapide de modèles acoustiques au vietnamien dans [Le 2005]. Le lecteur trouvera une explication plus détaillée dans l'annexe A.

On obtient ainsi une matrice $A(M, N)$ qui est la matrice de confusion de phonèmes, avec $0 \leq A_{i,j} \leq 1$, mesure de confusion entre le phonème t_j en langue cible et le phonème s_i en langue source. La distance entre phonèmes peut alors s'écrire simplement :

$$d(s_i, t_j) = A_{i,j} \tag{3.1}$$

où $A_{ij} \in [0, 1]$ et i=1..M, j=1..N

4.1.2. Nouvelle méthode à base de connaissances phonémiques

Les recherches dans les années récentes se sont limitées à la détermination des couples de phonèmes source/cible les plus proches ou bien à la construction d'un tableau de correspondance phonémique (*phone mapping table*) entre la langue source et la langue cible [Beyerlein 1999, Schultz 2001]. Cependant, aucune méthode à base de connaissance phonémique ne permet d'estimer objectivement la distance ou la similarité entre deux phonèmes quelconques. Dans cette partie, nous proposons une nouvelle méthode d'estimation automatique objective de distance entre deux phonèmes à partir de leur position dans le tableau de l'API.

En utilisant les unités phonétiques, les similitudes entre des phonèmes peuvent être

exprimées par l'API qui classifie des sons à partir de connaissances phonétiques. La figure 3.4 ci-dessous présente l'API pour les consonnes et les voyelles : les consonnes sont classifiées selon la position d'articulation, le type d'articulation, le voisement,… et les voyelles sont classifiées selon la position de la langue et l'arrondissement.

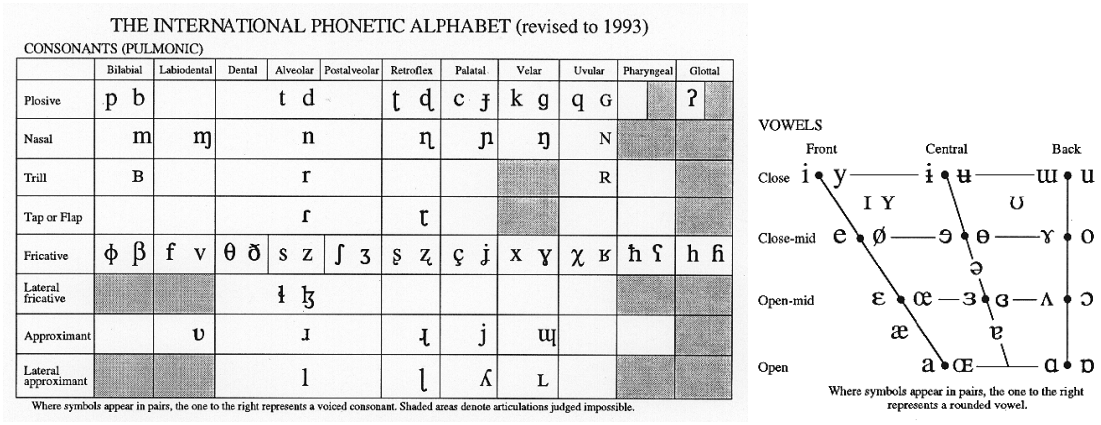


Figure 3.4 : API pour les consonnes et les voyelles [IPA 1999]

À partir des positions des phonèmes dans le tableau de l'API, nous pouvons estimer la distance entre deux phonèmes quelconques. Cette méthode consiste en deux étapes : la classification descendante (*top-down*) des phonèmes dans le graphe hiérarchique et l'estimation ascendante (*bottom-up*) de distance entre les phonèmes.

Etape 1: Classification descendante (top-down)

A base de la classification des phonèmes dans le tableau l'API, la figure 3.5 présente un graphe hiérarchique où chaque nœud est attaché à un groupe de phonèmes. Sur la base d'une connaissance de l'expert phonétique ou de l'expérimentation, chaque groupe de phonème est assigné à une valeur de similarité qui représente la similarité des éléments dans ce groupe. Tous les groupes de phonèmes sont placés en couches correspondent à leurs valeurs de similarité. Ainsi, les groupes dans la même couche obtiennent la même valeur de similarité. On note que si la valeur de similarité des phonèmes p_i et p_j est égale à 1, ils sont totalement différents. Par contre si cette valeur de similarité est égale à 0, les phonèmes p_i et p_j sont les mêmes.

Soit k le nombre de couches du graphe et G_i la valeur de similarité de la couche i avec i=0..k-1. Nous avons :

$$\begin{cases} G_i \in [0, 1] \\ G_i < G_j \text{ avec } i > j \end{cases} \quad (3.2)$$

Dans notre travail, nous essayons plusieurs valeurs différentes de k et G_i sur une expérimentation de modélisation acoustique crosslingue (système multilingue vers système vietnamien) [Le 2006] et les valeurs les plus appropriées sont G = {1 ; 0,9 ; 0,45 ; 0,25 ; 0,1 ; 0} avec k = 6 (comme illustré dans la figure 3.5).

Pour développer le graphe hiérarchique, nous commençons à la couche 0 contenant un seul groupe UNITE qui regroupe tous les phonèmes, les bruits et le silence. Ensuite, le groupe UNITE est divisé en trois groupes plus petits à la couche 1: PHONEME (contenant tous les phonèmes), BRUIT (contenant tous les types de bruits) et SILENCE. Le groupe PHONEME se subdivise à nouveau en deux groupes : CONSONNE et VOYELLE à la couche 2. Cette classification descendante des groupes est appliquée en augmentant et raffinant les critères de regroupement jusqu'à ce que chaque groupe ne consiste plus qu'en un seul phonème.

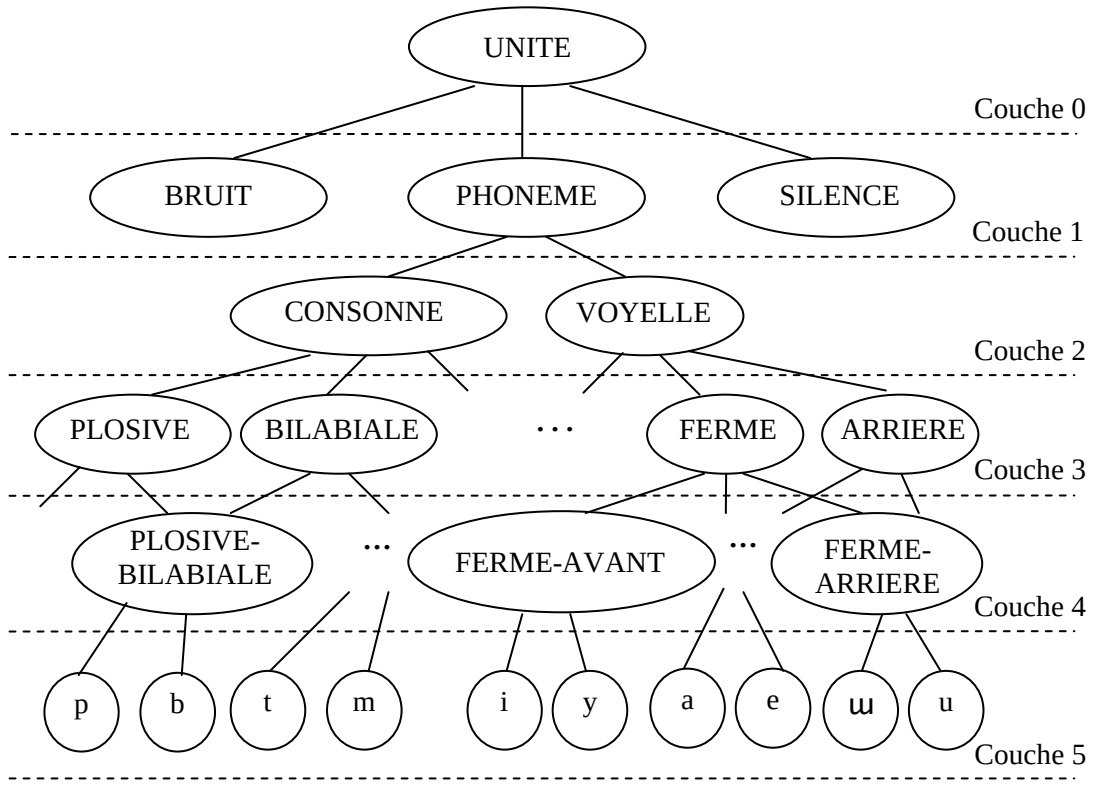


Figure 3.5 : Graphe hiérarchique pour l'estimation de distances entre phonèmes

Etape 2 : Estimation ascendante (bottom-up) de distance entre phonèmes

Pour estimer la distance entre deux phonèmes *s* et *t*, nous les localisons d'abord dans le graphe. Puis, à partir des feuilles du graphe qui leur correspondent, nous remontons dans le graphe jusqu'à ce qu'un nœud père commun aux deux phonèmes soit atteint. La distance entre le phonème *s* et le phonème *t* est alors égale à la valeur de similarité de la couche qui contient le nœud père de *s* et *t*. Nous avons :

$$d(s, t) = G_i \tag{3.3}$$

avec *i* indice de la couche qui contient le nœud père de *s* et *t*.

A titre d'exemple, le nœud père le plus proche des voyelles [i] et [u] est le groupe FERME sur la couche 3 (voir la figure 3.6). Alors : $d([i], [u]) = G_3$ (*égal à* 0,25 *dans l'exemple numérique)*.

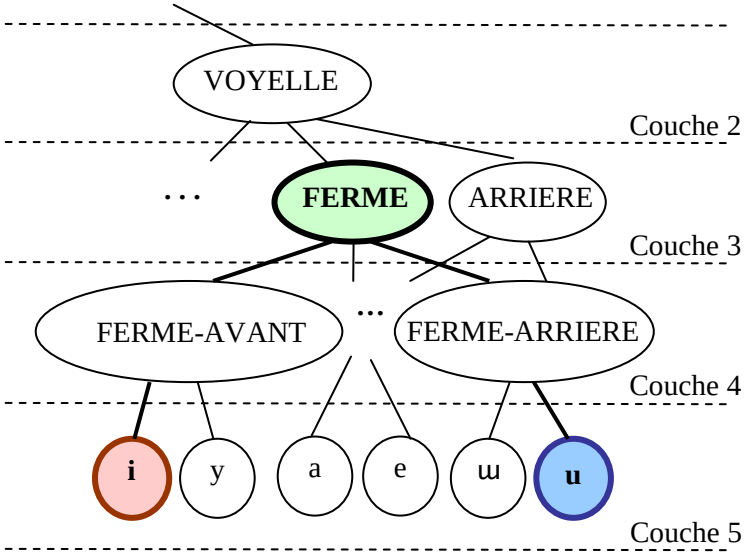


Figure 3.6 : Estimation d'une distance entre les voyelles [i] et [u]

4.2. Distance entre deux groupes de phonèmes source/cible

Définition 2 : Soit M le nombre de phonèmes en langue source et N le nombre de phonèmes en langue cible. Soit $S = (s_1, s_2, \ldots, s_M)$ l'ensemble des phonèmes en langue source et $T = (t_1, t_2, \ldots, t_N)$ l'ensemble des phonèmes en langue cible. Soit $C_S = (s_1, s_2, \ldots, s_m)$ une classe de m phonèmes en langue source ($C_S \subset S$) et $C_T = (t_1, t_2, \ldots t_n)$ une classe de n phonèmes en langue cible ($C_T \subset T$). La similarité entre deux groupes ou deux sous-ensembles de phonèmes est la distance phonétique $d(C_S, C_T)$ entre le groupe de phonèmes C_S en langue source et le groupe de phonèmes C_T en langue cible (voir la figure 3.7).

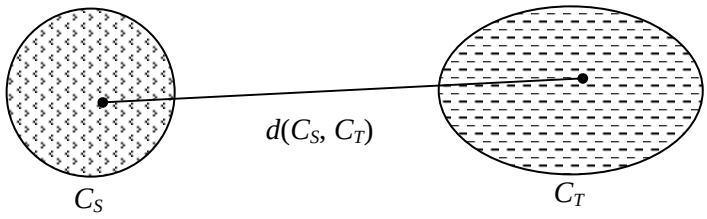


Figure 3.7 : Distance entre deux groupes de phonèmes

Nous appelons $d(s, C_T)$ la distance entre un phonème s en langue source et un groupe de phonèmes $C_T = (t_1, t_2, \ldots, t_n)$ en langue cible (voir la figure 3.8). Cette distance est calculée comme suit :

$$d(s, C_T) = \frac{\sum_{j=1}^{n} d(s, t_j)}{n} \tag{3.4}$$

où $d(s, t_j)$ la distance entre deux phonèmes s et t_j ($j = 1..n$), calculée par l'équation (3.1) ou (3.3)

dans la section précédente (méthode guidée par les données ou utilisant une connaissance phonétique).

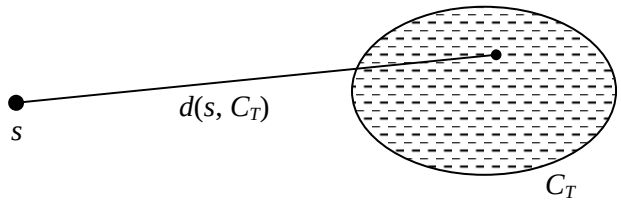


Figure 3.8 : Distance entre un phonème et un groupe de phonèmes

De la même manière, nous avons $d(t, C_S)$ la distance entre un phonème t en langue cible et un groupe de phonèmes $C_S = (s_1, s_2, \ldots, s_m)$ en langue source :

$$d(t, C_S) = \frac{\sum_{i=1}^{m} d(t, s_i)}{m} \tag{3.5}$$

À partir des distances entre un phonème et un groupe de phonèmes $d(s, C_T)$ et $d(t, C_S)$, nous proposons alors la distance entre groupes de phonèmes illustrée par l'équation (3.6) ci-dessous :

$$d(C_S, C_T) = \frac{\sum_{i=1}^{m} d(s_i, C_T)}{m} = \frac{\sum_{j=1}^{n} d(t_j, C_S)}{n} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d(s_i, t_j)}{m.n} \tag{3.6}$$

La distance entre deux groupes de phonèmes est égale la moyenne de toutes les distances partielles entre deux éléments (deux phonèmes) des ensembles C_S et C_T.

4.3. Distance entre deux polyphones source/cible

La distance entre deux polyphones source/cible est proposée et estimée en supposant que :

1) le contexte considéré d'un polyphone dans notre problème est symétrique et unique ; c'est-à-dire que la longueur du contexte gauche et du contexte droit des polyphones sont les mêmes ;

2) la longueur du contexte d'un polyphone en langue source est égale à celle du contexte en langue cible. Sinon, nous devons appliquer une procédure de normalisation en fonction de la longueur du contexte.

Définition 3 : Soit L la longueur des contextes gauche et droit d'un polyphone en langue source et en langue cible. Soit P_S un polyphone en langue source et P_T un polyphone en langue cible. La similarité entre deux polyphones est la distance phonétique $d(P_S, P_T)$ entre le polyphone P_S en langue source et le polyphone P_T en langue cible.

Nous supposons que :

$$P_S = (s_{-L}, s_{-L+1}, \ldots, s_{-1}, s_0, s_1, s_2, \ldots, s_L) \quad (3.7)$$

où s_0 phonème central modélisé dans son contexte gauche $s_{-L}, s_{-L+1}, \ldots, s_{-1} \in S$ et son contexte droit $s_1, s_2, \ldots, s_L \in S$. Et

$$P_T = (t_{-L}, t_{-L+1}, \ldots, t_{-1}, t_0, t_1, t_2, \ldots, t_L) \quad (3.8)$$

ou t_0 phonème central modélisé dans son contexte gauche $t_{-L}, t_{-L+1}, \ldots, t_{-1} \in T$ et son contexte droit $t_1, t_2, \ldots, t_L \in T$.

La distance entre le polyphone P_S en langue source et le polyphone P_T en langue cible est calculée à partir des distances entre les phonèmes source/cible dans les contextes correspondant en appliquant l'équation suivante :

$$d(P_S, P_T) = \alpha_0.d(s_0, t_0)+\alpha_1.[d(s_{-1}, t_{-1})+d(s_1, t_1)]+\ldots\alpha_L.[d(s_{-L}, t_{-L})+d(s_L, t_L)] \quad (3.9)$$

avec : - $\alpha_0, \alpha_1, \ldots, \alpha_L$: coefficients de distance suivant la position dans le contexte, ils sont aussi déterminés par l'utilisateur.

- $d(s_k, t_k)$: la distance entre deux phonèmes pour $k = -L,\ldots L$.

La figure 3.9 présente un exemple du calcul de la distance entre un polyphone P_S = (A B **C** D E) en langue source et un polyphone P_T = (a b **c** d e) en langue cible.

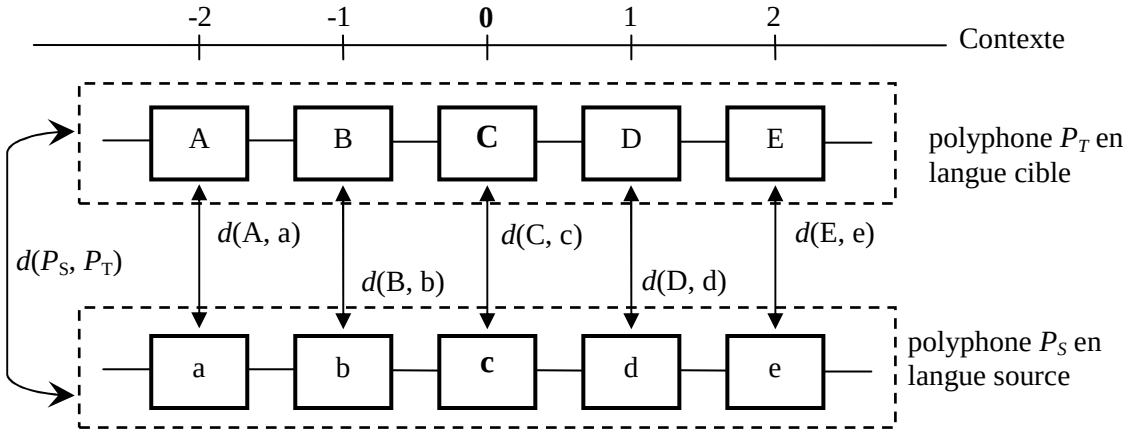


Figure 3.9 : Calcul de la distance entre deux polyphones

Dans le cas où les contextes droits et gauches de P_S et P_T ne sont pas des phonèmes mais des groupes de phonèmes, les équations (3.7) et (3.8) deviennent alors :

$$P_S = (CS_{-L}, CS_{-L+1}, \ldots CS_{-1}, s_0, CS_1, CS_2, \ldots, CS_L) \quad (3.10)$$

$$P_S = (CS_{-L}, CS_{-L+1}, \ldots CS_{-1}, s_0, CS_1, CS_2, \ldots, CS_L) \quad (3.11)$$

avec : - $CS_k \subset S$ un groupe (sous-ensemble) de phonèmes en langue source.

- $CT_k \subset T$ un groupe (sous-ensemble) de phonèmes en langue cible.

L'équation (3.9) devient :

$$d(P_S, P_T) = \alpha_0.d(s_0, t_0)+\alpha_1.[d(CS_{-1}, CT_{-1})+d(CS_1, CT_1)]+\ldots\alpha_L.[d(CS_{-L}, CT_{-L})+d(CS_L, CT_L)] \quad (3.12)$$

avec $d(CS_k, CT_k)$ la distance de deux groupes (sous-ensemble) de phonèmes ($k = -L..L$) calculée par l'équation (3.6).

Par conséquent, le polyphone P_S* en langue source sera le plus similaire d'un polyphone P_T en langue cible si et seulement si il satisfait l'expression suivante :

$$\forall P_S \in S,\ d(P_S^*, P_T) = \min[\ d(P_S, P_T)\ ] \quad (3.13)$$

4.4. Distance entre deux groupes de polyphones source/cible

Nous avons proposé les équations (3.9) et (3.12) pour calculer les distances entre deux polyphones dont les contextes gauches et droits sont des phonèmes ou des groupes de phonèmes. À partir de ces distances, nous pouvons chercher le polyphone en langue source le plus proche d'un polyphone en langue cible qui satisfait l'équation (3.13).

Cependant, dans les conditions réelles où le nombre de polyphones d'une langue est très large (par exemple, plus de 100 000 triphones en anglais), il est très difficile de collecter un corpus vocal d'apprentissage qui pourrait couvrir tous les polyphones. Pour la modélisation acoustique dépendante du contexte, nous avons donc besoin de choisir des modèles plus simples car nous ne pouvons pas modéliser tous les contextes possibles des phonèmes. En conséquence, nous pouvons regrouper les polyphones similaires (même contexte droit par exemple) dans un ensemble de polyphones (voir la figure 3.10) en utilisant une procédure de regroupement de polyphones par agglomération (*agglomerative polyphone clustering)* [Imperl 2000] ou à partir d'un arbre de décision (*decision tree-based clustering*) [Huang 2001]. Cet arbre de décision est construit en appliquant une question de contexte du type : Est-ce que le contexte gauche est une CONSONNE ? Est-ce que le contexte droit est le phonème B ? etc.

Définition 4 : Soit $\Phi_S = (P_{S1}, P_{S2}, \ldots, P_{Sm})$ un groupe de m polyphones en langue source et $\Phi_T = (P_{T1}, P_{T2}, \ldots, P_{Tm})$ un groupe de n polyphones en langue cible. La similarité entre deux groupes de polyphones est la distance acoustico-phonétique $d(\Phi_S, \Phi_T)$ entre le groupe de polyphone Φ_S en langue source et le groupe de polyphone Φ_T en langue cible.

De la manière que nous avons déterminé la distance entre deux groupes de phonèmes (équation 3.6), la distance entre deux groupes de polyphones est égale à la moyenne des distances entre tous les couples de polyphones composant les deux ensembles Φ_T et Φ_S. Nous avons :

$$d(\Phi_S, \Phi_T) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d(P_{Si}, P_{Tj})}{m.n} \quad (3.14)$$

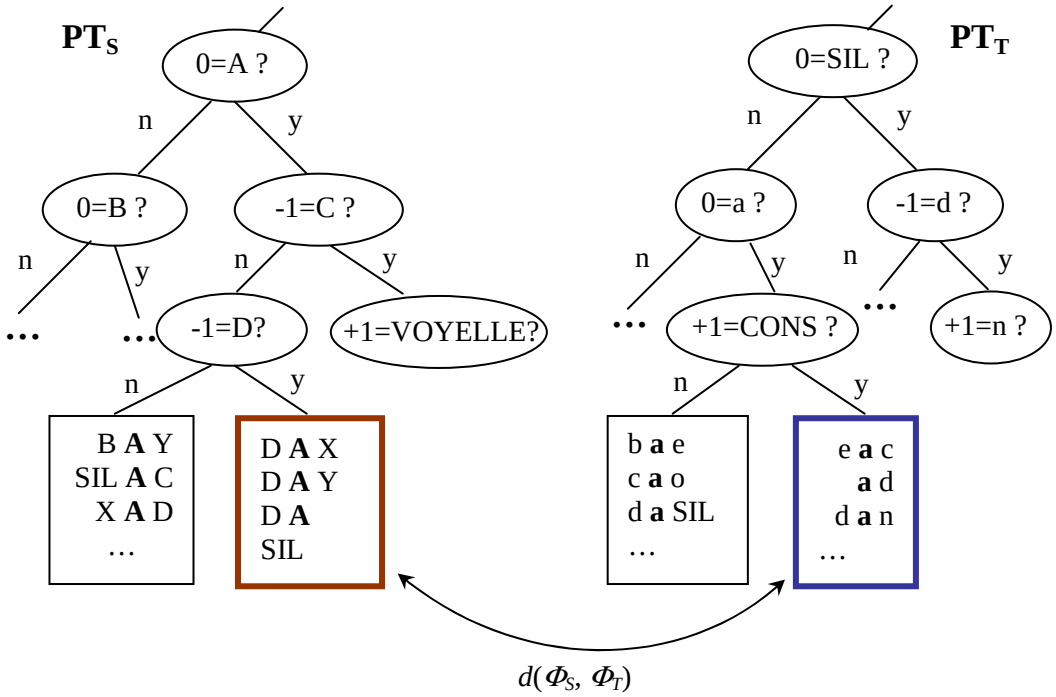


Figure 3.10 : Distance entre deux groupes polyphoniques dans les arbres de décision source/cible

Par conséquent, le groupe de polyphones Φ_S* en langue source le plus similaire du groupe de polyphones Φ_T en langue cible doit satisfaire cette expression :

$$\forall \Phi_S,\ d(\Phi_S^*, \Phi_T) = \min\ [d(\Phi_S, \Phi_T)] \quad (3.15)$$

Pour la modélisation acoustique, le groupe de polyphones Φ_S en langue source et le groupe de polyphones Φ_T en langue cible sera modélisé par un modèle HMM. Alors, les équations (3.14) et (3.15) pourront être utilisées pour estimation de la distance entre deux modèles acoustiques et la recherche des couples de modèles acoustiques source/cible les plus proches.

5. Construction et adaptation rapide de modèles acoustiques pour des langues peu dotées

Nous avons vu précédemment qu'il existe des solutions pour collecter rapidement des ressources orales et écrites dans une nouvelle langue. Dans le cas idéal, si ces ressources sont en grande quantité, et si un dictionnaire de prononciation est disponible pour la langue cible,

l'adaptation du système de reconnaissance peut correspondre alors à un simple réapprentissage des modèles sur ces nouvelles données.

Dans la réalité, la quantité de données collectées pour les langues peu dotées reste bien souvent inférieure à ce qu'elle est pour les langues bien dotées. La construction d'un système de reconnaissance automatique de la parole nécessite donc également des techniques d'adaptation rapide au niveau des modèles acoustiques comme cela a déjà été proposé dans [Beyerlein 1999, Schultz 2001] par exemple. Pour la portabilité et l'adaptation de modèles acoustiques *(cross-lingual acoustic modeling)* vers une nouvelle langue, des modèles acoustiques de la langue cible sont initialisés en empruntant des modèles acoustiques existant en langue source et ces modèles initiaux sont ensuite adaptés avec une quantité réduite de signaux en langue cible.

Avant de travailler sur la portabilité et l'adaptation de modèles acoustiques, nous discutons d'abord le problème de la construction de tableaux de correspondance phonémique source/cible.

5.1. Construction de tableaux de correspondance phonémique

La portabilité de modèles acoustiques multilingues *indépendants du contexte* vers une nouvelle langue nécessite de construire un tableau de correspondances phonémiques (*phone mapping table*) entre une langue (cas monolingue) ou plusieurs (cas multilingue) langues sources et la langue cible. Pour cela, on distingue les méthodes manuelles à base de connaissances (*knowledge-based*) et les méthodes automatiques (*data-driven*). Les méthodes manuelles consistent à chercher les couples de phonèmes source/cible les plus proches dans le tableau d'API et nécessitent des connaissances acoustiques et phonétiques des deux langues (source et cible) [Beyerlein 1999, Schultz 2001]. Une approche automatique consiste plutôt à disposer d'un corpus vocal en quantité limitée en langue cible et étiqueté (quelques minutes peuvent suffire), puis à chercher les couples de phonèmes source/cible les plus proches suivant une fonction de distance entre phonèmes [Kohler 1996] ou suivant la matrice de confusion de phonèmes source/cible [Andersen 1994, Beyerlein 1999, Le 2005].

Pour la modélisation acoustique indépendante du contexte (monophones), avec P_S, P_T deux monophones en langue source et cible, l'équation (3.12) se réécrit :

$$d(P_S, P_T) = d(s_0, t_0) \tag{3.16}$$

avec $\propto_0 = 1$ et $d(s_0, t_0)$ est la distance entre deux phonèmes source/cible calculée par l'équation (3.1) ou l'équation (3.2). L'équation (3.13) devient :

$$\forall P_S \in S,\ d(P_S{*}, P_T) = \min [\ d(P_S, P_T)\ ] = \min [\ d(s_0, t_0)\ ] \tag{3.17}$$

En appliquant l'équation (3.17), porter des modèles acoustiques indépendants du contexte multilingues vers une nouvelle langue cible consiste à déterminer les couples de phonèmes les plus similaires ou bien à construire le tableau de correspondances phonémiques (*phone mapping table*) source/cible. Le tableau 3.2 présente des exemples de correspondances phonémiques

entre la langue cible (vietnamien) et la langue source (français, ou modèle multilingue issu de sept langues).

Phonème vietnamien	Phonème français		Phonème issu d'un ensemble multilingue (*GlobalPhone*)	
	Obtenu manuellement	*Obtenu automatiquement*	*Obtenu manuellement*	*Obtenu automatiquement*
t	t	t	t	t
ɣ	g	g	ɣ	g
χ	k	k	χ	χ
ŋ	ŋ	ŋ	ŋ	ŋ
ʂ	s	s	ʂ	ʂ
w	w	w	w	au
e	e	e	e	e
uo	uœ	o	uɔ	u
ie	jø	i	iɛ	i
...	...	...	...	...

Tableau 3.2 : Tableau de correspondances phonémiques
langue source : français et multilingue ; langue cible : vietnamien

5.2. Portabilité et adaptation de modèles acoustiques crosslingues

Cette section présente notre apport sur la portabilité et l'adaptation de modèles acoustiques vers de nouvelles langues. Quand la plupart des recherches dans ce domaine se sont limitées à la portabilité de modèles acoustiques indépendant du contexte (modèles monophoniques), nous proposons une nouvelle méthode de portabilité de modèles acoustiques dépendants du contexte multilingues vers une nouvelle langue qui utilise les distances entre deux groupes de polyphones source/cible présentées précédemment.

5.2.1. Portabilité et adaptation de modèles acoustiques indépendants du contexte

La figure 3.11 ci-dessous résume le processus pour la portabilité et l'adaptation de modèles acoustiques indépendants du contexte vers une nouvelle langue cible. Après avoir obtenu le tableau de correspondances phonémiques construit dans la section précédente, les modèles acoustiques indépendants de contexte (monophones) en langue source peuvent être dupliqués pour obtenir des modèles acoustiques en langue cible. L'avantage d'une telle approche est qu'elle ne nécessite pas ou peu de signaux d'apprentissage en langue cible puisque les modèles acoustiques du système de reconnaissance en langue cible sont en fait ceux d'une autre langue.

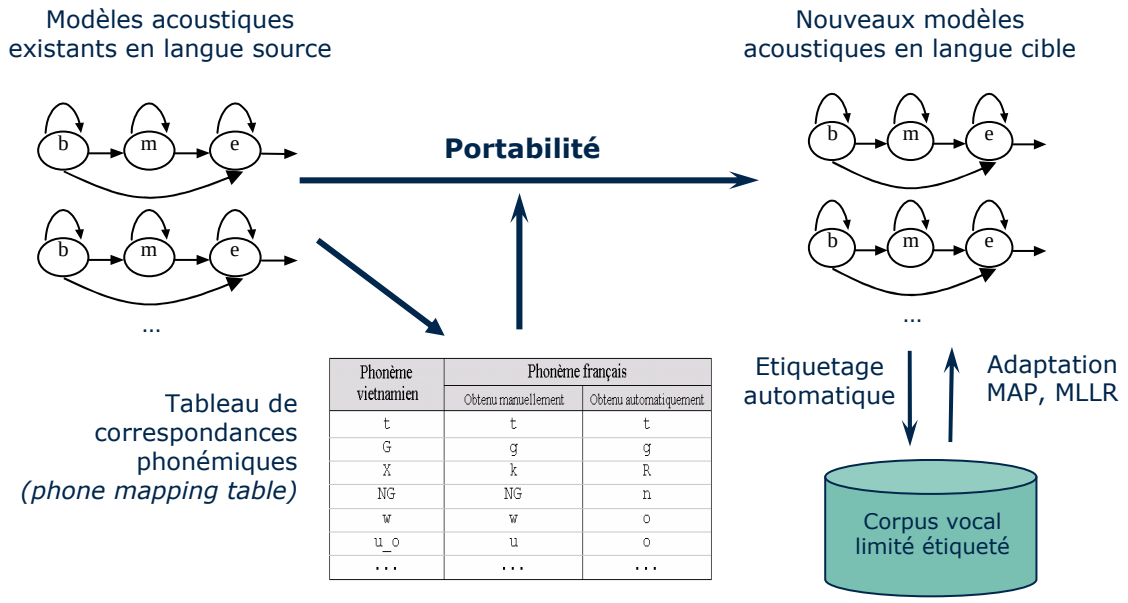


Figure 3.11 : Portabilité et adaptation de modèles acoustiques indépendants du contexte

Par exemple, les modèles de phonèmes [c] et [tʰ] du vietnamien sont dupliqués à partir du modèle [t] du français (figure 3.12).

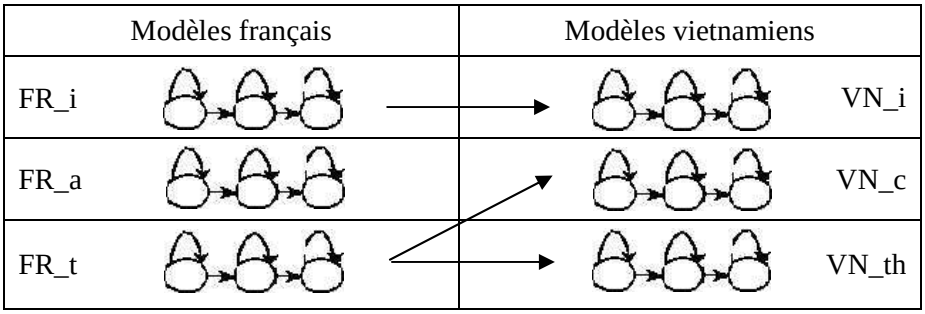


Figure 3.12 : Portabilité de modèles acoustiques français vers des modèles vietnamiens

Cependant, on retrouve dans cette approche les mêmes défauts que ceux mentionnés dans la méthode de construction automatique d'un dictionnaire de prononciation à savoir le problème de la couverture phonémique (i.e. reconnaître du vietnamien avec des modèles acoustiques appris sur du français). De tels systèmes peuvent cependant être améliorés en adaptant, par exemple, les modèles acoustiques avec une quantité réduite de signaux en langue cible. Nos résultats expérimentaux sur ce point sont présentés en détail dans [Le 2005] et aussi dans le chapitre 4 sur la construction rapide d'un système de reconnaissance automatique de la parole en langue vietnamienne.

5.2.2. Portabilité et adaptation de modèles acoustiques dépendant du contexte

La couverture phonémique source/cible diminue dramatiquement en élargissant le contexte des polyphones modélisés. Par exemple, le taux de couverture des monophones portugais est de

91% pour un modèle multilingue contenant 9 langues, mais la couverture diminue à 73% pour des triphones et 47% pour des quintphones, respectivement [Schultz 2001].

On trouve assez peu de recherches sur la portabilité de modèles acoustiques multilingues dépendants du contexte vers une nouvelle langue. J. Köhler utilise la distance entre deux modèles de phonèmes HMM pour la portabilité de modèles indépendants du contexte [Kohler 1996]. Il mentionne, comme une perspective, que cette technique pourrait être appliquée à la portabilité de modèles dépendants du contexte. Une méthode d'estimation de similarité entre deux triphones est proposée par B. Imperl et utilisée pour regrouper des triphones dans un système multilingue [Imperl 2003].

Nous constatons qu'un problème important pour la portabilité de modèles acoustiques dépendant du contexte est la disparité de contextes (*context mismatch*) à travers les langues qui augmente pour des contextes plus larges. Une méthode efficace nommée PDTS (*Polyphone Decision Tree Specialization*) est proposée par T. Schultz et appliquée sur l'arbre de décision multilingue pour résoudre ce problème. Dans la méthode PDTS, on adapte l'arbre de décision de polyphones multilingues existante en utilisant un corpus vocal en quantité limitée en langue cible [Schultz 2001].

Dans cette partie, nous proposons une autre méthode de portabilité de modèles acoustiques dépendants du contexte multilingue vers une nouvelle langue, en utilisant notre distance entre deux groupes de polyphones source/cible présentée dans la section 4 précédente. La figure 3.13 présente le schéma général de notre technique.

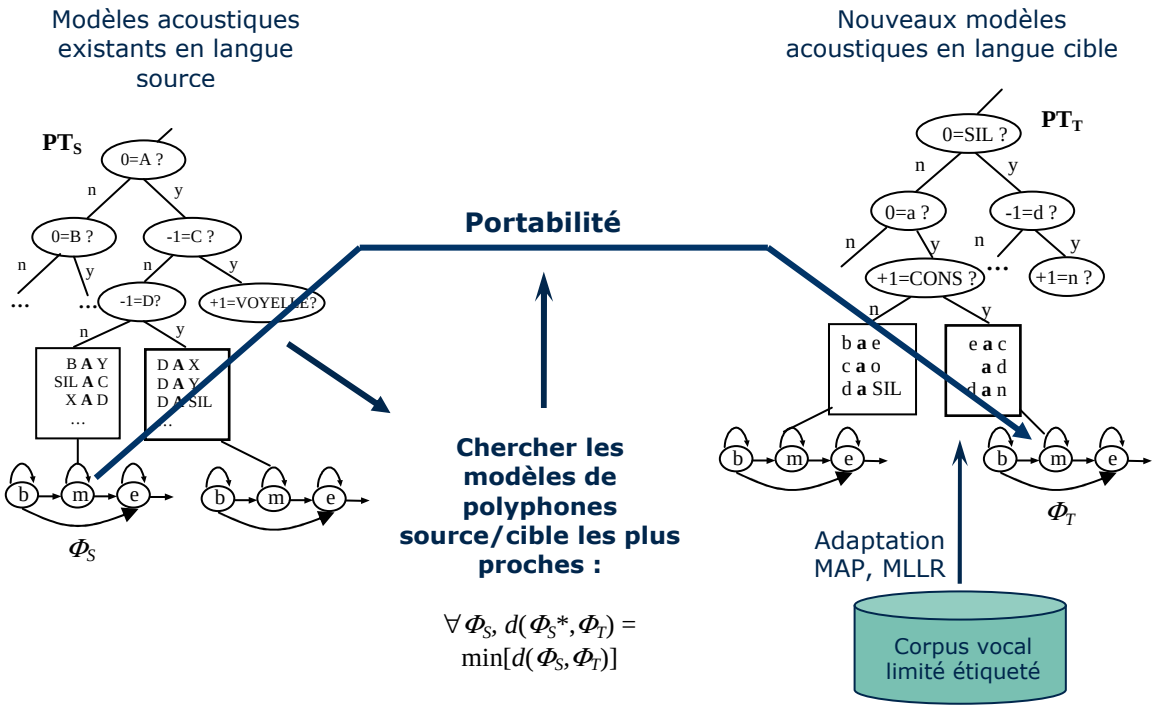


Figure 3.13 : Portabilité et adaptation de modèles acoustiques dépendants du contexte

Tout d'abord, nous utilisons un corpus de parole en quantité limitée en langue cible pour développer un arbre de décision en langue cible (PT_T). Nous supposons qu'un arbre de décision en langue source (PS_S), ainsi que des modèles acoustiques entraînés sur une base de données de signaux multilingues, est disponible. La figure 3.14 présente deux arbres de décision PT_T en langue cible et PT_S en langue source où les polyphones sont groupés et placés sur les nœuds des arbres. Ensuite, en appliquant l'équation (3.12) ou (3.14) nous pouvons estimer les distances entre tous les couples de 2 polyphones ou de 2 groupes de polyphones source/cible. Ceci nous permet, en appliquant l'équation (3.13) ou (3.15), de déterminer, pour un polyphone (ou un groupe de polyphones) en langue cible, le polyphone (ou le groupe de polyphones) le plus proche (au niveau phonétique) en langue source. Enfin, les modèles HMM des polyphones en langue source peuvent être dupliqués pour obtenir des modèles acoustiques en langue cible. De tels modèles initiaux peuvent cependant être améliorés en adaptant (apprentissage Viterbi, adaptation MLLR, MAP, etc.) avec une quantité réduite de signaux en langue cible.

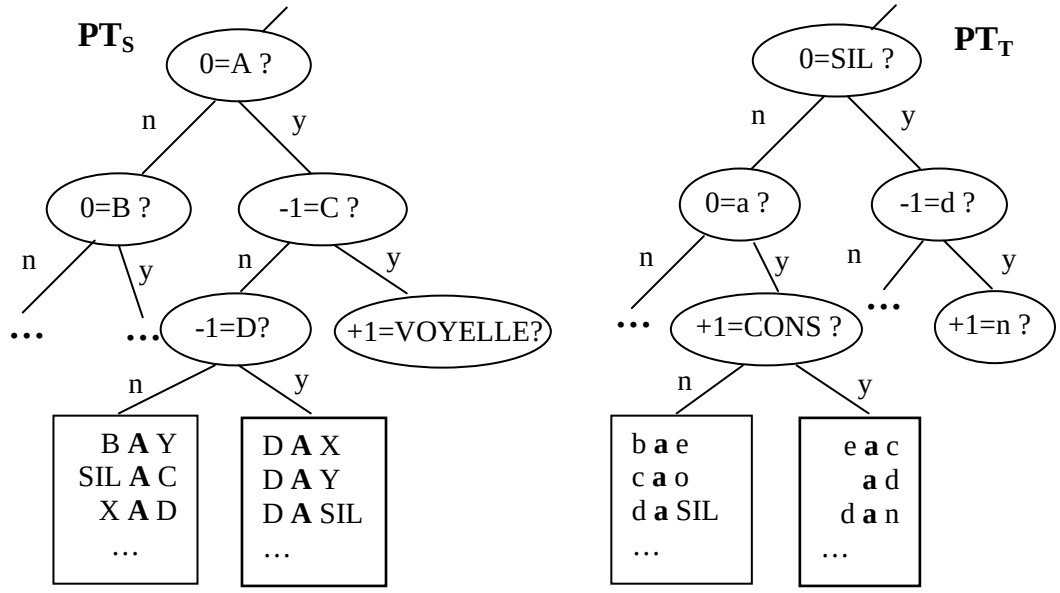


Figure 3.14 : Arbres de décision source/cible

6. Modélisation acoustique à base de graphèmes

Nous rappelons que pour l'apprentissage des modèles acoustiques pour un système de reconnaissance automatique de la parole continue à grand vocabulaire dans une nouvelle langue, il est souvent nécessaire de rassembler une grande quantité de ressources, contenant à la fois des signaux de parole étiquetés mais également un dictionnaire de prononciation. Ces ressources sont souvent indisponibles pour des langues peu dotées.

Récemment, des travaux sur la modélisation acoustique à base de graphèmes ont vu le jour. Le système de reconnaissance à base de graphèmes peuvent être multilingue comme dans [Kanthak 2002] ou monolingue tels que pour l'allemand, l'anglais et l'espagnol [Killer 2003b], pour l'arable [Abdou 2004] ou pour le russe [Stuker 2004]. La représentation orthographique

(caractère ou graphème) est utilisée comme l'unité de modélisation acoustique au lieu du phonème. La performance du système de reconnaissance automatique de la parole à base de graphèmes dépend fortement de la différence graphème-phonème de la langue considérée. Les résultats d'expérimentations présentés dans [Kanthak 2002, Killer 2003b, Abdou 2004] montrent que, si la modélisation indépendante du contexte à base de graphèmes n'est pas efficace, la modélisation dépendante du contexte obtient en revanche des résultats comparables avec les approches à base de phonèmes pour des langues telles que l'espagnol et surtout l'anglais, qui présente une faible différence graphèmes/phonèmes. En conséquence, pour les langues peu dotées qui ne possèdent pas encore un dictionnaire phonétique, cette approche de modélisation acoustique à base de graphèmes présente un potentiel intéressant pour construire rapidement un système de reconnaissance automatique de la parole.

Dans cette partie, nous présentons notre travail sur la modélisation acoustique à base de graphèmes pour des langues peu dotées. Une méthode d'initialisation des modèles acoustiques plus rapide et plus efficace que l'état de l'art sera aussi présentée.

6.1. Initialisation de modèles acoustiques graphémiques

Normalement, pour entraîner rapidement des modèles acoustiques, nous avons besoin d'un corpus de parole étiqueté (aligné temporellement) [Wheatley 1994, Schultz 1999]. C'est-à-dire que chaque trame des signaux doit être associée à une étiquette soit phonémique (système à base de phonèmes), soit graphémique (système à base de graphèmes). En pratique, nous utilisons une procédure d'alignement temporel (*Automatic Time Alignment*) telle que l'algorithme Viterbi à l'aide des modèles acoustiques.

Si les modèles acoustiques n'existent pas au départ, nous pouvons utiliser des stratégies d'initialisation de modèles acoustiques : démarrage aléatoire (*random start*), démarrage uniforme (*flat start*), …Ces modèles initiaux sont ensuite appris sur des données qui sont segmentées uniformément [Killer 2003a].

La figure 3.15 montre un exemple de segmentation uniforme pour initialiser des modèles acoustiques de graphèmes. On voit qu'une telle initialisation présente beaucoup d'erreurs d'étiquetage. Les signaux de silence sont étiquetés par un graphème et inversement.

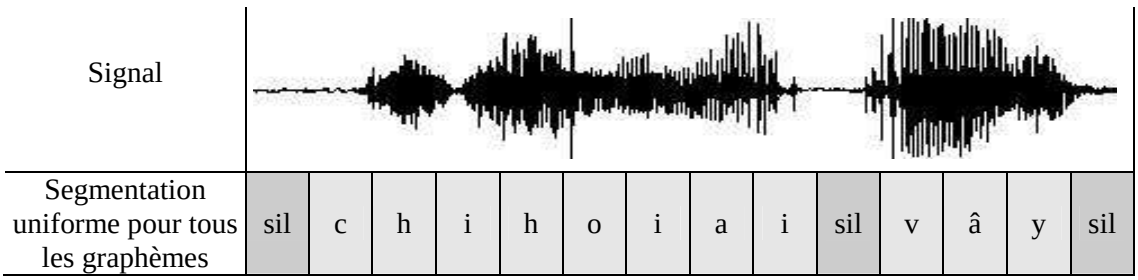


Figure 3.15 : Segmentation et étiquetage uniforme des données pour la phrase « Chị hỏi ai vậy ? » en vietnamien

Pour réduire ces erreurs d'étiquetage, après avoir appris des modèles acoustiques initiaux, les signaux d'apprentissage sont réalignés temporellement à l'aide de modèles acoustiques initiaux. Ensuite, les modèles acoustiques sont ré-entraînés à partir du corpus de signaux étiquetés (apprentissage à base d'étiquettes) et l'on réitère le cycle. La procédure d'apprentissage est arrêtée quand le système atteint un état stable.

Pour la modélisation à base de phonèmes, nous pouvons emprunter des modèles acoustiques existant en langue source, ensuite les données sont étiquetées automatiquement en appliquant l'algorithme d'alignement temporel de Viterbi. Cependant, pour la modélisation à base de graphèmes, les techniques de portabilité de modèles acoustiques ne sont pas efficaces car les similarités des graphèmes entre langues sont très faibles. En effet, la technique d'initialisation à partir des modèles acoustiques graphémiques *multilingues* a été utilisée précédemment dans [Killer 2003a]. Avec cette technique crosslingue, seul le temps d'apprentissage du système est réduit et le taux de reconnaissance du système n'est pas augmenté. Cela montre que la qualité des étiquettes graphémiques n'est pas vraiment améliorée après quelques cycles d'apprentissage.

Nous proposons dans cette section une méthode d'amélioration d'initialisation des paramètres des modèles acoustiques en utilisant une détection de frontières des mots. En fait, nous pouvons améliorer et affiner des erreurs d'alignement temporel en pré-détectant des frontières entre les mots et entre mots et silences. Pour cela, nous pouvons segmenter les données vocales en mot et en silence. Ensuite, pour chaque mot segmenté, nous segmentons *uniformément* en graphèmes.

La figure 3.16 illustre la segmentation de la même phrase que la figure 3.15 par notre méthode à base d'une détection de frontière des mots. En effet, cette méthode d'initialisation nous permet, pour chaque phrase d'énoncés en entrée, de réduire les erreurs de segmentation globales (les erreurs « *inter-mot* »). Il existe encore des erreurs locales « *intra-mot* ». Cependant, ces erreurs locales peuvent facilement et rapidement disparaître après quelques boucles de « bootstrapping » par les algorithmes d'apprentissage.

Décodage des frontières de mots par des modèles « *mot/silence* »	SIL	MOT			MOT			MOT			SIL	MOT		
Récupérer la transcription du mot	**sil**	**chị**			**hỏi**			**ai**			**sil**	**vậy**		
Segmentation *uniforme* pour chaque mot	sil	c	h	i	h	o	i	a	i		sil	v	â	y

Figure 3.16 : Segmentation et étiquetage initial en utilisant une détection de frontières des mots pour la phrase « Chị hỏi ai vậy ? »

Les sections suivantes présentent en détail notre méthode d'initialisation des modèles acoustiques graphémiques à base d'une détection de frontière de mots.

6.1.1. Détection de frontière de mots

La détection de la frontière des mots (*Word Boundary Detection*) est un problème fondamental pour la reconnaissance automatique de la parole continue. Pour la parole continue spontanée, et contrairement au cas des mots connectés ou enchaînés, la frontière entre mots est floue et elle devient difficile à détecter [Rabiner 1993]. La plupart des méthodes utilisent des informations temporelles ou fréquentielles pour détecter la frontière entre silence/bruit et parole ou entre mots : énergie du signal [Gu 2002], taux croissant par zéro (ZCR), pitch, durée du signal [Rao 1992, Rao 1996], entropie [Waheed 2002], etc. Par ailleurs, les frontières des mots peuvent être détectées par l'algorithme d'alignement temporel *Viterbi* sur les modèles HMMs qui modélisent tous les mots dans le vocabulaire. Cette méthode montre une bonne performance en présence de cas de bruits et de frontières floues en parole continue [Wilpon 1987].

Nous employons, dans notre travail, la méthode de détection de frontière des mots à l'aide de modèles HMM de mots. Cependant, quand la modélisation de tous les mots dans le vocabulaire (chaque mot est modélisé par un modèle HMM) devient impossible pour des systèmes de reconnaissance automatique de la parole grand vocabulaire, nous décidons de n'utiliser qu'un modèle de mot « générique » pour tous les mots et un modèle de silence pour le silence et les bruits. Nous avons alors une modélisation « mot/silence ». Nous espérons qu'avec cette modélisation « mot/silence », le modèle du mot générique modélise correctement le début et la fin (au niveau signal) du mot. Ainsi, les modèles « mots/silence » peuvent être appliqués à la détection de frontière de mots.

a) Modélisation « mot/silence »

Nous construisons deux modèles acoustiques généraux à partir d'un corpus d'apprentissage : modèle de mot générique et modèle de silence (voir la figure 3.17). Un mot est modélisé par un HMM à trois états gauche-droit. Le modèle HMM silence consiste en un état unique.

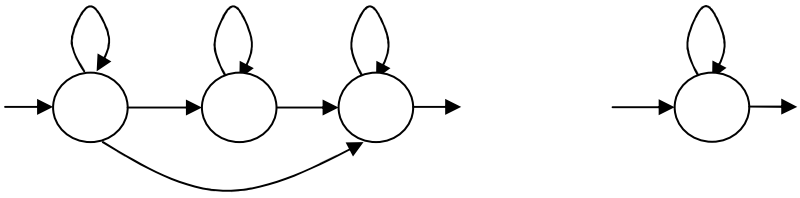


Figure 3.17 : Modélisation « mot/silence »

Pour l'apprentissage de modèles mot/silence, nous pouvons utiliser les données qui sont segmentées en mot et en silence. Cette segmentation en mots peut être effectuée par une concaténation des étiquettes des phonèmes (obtenues par les modèles acoustiques phonémiques) comme illustré dans la figure 3.18.

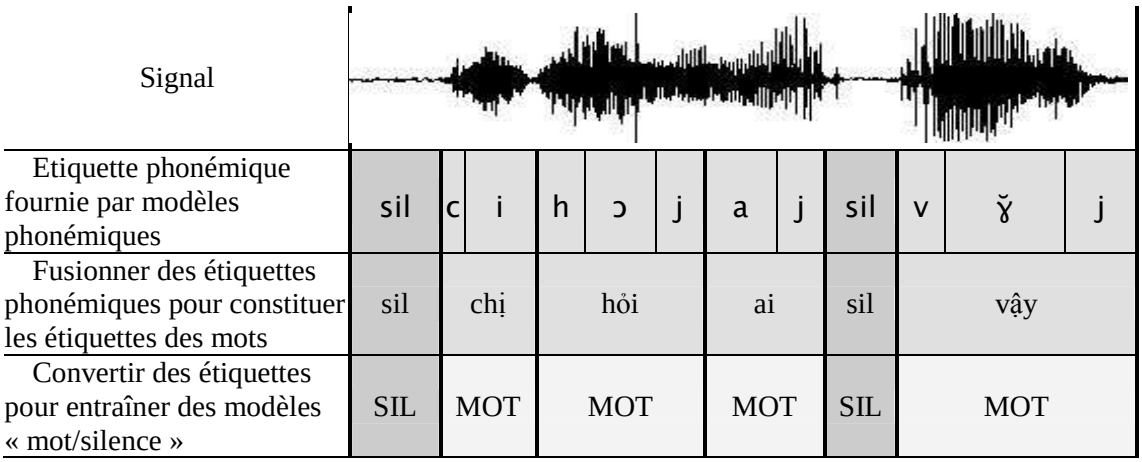


Figure 3.18 : Entraîner les modèles « mot/silence » à base d'une concaténation des étiquettes phonémiques

Après avoir segmenté des données de mots et des silences, les modèles « mot/silence » sont entraînés par quelques cycles de *bootstrapping*. La procédure d'apprentissage est arrêtée quand le système atteint un état stable, normalement après 6-8 itérations de la procédure d'apprentissage.

b) Détection de frontière de mots à base de modélisation « mot/silence »

Pour chaque fichier signal et fichier de transcription en entrée, les modèles HMM de mot et silence correspondants sont combinés consécutivement pour obtenir un modèle HMM entier (voir la figure 3.19). En appliquant l'algorithme d'alignement temporel de Viterbi sur le signal d'entrée, les états les plus probables correspondant à chaque trame de signal sont décodés.

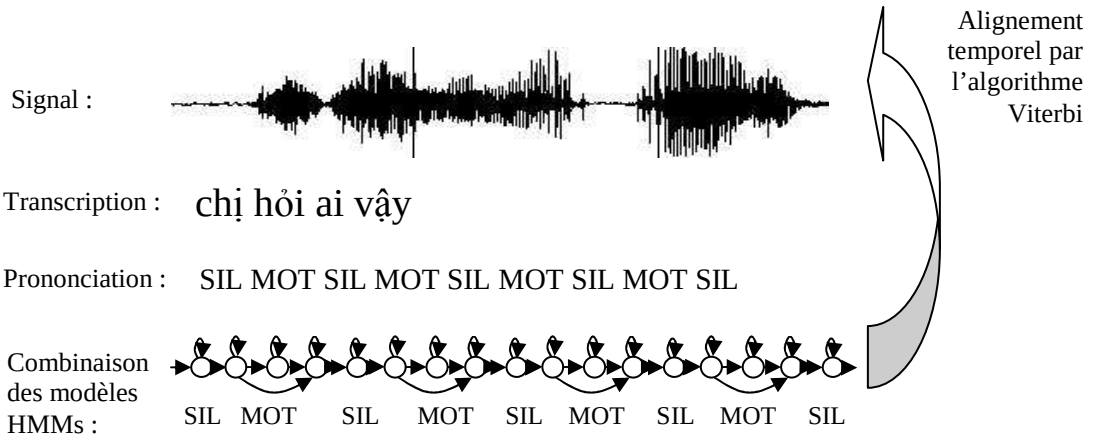


Figure 3.19 : Décodage les frontières des mots et silences par l'algorithme Viterbi

La figure 3.20 illustre un résultat de détection de frontières des mots dans notre expérimentation pour une phrase en vietnamien « *Cám ơn anh !* ». Nous trouvons que les modèles « mots/silence » peuvent détecter la frontière des mots.

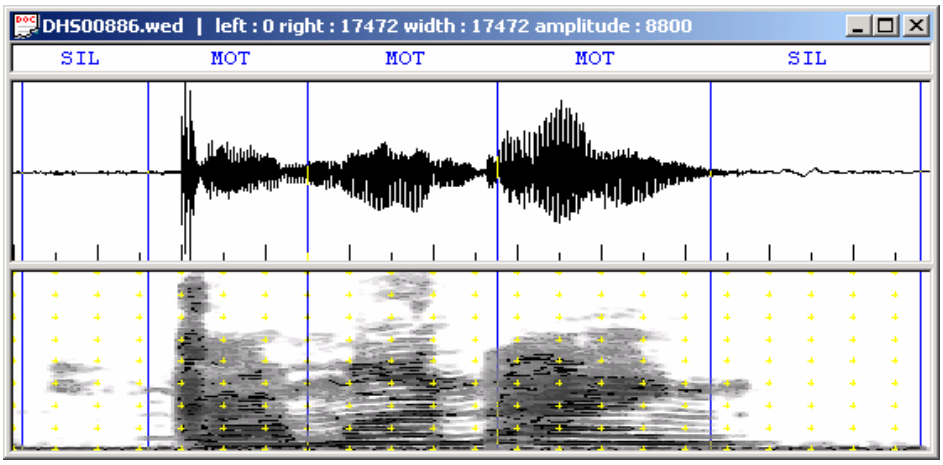


Figure 3.20 : Exemple de détection de frontières des mots pour la phrase « cám ơn anh »

6.1.2. Initialisation et apprentissage de modèles acoustiques graphémiques

Pour une phrase en entrée, nous pouvons, en utilisant les modèles « mot/silence » construits dans la section précédente, décoder les frontières des mots par l'algorithme d'alignement temporel Viterbi. Ensuite, chaque morceau de signal correspondant à un mot est segmenté uniformément suivant le nombre de graphèmes qui composent la transcription de ce mot (selon la figure 3.16). Ceci constitue l'initialisation pour l'entraînement de nos modèles acoustiques à base de graphèmes.

Après avoir appris des modèles acoustiques initiaux avec l'étiquetage de départ, nous utilisons ces modèles pour ré-aligner temporellement les signaux d'apprentissage. Ensuite, les modèles acoustiques graphémiques sont ré-entraînés (apprentissage à base d'étiquettes) et nous ré-itérons le cycle. La procédure d'apprentissage est arrêtée quand le système atteint un état stable, normalement après 6-8 itérations de la procédure d'apprentissage.

La figure 3.21 illustre un résultat d'alignement temporel par les modèles acoustiques graphémiques après 6 itérations d'apprentissage.

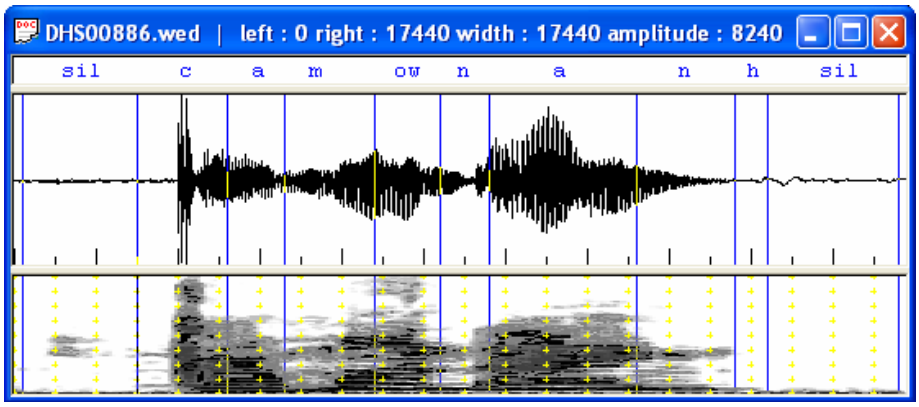


Figure 3.21 : Exemple d'alignement temporel par des modèles acoustiques graphémiques pour la phrase « cám ơn anh »

Pour montrer l'efficacité de notre méthode d'initialisation, nous avons effectué une comparaison de deux méthodes d'initialisation des modèles graphémiques :

- méthode « baseline » : initialisation uniforme ;

- notre méthode d'initialisation à base de détection de frontières des mots.

Le tableau 3.3 présente les taux d'exactitude en syllabes des méthodes d'initialisation et du système de référence à base de phonèmes (système phonémique). Les détails expérimentaux concernant ce résultat sont presentés dans le chapitre suivant.

Langue	**Système phonémique**	**Système graphémique**	
		Initialisation uniforme	Initialisation à base de détection de frontières de mots
Vietnamien	57,4 %	31,5 %	44,4 %

Tableau 3.3 : Comparaison des taux d'exactitude en syllabes de deux méthodes d'initialisation des modèles acoustiques graphémiques

À partir des résultats obtenus, nous constatons que notre méthode d'initialisation des modèles graphémiques à partir des modèles « *mot/silence* » améliore les performances par rapport à une initialisation uniforme.

6.1.3. Portabilité vers une nouvelle langue

Les sections précédentes abordent notre méthode d'initialisation de modèles acoustiques graphémiques à base de détection de frontières des mots. Nous supposons (et avons pu vérifier) que nos modèles « mot/silence » sont relativement indépendants de la langue. Pour une nouvelle langue, nous pouvons appliquer des modèles « mot/silence » appris sur une autre langue pour détecter la frontière des mots.

En effet, nous avons effectué des expérimentations « *crosslingues* » vietnamien / khmer. Pour une phrase khmère en entrée, nous pouvons, en utilisant les modèles « mot/silence » construits précédemment pour le vietnamien, décoder la frontière des mots par l'algorithme de Viterbi. Par exemple, la figure 3.23 illustre un résultat de détection de frontières des mots pour une phrase en khmer en utilisant les modèles « crosslingues » vietnamien/khmer. Bien qu'il reste encore des erreurs de détection, la plupart de frontières des mots et silences sont bien détectées.

Ensuite, chaque morceau de signal correspondant à un mot est segmenté uniformément suivant le nombre de graphèmes du mot. Ces segments de données vocales ainsi obtenus sont alors utilisés pour apprendre les modèles graphémiques correspondants du khmer.

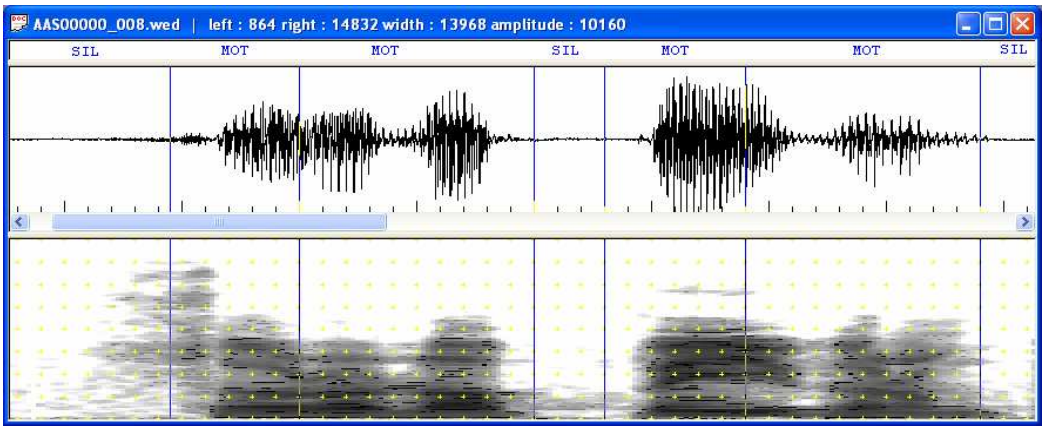


Figure 3.22 : Exemple de détection de frontières des mots en khmer à l'aide des modèles « mot/silence » en vietnamien

6.2. Modélisation acoustique graphémique dépendante du contexte

Pour la modélisation de graphèmes dépendant du contexte, l'unité de modélisation est appelée un *polygraphème*. Similairement à la modélisation de phonèmes, l'une des difficultés de la modélisation dépendante du contexte à base de graphèmes est la génération des questions linguistiques afin de regrouper les unités « similaires » dans un modèle acoustique en utilisant un arbre de décision. Nous utilisons deux méthodes de génération d'arbre de décision :

- Méthode de « singleton » ;
- Méthode à base de relation « graphème - phonème ».

Avec la méthode de « singleton », la génération d'un arbre de décision de *polygraphèmes* consiste simplement à demander quel est le graphème du contexte gauche ou droit (figure 3.23). Chaque question linguistique consiste en un seul graphème. En conséquence, l'ensemble de questions linguistiques s'appelle singleton (pour plus de détail, voir [Killer 2003a]).

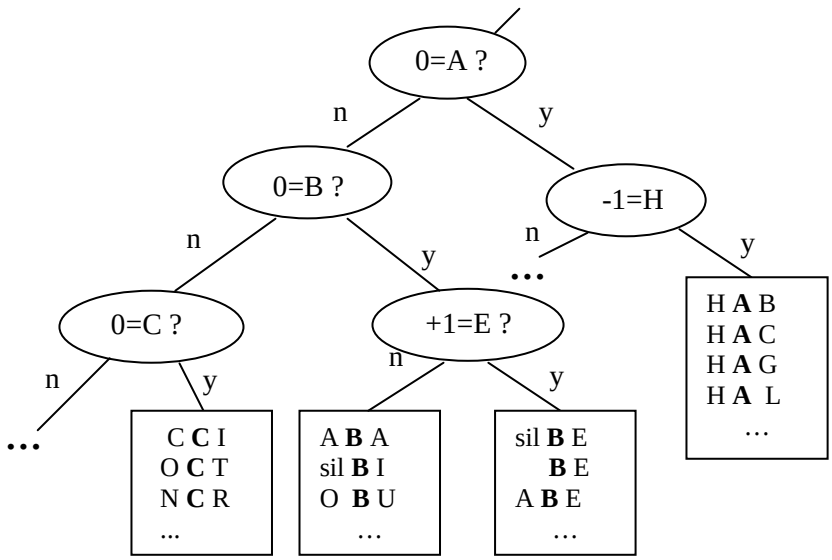


Figure 3.23 : Arbre de décision des graphèmes en français

Avec la méthode à base de relation « graphème-phonème », la classification des graphèmes est basée sur la classification des phonèmes correspondants. Alors, l'ensemble des questions phonémiques du système à base de phonèmes est converti en remplaçant chaque phonème par les graphèmes correspondants. A titre d'exemple, le tableau 3.4 présente un exemple de conversion des phonèmes dans l'ensemble de questions phonémiques du vietnamien aux graphèmes correspondants.

Questions linguistiques	Phonèmes	Graphèmes
SILENCES	SIL	SIL
BILABIAL	p b m	p b m
LABIODENTAL	f v w	v
ALVEOLAR	t tʰ d n s z l	t d đ n x l s r
RETROFLEX	ʈ ʂ ʐ	s r
VELAR	k χ ŋ ɣ	c k g
GLOTTAL	h	h
PLOSIVE	p b t tʰ d ʈ c k	p b t đ k c h
NASAL	m n ɲ ŋ	m n
FRICATIVE	f v s z ʂ ʐ χ ɣ h	v s x d r k g h
APPROXIMANT	w j	u o i y
FRONT	i e ɛ ɛ̆ a ă	y i ê e a ă
BACK	ɯ u ɤ ɤ̆ o ɔ ɔ̆	uw u ơ â oo o
CLOSE	i ɯ u	y i ư u
CLOSE-MID	e ɤ ɤ̆ o	ê ơ â ô
OPEN-MID	ɛ ɛ̆ ɔ ɔ̆	e a o
OPEN	a ă	a ă
…	…	…

Tableau 3.4 : Conversion des questions phonème-graphème pour le vietnamien

Cependant, l'efficacité de la méthode à base de relation « graphème- phonème » dépend fortement de la connaissance phonétique et de la relation graphème-phonème pour la langue considérée.

Dans le chapitre IV, nous comparerons la performance de ces méthodes de génération d'arbre de décision sur la langue khmère. Le lecteur trouvera d'autres méthodes de génération d'arbre de décision graphémiques dans [Killer 2003a].

7. Conclusion du chapitre

Dans ce chapitre, nous avons présenté nos méthodes de recueil de ressources acoustiques et construction rapide de modèles acoustiques pour des langues peu dotées. Pour la collecte facilitée des signaux de parole en langue peu dotée, nous avons adapté la plateforme du logiciel EMACOP pour manipuler respectivement les caractères dans plusieurs langues (la plateforme

Unicode). Pour la construction d'un dictionnaire de prononciation dans une langue peu dotée, nous avons abordé trois approches automatiques différentes. La sélection de méthodes pertinentes pour chaque langue dépend fortement de la connaissance phonétique de la langue.

Le travail le plus important sur la portabilité des systèmes de reconnaissance automatique de la parole existant en langue source vers une langue cible consiste à déterminer les correspondances phonémiques source/cible. Nous avons ainsi proposé des méthodes d'estimation de similarité entre des unités phonémiques différentes telles que : le phonème (monophone), le groupe de phonèmes, le polyphone, etc. A base de correspondances phonémiques entre langue source et langue cible déterminées par nos méthodes, les modèles acoustiques crosslingues sont recueillis rapidement à partir des modèles acoustiques existants en langue source. Ces modèles acoustiques crosslingues peuvent être ensuite adaptés avec une quantité réduite de signaux en langue cible.

Pour les langues peu dotées qui ne possèdent pas un dictionnaire phonétique, l'approche de modélisation acoustique à base de graphèmes présente un potentiel intéressant pour construire rapidement un système de reconnaissance automatique de la parole. Nous avons proposé une méthode d'initialisation rapide de modèles acoustiques graphémiques à base d'un modèle acoustique crosslingue « mot/silence ».

Les différentes techniques proposées dans ce chapitre et dans le chapitre précédent vont maintenant être validées sur le vietnamien (chapitre 4 suivant) et le khmer (chapitre 5).

Chapitre 4

Application au vietnamien

Ce chapitre présente l'application au vietnamien de nos travaux sur la construction rapide d'un système de reconnaissance vocale. Tout d'abord, pour réaliser un système de reconnaissance automatique de la parole, une connaissance minimale des caractéristiques linguistiques et phonétiques de la langue vietnamienne est nécessaire. Ensuite, nous présentons nos travaux concernant la collecte de ressources textuelles et acoustiques pour le vietnamien. La modélisation du langage et les techniques de portabilité rapide des modèles acoustiques multilingues vers une nouvelle langue seront aussi présentées. Des résultats d'expérimentations et les comparaisons des performances des systèmes obtenus sont aussi détaillés dans ce chapitre. Pour finir, une technique de modélisation acoustique à base de graphèmes est proposée pour le vietnamien.

1. Phonologie et phonétique du vietnamien

1.1. Généralités

La langue vietnamienne est parlée par environ 67,4 millions de personnes au Vietnam et à l'étranger[1]. Son origine est toujours sujette à débat parmi les linguistes. Il est cependant généralement admis qu'elle a des racines communes fortes avec le môn-khmer qui fait partie de la branche austro asiatique. C'est une langue tonale qui possède six tons (dialecte standard du nord) [Nguyen 2002]. L'orthographe est latine depuis le XVIIè siècle, avec des caractères accentués pour les tons.

Le Vietnam est divisé en 3 régions dialectiques : les dialectes du Nord, du Centre et du Sud. Ceux qui habitent une même région peuvent parler vietnamien avec des accents différents, mais ces différences ne causent pratiquement pas de difficulté dans la communication. Par contre, d'une région à l'autre, la communication par la parole peut s'avérer difficile. Malgré les divergences de prononciation et de vocabulaire, on considère le vietnamien comme une seule et même langue dans la mesure où la communication est effectivement assurée [Pham 1969]. Par conséquent, la forme parlée considérée comme la plus recommandable par tous les vietnamiens est toujours celle qui est la plus proche de l'écriture officielle, c'est-à-dire, celle que parlent la plupart des vietnamiens du nord.

[1] On trouve notamment des communautés vietnamiennes en Australie, au Cambodge, au Canada, en Chine, en France, au Laos, et aux Etats-Unis (*source : Ethnologue 1999*)

1.2. Évolution historique de la langue vietnamienne

Nous présentons dans cette section l'évolution de la langue vietnamienne. En effet, il existe des écritures différentes et des dialectes divers en langue vietnamienne. L'histoire du Vietnam se traduit assez fidèlement à travers plusieurs étapes par l'histoire de la langue vietnamienne.

Le premier système d'écriture vietnamienne s'appelle *chữ nho* (ou *chữ Hán*). Ses caractères sont issus du système chinois, mais la prononciation est vietnamienne. En effet, l'écriture par idéogrammes a ceci de particulier que le même signe se prononce de façon différente suivant qu'il est lu par un chinois, un japonais ou un vietnamien. Dans l'ancien Vietnam, pendant de longs siècles, du I^{er} siècle avant Jésus-Christ au X^{e} siècle après Jésus-Christ, la langue parlée et écrite du chinois classique (le *Hán*) dominait tout le pays alors que la langue du pays, n'ayant pas un système d'écriture, se développait en parallèle dans le peuple. En effet, le chinois n'a jamais été utilisé comme moyen de communication orale par la population vietnamienne [Nguyen-Thi 2000]. En revanche, il était employé pour tous les documents écrits officiels.

Le deuxième système d'écriture est le *chữ nôm* (écriture démotique), système de transcription des mots vietnamiens au moyen de caractères chinois simples ou combinés entre eux pour noter soit uniquement le son soit le sens et le son combinés d'un mot vietnamien. À partir du X^{e} siècle, marquant le début d'une période d'indépendance, les vietnamiens inventent une écriture propre pour transcrire les mots de leur langue vernaculaire. C'est la naissance du *chữ nôm*, écriture démotique issue de l'écriture chinoise *Hán* et représentant une transcription mi-phonétique mi-idéographique du vietnamien. Les contacts fréquents entre le vietnamien et le *Hán***,** dont la prédominance existait toujours, ont eu pour conséquence inévitable l'existence d'un grand nombre d'emprunts *Hán* dans le lexique du vietnamien. La prononciation du chinois au Vietnam, éloigné du pays d'origine, aurait évolué selon les propriétés phonétiques propres au vietnamien. Les deux systèmes (le *chữ nho* et le *chữ nôm*) ont cohabité jusqu'au XX^{e} siècle : utilisation du *chữ nho* pour l'enseignement et les documents officiels mais la littérature populaire a toujours été écrite avec le *chữ nôm*.

Au XVII^{e} siècle, les missionnaires européens inventent une autre forme d'écriture, le *chữ quốc ngữ* (écriture de la langue nationale). C'est une écriture vietnamienne latinisée qui est une transcription du vietnamien parlé en signes alphabétiques empruntés aux langues d'origine latine (au portugais, à l'italien, à l'espagnol et au français), les langues maternelles des missionnaires. Ils ont utilisé 6 accents différents pour transcrire 6 tons vietnamiens. Le premier document imprimé avec cette nouvelle écriture est le *Dictionarium annamiticum, lusitanum et latinum* (dictionnaire vietnamien-portugais-latin) publié par Alexandre de Rhodes en 1651 [Pham 1969]. Aujourd'hui, le *chữ quốc ngữ* est utilisé dans toutes les circonstances de la vie et il est l'écriture de l'administration et l'éducation.

Concernant le lexique, le vietnamien moderne dispose toujours de mots empruntés au lexique chinois moderne. La prononciation des mots est soumise aux règles de correspondance qui régissent les emprunts sino-vietnamiens dits « *Hán - Việt* ». Parallèlement, on adopte aussi

des termes scientifiques et techniques français arrivés par voie de transcription phonétique. Par exemple : *bánh mì* (pain de mie), *vi-ô-lông* (violon), *bi-đông* (bidon), pê-đan (pédale), *xăng-đan* (sandales).

1.3. Système phonétique et structure syllabique du vietnamien

Malgré les différences phonologiques régionales, on peut établir une prononciation standard d'après l'orthographe traditionnelle. Par conséquent, la plupart des chercheurs en linguistique vietnamiens [Doan 1999] proposent le « parler » du Nord comme standard pour la phonétique du vietnamien, en ajoutant quelques sons manquants qui existent dans d'autres zones géographiques [Nguyen-Thi 2000].

1.3.1. Système phonétique du vietnamien

La langue vietnamienne possède 23 consonnes. Le tableau 4.1 présente l'Alphabet Phonétique International (API) [IPA 1999] pour les consonnes du vietnamien qui sont classifiées selon la position d'articulation, le type d'articulation, le voisement,…

	Bilabiales	Labio-dentales	Dentales	Alvéolaires	Post-alvéolaire	Rétroflexes	Palatales	Vélaires	Uvulaires	Pharyngales	Glottales
Occlusives	p b		t tʰ d			ʈ	c	k			ʔ
Nasal	m		n				ɲ	ŋ			
Trilles											
Battues											
Fricatives		f v		s z		ʂ ʐ		ɣ	χ		h
Fricatives latérales											
Approx.											
Approx. latérales			l								

Tableau 4.1 : Classification des consonnes vietnamiennes

Contrairement aux personnes habitant dans d'autres régions au Vietnam, les vietnamiens du nord ne distinguent pas quelques couples de consonnes : s/ʂ, c/ ʈ, z/ʐ dans la langue parlée. Cela signifie que ces couples de phonèmes sont prononcés identiquement malgré le fait qu'ils sont distincts dans la langue écrite.

Il y a 9 voyelles longues, 4 voyelles courtes (figure 4.1) et 3 diphtongues pour le vietnamien. Une voyelle longue peut apparaître individuellement sans son final mais une voyelle courte est toujours combinée avec un son final.

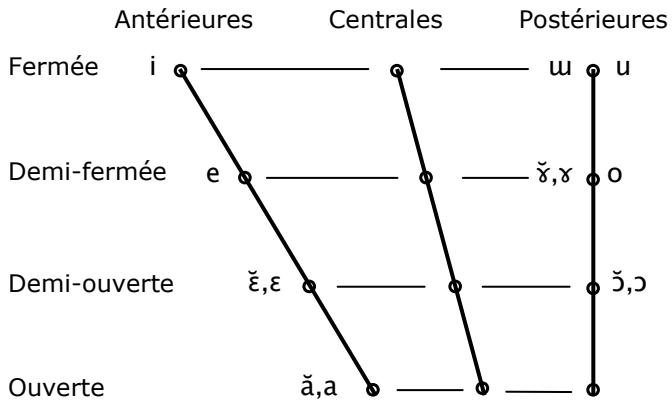


Figure 4.1 : Lieu d'articulation des voyelles vietnamiennes

1.3.2. Structure syllabique du vietnamien

La notion de syllabe est fréquemment utilisée dans le traitement de nombreux problèmes linguistiques. La syllabe est définie comme une suite de segments sonores dont le groupement est soumis à des contraintes particulières.

D'autre part, la langue vietnamienne est un des exemples les plus typiques de langue isolante, dans laquelle chaque mot a une forme unique et ne peut pas être modifié par dérivation ou flexion [Nguyen-Thi 2001, Nguyen 2004]. La syllabe de la langue vietnamienne a une position linguistique très importante. La syllabe est prise pour unité morphologique et phonologique de base dans l'analyse phonologique.

Au niveau phonologique, la structure complète d'une syllabe du vietnamien (une syllabe avec ton) consiste en cinq sous-parties [Doan 1999] : son initial (une consonne), prétonal (une semi-voyelle), noyau (une voyelle ou une diphtongue), coda (une consonne ou une semi-voyelle) et ton. Sauf la consonne initiale (qui s'appelle partie **INITIALE**), le reste de la syllabe (prétonal, noyau et coda) s'appelle partie **RIME** ou **FINALE** (voir la figure 4.2). Le ton vietnamien n'influence que la partie FINALE de la syllabe [Tran 2005].

<table>
<tr><td colspan="4">SYLLABE AVEC TON (6 686)</td></tr>
<tr><td colspan="4">SYLLABLE DE BASE (2 376)</td></tr>
<tr><td rowspan="3">SON INITIAL (22)</td><td colspan="3">TON (6)</td></tr>
<tr><td colspan="3">RIME (155)</td></tr>
<tr><td>Prétonal (1)</td><td>Noyau (16)</td><td>Coda (8)</td></tr>
</table>

Figure 4.2 : Structure phonologique d'une syllabe en vietnamien avec le nombre d'occurrences différentes existant pour chaque unité phonétique

Dans le tableau 4.2, nous présentons la liste des 21 consonnes qui peuvent être des sons initiaux de la syllabe auxquelles on ajoute le cas qui représente l'absence de son initial dans une syllabe que nous appellerons *phonème zéro.* Le système des sons initiaux du vietnamien contient donc 22 phonèmes différents [Doan 1999, Nguyen 2002].

API	Caractère	API	Caractère
ŋ	ngh, ng	t	t
tʰ	th	b	b
ʈ	tr	m	m
c	ch	n	n
ɲ	nh	l	l
f	ph	s	x
χ	kh	ʂ	s
z	d, gi	h	h
ɣ	g, gh	v	v
k	c, k, q	ʐ	r
d	đ	ʔ	

Tableau 4.2 : Consonnes initiales vietnamiennes

La semi-voyelle labiale /w/ de la langue vietnamienne est considérée comme un son prétonal (tableau 4.3). Cependant, l'apparition de ce son prétonal dans la structure de la syllabe est facultative et il n'apparaît jamais derrière les consonnes initiales /b, m, f, v, n, ʐ/.

API	Caractère
w	o, u

Tableau 4.3 : Prétonal vietnamien

Le tableau 4.4 présente les voyelles courtes, les voyelles longues et les diphtongues vietnamiennes. Chacune des trois diphtongues se compose de deux éléments vocaliques inséparables. Les voyelles et diphtongues vietnamiennes jouent un rôle de son noyau ou son principal de la syllabe. Par conséquent, l'apparition d'un noyau dans une syllabe est obligatoire.

API	Caractère	API	Caractère
i	i, y	ɔ	o
ɯ	ư	ɤ̆	â
u	u	ɛ̆	a(anh, ach)
e	ê	ă	ă, a(au, ay)
ɤ	ơ	ɔ̆	o(ong, oc)
o	ô	ie	iê, ia, yê, ya
ɛ	e	ɯɤ	ươ, ưa
a	a	uo	uô, ua

Tableau 4.4 : Voyelles et diphtongues vietnamiennes

Le son final (coda) est l'élément qui caractérise le mode de terminaison de la syllabe vietnamienne et il est étroitement lié au noyau de la syllabe. La coda de la syllabe contient 8 phonèmes dont 6 consonnes et 2 semi-voyelles (tableau 4.5). Cependant, l'apparition d'une coda pour une syllabe est facultative.

API	Caractère	API	Caractère
p	p	n	n
t	t	ŋ	nh, ng
k	ch, c	j	i, y
m	m	w	o, u

Tableau 4.5 : Sons finaux (coda) vietnamiens

Le prétonal, le noyau et la coda constituent une rime dans la structure syllabique vietnamienne. La liste des 155 rimes du vietnamien peut être trouvée dans [Nguyen 2002].

Dans une langue tonale comme le vietnamien, les tons influencent la structure syllabique, car toutes les syllabes vietnamiennes possèdent l'un de ces six tons. Ainsi, nous ne pouvons pas étudier les tons sans nous passer de l'analyse de la structure syllabique du vietnamien. Le tableau 4.6 présente les 6 tons vietnamiens.

API	Signe diacritique	Nom français	Exemple
1		ton plat	ba
2	`	ton descendant	bà
3	̉	ton interrogatif	bả
4	~	ton retombant	bã
5	´	ton montant	bá
6	̣	ton grave	bạ

Tableau 4.6 : Tons vietnamiens

La figure 4.3 présente très schématiquement l'évolution temporelle des 6 tons et leur hauteur relative. Ce schéma a été calculé par N-D. Andreev et M-V. Gordina [Doan 1999]. Les analyses, les caractéristiques et les méthodes de reconnaissance des tons vietnamiens sont présentées en détail dans [Doan 1999] et [Nguyen 2002].

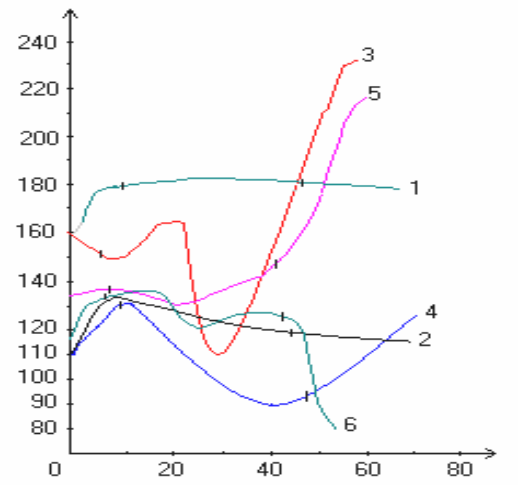


Figure 4.3 : Evolution temporelle des tons du vietnamien [Doan 1999]

2. Recueil de ressources linguistiques

2.1. Vocabulaire

Dans notre travail, un vocabulaire peut être défini comme une liste close d'unités lexicales qui peuvent être reconnues par un système de reconnaissance automatique de la parole. La taille du vocabulaire et la sélection des unités lexicales dans le vocabulaire influencent fortement les performances du système de transcription automatique (la perplexité des modèles de langages, l'espace de recherche, le taux de reconnaissance, …) puisque toutes les unités hors-vocabulaire ne peuvent pas être reconnues par le système de reconnaissance.

L'écriture vietnamienne étant segmentée en syllabes mais pas en mot (sauf pour quelques mots empruntés franco-vietnamiens), il existe toujours une espace entre deux syllabes dans une phrase (figure 4.4). La segmentation en mots n'est donc pas triviale pour la langue vietnamienne.

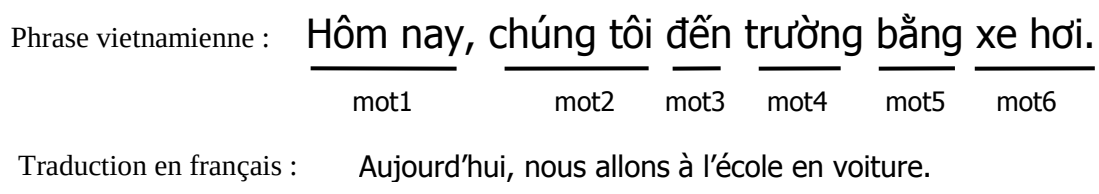


Figure 4.4 : Langue vietnamienne – langue segmentée en syllabes

Pour obtenir des ressources lexicales répondant à nos besoins, nous avons utilisé les ressources issues de communautés « ouvertes » en ligne. Ces ressources sont cependant très variées en format, en quantité, en qualité, en information linguistique, ... A titre d'exemple, dans le contexte du projet Papillon[1], une base lexicale multilingue comprenant entre autres l'allemand, l'anglais, le français, le japonais, le malais, le lao, le thaï, le vietnamien et le chinois est en cours de construction. Ce projet est totalement ouvert pour permettre l'élargissement arbitraire du nombre de langues, et mettre en œuvre un schéma de construction coopérative et d'utilisation mutualisée [Boitet 2001]. À partir du projet Papillon, nous avons récupéré un dictionnaire bilingue franco-vietnamien.

Pour développer un système de reconnaissance automatique de la parole, le dictionnaire bilingue est filtré pour obtenir finalement une liste de mots (un vocabulaire) de plus de 40 000 mots en vietnamien. Le vocabulaire obtenu contient des mots composés, des mots isolés et des mots empruntés franco-vietnamien.

Dans le contexte de construction d'un système de reconnaissance automatique de la parole en langue vietnamienne, nous avons identifié deux pistes possibles :

- la première consiste à faire un système de reconnaissance syllabique où les cooccurrences modélisées sont des suites de syllabes ;

[1] http://www.papillon-dictionary.org

- la seconde consiste à faire un système de reconnaissance de mots où les cooccurrences modélisées sont des suites de mots.

En fonction de ces deux pistes, nous devons construire deux types de vocabulaires : un vocabulaire de syllabes et un vocabulaire de mots.

2.1.1. Vocabulaire de syllabes

Pour générer un vocabulaire de syllabes, nous avons filtré le vocabulaire de 40 000 mots en vietnamien à partir d'un dictionnaire franco-vietnamien issu du projet Papillon. La taille du vocabulaire de syllabes est de 6 686 syllabes. Nous notons que ce vocabulaire couvre toutes les syllabes de la langue servant à constituer des mots (mots isolés et mots composés) de la langue vietnamienne. Le vocabulaire de syllabes obtenu est comparable au vocabulaire de syllabes vietnamiennes publié dans [Hoang 2004]. Comme cela est dit dans le chapitre 2 – section 1.2, le vocabulaire de syllabes est enrichi en ajoutant 114 nouvelles formes lexicales les plus fréquentes observées sur un grand corpus de texte. On obtient enfin un vocabulaire de syllabes de 6 800 entrées lexicales.

2.1.2. Vocabulaire de mots

Les mots vietnamiens sont soit monosyllabiques (mots simples), ou plurisyllabiques (mots composées). La figure 4.5 présente la statistique sur le nombre de mots vietnamiens selon le nombre de syllabes par mot dans un vocabulaire de mots vietnamiens [Dinh 2003]. Nous constatons que les mots bisyllabiques (2 syllabes par mot) possèdent un grand pourcentage de mots du vocabulaire (environ 70%). A cause de la grande fréquence des mots bisyllabiques et multisyllabiques, la segmentation des textes en vietnamien est compliquée et il y a beaucoup d'ambiguïtés de segmentation.

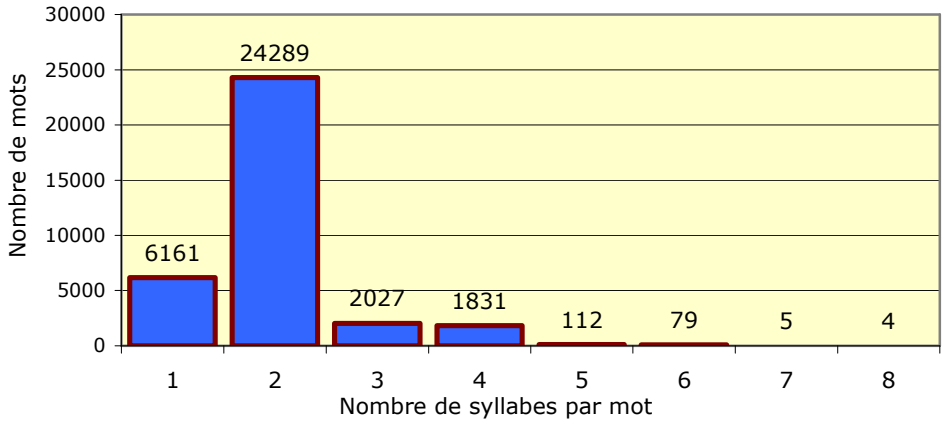


Figure 4.5 : Nombre de mots vietnamiens selon le nombre de syllabes par mot

D'autre part, la langue vietnamienne comprend dans son vocabulaire un grand nombre d'homophones ou de mots ayant plusieurs significations. Pour distinguer entre eux les homophones, comme pour distinguer les différentes significations d'un même mot, la langue vietnamienne recours à des mots composés, soit en juxtaposant deux synonymes, soit en ajoutant au mot simple une syllabe asémantique [Truong 1970].

Nous avons effectué une analyse de la couverture lexicale du vocabulaire de 40 000 mots en vietnamien sur un grand corpus de texte collecté à partir du Web. La figure 4.6 présente la couverture de mots monosyllabiques et plurisyllabiques dans le corpus de texte. Bien que les mots bisyllabiques possèdent un grand pourcentage dans un vocabulaire (environ 70%), le nombre de mots bisyllabiques n'est que 21,2% dans le corpus de texte. Les mots monosyllabiques sont ainsi les plus utilisés.

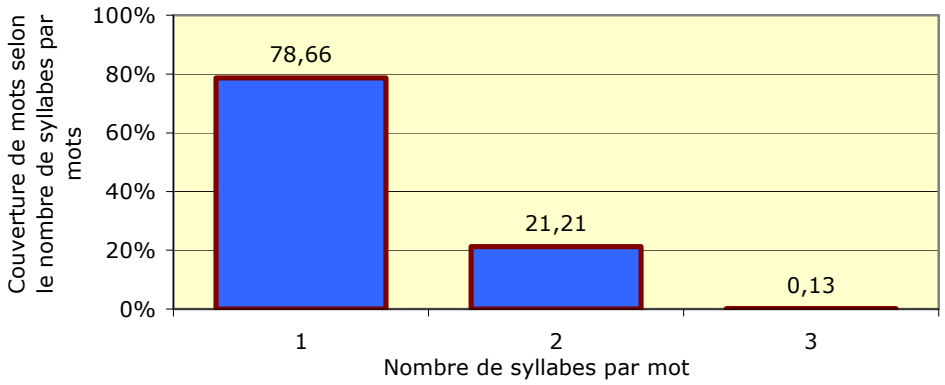


Figure 4.6 : Couverture de mots monosyllabiques et plurisyllabiques présentée dans un grand corpus de texte du Web

Par ailleurs, les 20 000 mots les plus fréquents du corpus de texte couvrent déjà 99,5% des mots du corpus de texte collecté sur le Web (figure 4.7). Par conséquent, dans notre expérimentation, nous avons décidé de réduire la taille du vocabulaire à 20 000 mots en gardant les mots les plus fréquents du grand corpus de texte.

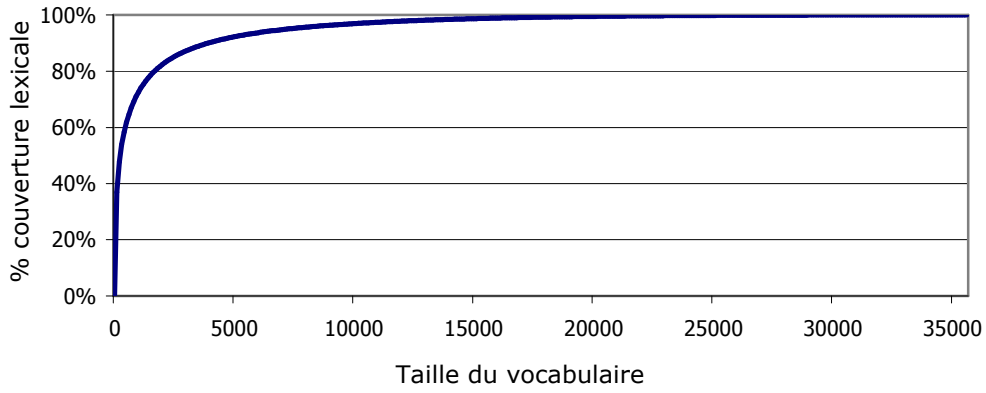


Figure 4.7 : Couverture lexicale sur le corpus tiré du Web selon la taille du vocabulaire

2.2. Recueil d'un corpus de texte à partir du Web

Notre méthodologie de récupération rapide d'un corpus de texte à partir du Web a été appliquée au vietnamien. La quantité de pages Web collectées était de 2,5 Go. Après filtrage, la quantité de données textuelles pouvant servir à l'apprentissage d'un modèle de langage statistique était d'environ 400 Mo (5 millions de phrases). A titre de comparaison, une année complète du journal « Le Monde » en français correspond à 120 Mo en moyenne.

Par ailleurs, la langue vietnamienne possède une écriture segmentée en syllabe avec une grande fréquence des mots composés (ou mots polysyllabiques) dans le vocabulaire de mots. Ainsi, la segmentation d'une phrase en mots n'est pas triviale. En effet, pour résoudre le problème de la segmentation en mots, une méthode de segmentation à base d'automates d'états finis est appliquée sur un corpus de texte vietnamien [Nguyen 2004]. Dans le cas d'une ambiguïté, une décision humaine est nécessaire. Une autre méthode appliquée au vietnamien consiste à utiliser un algorithme d'apprentissage, à base d'un réseau de neurones [Dinh 2001].

Dans un premier temps, nous avons essayé deux algorithmes à base de vocabulaire des mots : la segmentation « plus longue chaîne d'abord » (*longest matching*) qui maximise la taille des mots et la segmentation « en plus petit nombre de mots » (*maximal matching*). Dans le futur, nous envisageons d'améliorer la segmentation en mots à l'aide d'un algorithme probabiliste, à l'aide de modèles de langage syllabiques.

3. Solutions pour la modélisation statistique du langage

3.1. Filtrage des informations redondantes

Comme cela est présenté dans le chapitre 2 - section 4.2.2, les pages récupérées sur le Web peuvent contenir de l'information redondante (menus, références, annonces,...) qui est répétée dans plusieurs pages. Cette redondance a une influence sur la qualité du corpus de texte collecté (revoir la figure 2.11 du chapitre 2).

En fait, nous avons estimé l'influence de ces informations redondantes sur des documents html collectées sur un site Web de nouvelles en vietnamien. En filtrant les informations redondantes, la taille du corpus de texte est réduite de 54% dans notre expérimentation. Par ailleurs, la perplexité des modèles de langage construits sur ce corpus de texte est améliorée significativement de 26% en supprimant les informations redondantes [Le 2003b].

Par conséquent, pour obtenir un corpus de texte de meilleure qualité, les informations redondantes contenues dans les documents html doivent être enlevées et filtrées.

3.2. Comparaison des filtrages des phrases et des expressions langagières

Concernant les modèles de langage du vietnamien, nous avons choisi deux systèmes possibles :

- système à base de syllabes où les cooccurrences modélisées sont des suites de syllabes ;
- système à base de mots où les cooccurrences modélisées sont des suites de mots. Dans ce cas, le corpus de texte doit être segmenté en mots par un outil de segmentation à base d'un vocabulaire.

De plus, pour un corpus de texte, nous essayons des méthodes différentes de filtrage des phrases comme celles présentées dans le chapitre 3 :

- **filtrage 1 :** prendre toutes les phrases (ne pas appliquer de filtrage) ;
- **filtrage 2 :** prendre les phrases entières ayant au moins *N* mots et dont tous les mots appartiennent au vocabulaire ;
- **filtrage 3 :** prendre une séquence consécutive (un bloc) d'au moins *M* mots appartenant au vocabulaire ;
- **filtrage 4 :** utiliser une méthode hybride qui consiste à prendre les phrases entières ayant au moins *N* mots appartenant au vocabulaire (filtrage 2) et appliquer le filtrage par blocs minimaux de taille *M* (filtrage 3) sur les phrases rejetées.

Après avoir filtré le corpus de texte à partir du Web, les tailles du corpus de texte obtenu par ces méthodes de filtrage avec les tailles du vocabulaire correspondant (le nombre de mots différents) sont présentées dans le tableau 4.7.

Unité du vocabulaire	Nombre d'unités	**Filtrage 1**	**Filtrage 2** (*N*=1)	**Filtrage 3** (*N*=1)		**Filtrage 4** (*N*=1)	
				M=3	*M*=5	*M*=3	*M*=5
Mot	Nombre de mots	4 179 981	2 680 835	4 310 321	3 520 238	4 534 042	4 109 280
	Nombre de mots différents	36 364	20 000	20 000	20 000	20 000	20 000
Syllabe	Nombre de syllabes	4 180 206	2 839 776	4 338 904	3 718 052	4 501 649	4 185 950
	Nombre de syllabes différentes	6 616	6 581	6 616	6 613	6 616	6 616

Tableau 4.7 : Taille du corpus de texte selon la méthode de filtrage

Pour apprendre les modèles de langage, nous utilisons la boîte à outils SRILM [Stolcke 2002] en utilisant la méthode de Good-Turing pour le lissage avec le repli de Katz [Katz 1987]. Le tableau 4.8 présente la taille des modèles de langage (1-gramme, 2-grammes et 3-grammes) selon les méthodes de filtrages utilisées et le type d'unité choisie dans le vocabulaire.

Unité du vocab.	ML	**Filtrage 1**	**Filtrage 2** (*N*=1)	**Filtrage 3** (*N*=1)		**Filtrage 4** (*N*=1)	
				M=3	*M*=5	*M*=3	*M*=5
Mot	2-grammes	5 885 253	4 165 397	5 517 800	5 388 562	5 527 776	5 462 327
	3-grammes	5 125 710	3 305 022	5 039 753	4 856 487	5 061 023	4 978 587
Syllabe	2-grammes	2 676 426	2 282 867	2 671 418	2 641 140	2 672 784	2 658 875
	3-grammes	5 823 122	4 296 780	5 811 029	5 715 107	5 819 752	5 782 900

Tableau 4.8 : Nombre de n-grammes des modèles de langage selon les méthodes de filtrages utilisées et l'unité choisie dans le vocabulaire

Pour évaluer des modèles de langage, nous estimons la perplexité sur un corpus de test de 136 phrases de type dialogue court, soit 1483 syllabes ou 1286 mots. Le tableau 4.9 présente les valeurs de perplexité (PPL) des modèles de langage obtenus et le taux d'unités hors vocabulaire (MHV) du corpus de test selon les modèles de langage et le type d'unité de vocabulaire choisis.

Unité du vocabulaire		**Filtrage 1**	**Filtrage 2** (*N*=1)	**Filtrage 3** (*N*=1)		**Filtrage 4** (*N*=1)	
				M=3	*M*=5	*M*=3	*M*=5
Mot	%MHV	0,5	0,5	0,5	0,5	0,5	0,5
	PPL	202,8	191,8	193,4	200,0	**191,7**	192,6
Syllabe	%MHV	0	0	0	0	0	0
	PPL	109,0	108,6	109,0	111,2	**108,3**	108,5

Tableau 4.9 : Perplexités des modèles trigrammes calculées sur le corpus de test

À partir des valeurs de perplexités, nous constatons que la méthode de filtrage 4 (méthode hybride) obtient la meilleure valeur de perplexité bien que les différences ne soient pas très significatives. Nous utiliserons toutefois les modèles de langage construits à partir de corpus issus de la méthode de filtrage 4 (avec la taille minimale de blocs égale à 3) pour évaluer notre système de reconnaissance final en vietnamien. Par ailleurs, le taux d'unités hors-vocabulaire dans les modèles de langage à base de syllabes est égal à 0% dans notre évaluation bien que ce taux dans les modèles à base de mots soit égal à 0,5%. Cela montre que le vocabulaire de 6 800 syllabes a une couverture lexicale complète de la langue vietnamienne.

Nous notons que les valeurs de perplexités plus faibles pour les modèles de langage de syllabes ne signifient pas forcement que les modèles de syllabes obtiendront un meilleur taux de reconnaissance que les modèles de langage de mots dans un système de reconnaissance automatique de la parole en vietnamien car la taille et l'unité du vocabulaire dans les deux modèles de langage sont totalement différentes. De plus, sur un même corpus d'apprentissage, un modèle trigramme de mots couvre plus d'informations qu'un modèle trigramme de syllabes. Par conséquent, la sélection de mot ou de syllabe comme unité de reconnaissance dans notre système de reconnaissance automatique de la parole en langue vietnamienne est fondée plutôt sur la couverture lexicale désirée et la taille du corpus de texte d'apprentissage du modèle de langage.

En effet, avec la taille du corpus de texte d'apprentissage que nous obtenons sur le vietnamien, des tests préliminaires sur le système de reconnaissance vocale montrent que les modèles de langage de syllabes sont meilleurs que les modèles de mots. Par conséquent, nous choisissons la syllabe comme unité de reconnaissance dans des expérimentations de reconnaissance automatique de la parole en langue vietnamienne. Ainsi, le système de reconnaissance de la parole à base de syllabes peut être considéré comme *un système de reconnaissance automatique de la parole continue à grand vocabulaire (LVCSR).*

4. Construction d'un dictionnaire phonétique à base de règles

Un dictionnaire de prononciation (ou dictionnaire phonétique) est une ressource essentielle aux tâches de synthèse et de reconnaissance automatique de la parole. Mais au début de ce travail, il n'existait à notre connaissance aucun dictionnaire phonétique sous forme électronique pour le vietnamien. Par conséquent, nous avons construit un dictionnaire qui n'est pas utilisé seulement pour nos travaux mais également pour d'autres applications dans le domaine du traitement de la langue vietnamienne.

Dans notre travail, un analyseur phonétique (*VNPhoneAnalyzer*) [Le 2004] à base de règles a été développé pour obtenir automatiquement un dictionnaire de prononciation vietnamien.

Tout d'abord, à partir du vocabulaire de 6 686 syllabes, nous avons extrait les 22 parties initiales (les consonnes initiales), 155 parties finales (les rimes) et 6 tons de la syllabe vietnamienne. Ce nombre relativement réduit d'unités (183) a pu être phonétisé manuellement en utilisant les symboles de l'Alphabet Phonétique International (figure 4.8). Les correspondances graphème-phonème de ces 183 unités sont ensuite placées dans un tableau de référence que nous nommons *« IPA Reference Table for Sub-word Units »* (ou simplement le tableau IPATU). Nous notons que chaque unité est phonétisée par une seule prononciation. C'est-à-dire, nous n'intégrerons pas des variantes de prononciation dans le dictionnaire phonétique.

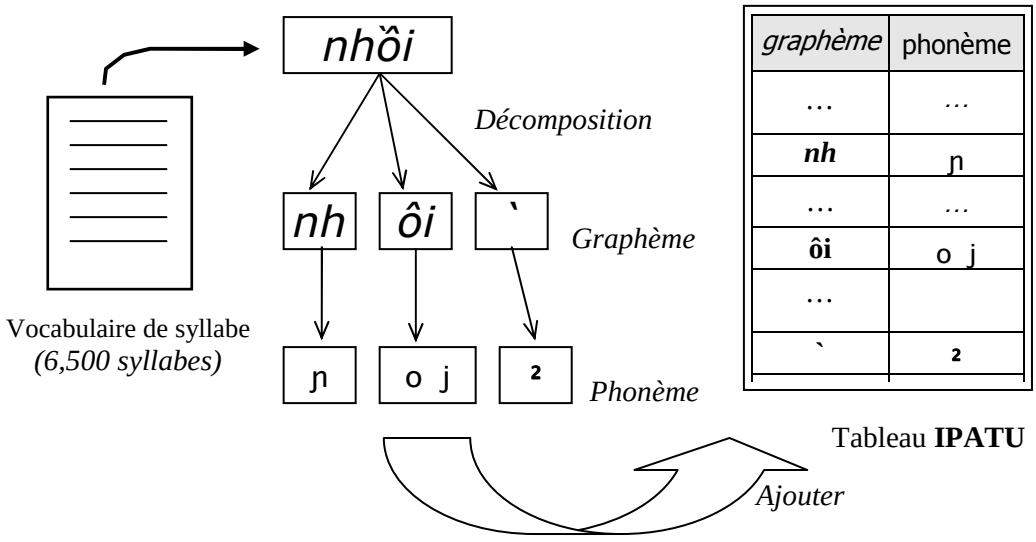


Figure 4.8 : Transcription manuelle des 22 parties initiales, 155 parties finales et 6 tons constituant les syllabes du vietnamien

À partir du tableau IPATU, nous avons construit un phonétiseur automatique (appelé *VNPhoneAnalyzer*). Ce phonétiseur utilise un algorithme de concaténation des unités phonétiques qui est décrit dans la figure 4.9.

Entrée : Soit W une syllabe tonale à phonétiser.

Sortie : La transcription phonétique de cette syllabe : P_W

Début

1. Décomposer la syllabe W en partie initiale (I), partie finale (F) et ton (T) :

$$W \rightarrow I \mid F \mid T$$

2. En utilisant le tableau IPATU, nous pouvons consulter respectivement les transcriptions phonétiques P_I, P_F et P_T pour les parties I, F et T. Nous avons :

$$I \rightarrow P_I,\ F \rightarrow P_F,\ T \rightarrow P_T$$

3. Concaténer ces transcriptions phonétiques pour obtenir la transcription phonologique complète pour la syllabe W.

$$P_W \leftarrow P_I \mid P_F \mid P_T$$

4. Vérifier et corriger les cas particuliers. Par exemple :

gi → / z i / ou qua → / k w a /

Fin

Figure 4.9 : Algorithme de concaténation des unités phonétiques

La figure 4.10 illustre aussi l'algorithme de concaténation des unités phonétiques que nous avons utilisée dans notre phonétiseur. Par ailleurs, *VNPhoneAnalyzer* peut produire en sortie des formats différents de transcription : symboles API, numéros API, symboles SAMPA[1], symboles JANUS, …

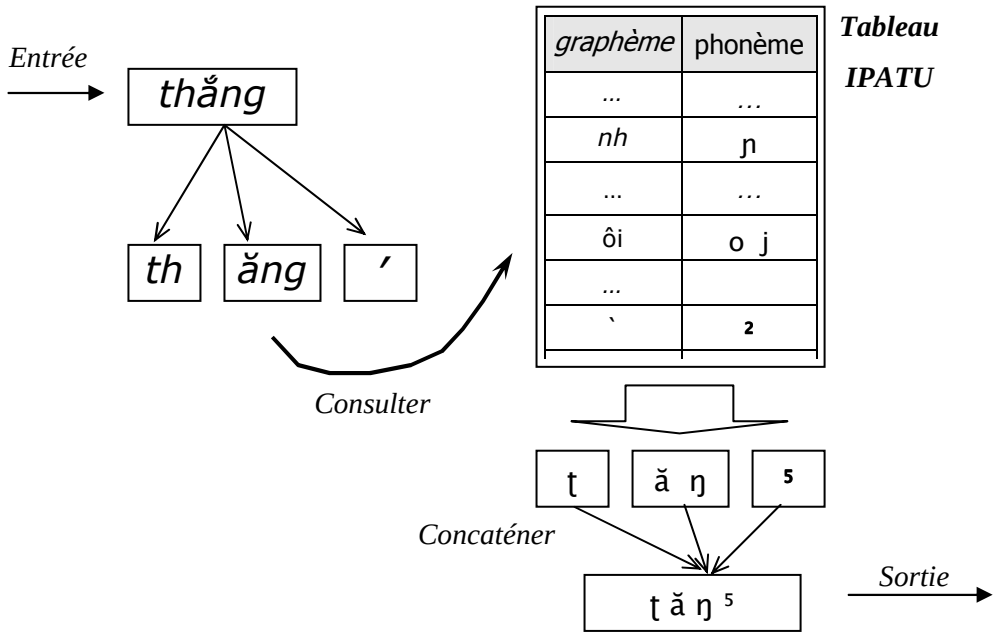


Figure 4.10 : Algorithme de concaténation des unités phonétiques dans VNPhoneAnalyzer

[1] http://www.phon.ucl.ac.uk/home/sampa

Ainsi, deux dictionnaires phonétiques du vietnamien sont construits en appliquant le *VNPhoneAnalyzer* sur toutes les entrées lexicales du vocabulaire de 6 686 syllabes et du vocabulaire de 40 000 mots vietnamiens. Pour les mots étrangers, une procédure de transcription manuelle est nécessaire. Le tableau 4.10 présente un exemple du dictionnaire phonétique à base de mots que nous avons généré.

Mot	Prononciation	Mot	Prononciation
a dua	a ¹ z uo ¹	chuyên gia	c w ie n ¹ z a ¹
am tường	a m ¹ t ɯɤ ŋ ²	dần	z ɤ̆ n ²
anh	ɛ̆ ŋ ¹	gỉ	z i ³
ban hành	b a n ¹ h ɛ̆ ŋ ²	giã	z a ⁴
ban trưa	b a n ¹ ʈ ɯɤ ¹	hồng quân	h o ŋ ² k w ɤ̆ n ¹
bút pháp	b u t ⁵ f a p ⁵	mạnh thường quân	m ɛ̆ ŋ ⁶ tʰ ɯɤ ŋ ² k w ɤ̆ n ¹
bơi thuyền	b ɤ j ¹ tʰ w ie n ²	quạng	k w a ŋ ⁶
bằng	b ă ŋ ²	rập khuôn	ʐ̩ ɤ̆ p ⁶ χ uo n ¹

Tableau 4.10 : Exemple du dictionnaire phonétique vietnamien

Enfin, ces dictionnaires phonétiques ont été vérifiés par des experts de l'Institut Linguistique du Vietnam [Le 2004].

5. *VNSpeechCorpus :* un corpus de parole en vietnamien

Un corpus de parole vietnamien est toujours en cours d'enregistrement au Centre MICA. En 2005, il contenait 39 locuteurs, 19 femmes et 20 hommes, venant des régions nord, centre et sud du Vietnam. Chaque locuteur a enregistré environ 1 heure de parole, ce qui fait un total de 39 heures. Le corpus contient non seulement des séquences de phonèmes, de nombres et de mots isolés, mais aussi la lecture de phrases complètes et de paragraphes.

Dans cette section, nous présentons la structure du corpus *VNSpeechCorpus*, la collection d'énoncés, l'enregistrement du corpus, l'évaluation et l'utilisation du corpus obtenu. Des détails supplémentaires sur les ressources collectées pour le vietnamien se trouvent dans [Tran 2003] et [Le 2004].

5.1. Organisation du corpus

VNSpeechCorpus comprend 5 types de données différentes :

- phonèmes ;
- mots avec six tons différents ;

- chiffres et nombres ;
- commande ;
- phrases de dialogue et paragraphes courts de texte.

Les phonèmes sont lus par tous les locuteurs. Les voyelles et les diphtongues peuvent être lus indépendamment sauf les voyelles *ă* [ă] et *â* [ɤ̆], parce qu'ils n'existent que dans le contexte d'un mot, par exemple : *ngă̲n* (court), *tâ̲n* (nouveau)... Pour les consonnes, nous les combinons avec la voyelle *ơ* [ɤ] dans les énoncés.

Une syllabe peut être combinée avec chacun des six tons de la langue, ce qui lui donne alors six significations différentes [Doan 1999], par exemple : *ba* (trois), *bà* (grand-mère), bá (roi), *bả* (appas), bã (déchet), *bạ* (n'importe). Donc, les locuteurs prononcent également des mots avec des tons différents.

Le corpus de chiffres et nombres vietnamiens se compose des chiffres de 0 à 9 et de nombres comme le numéro téléphonique, le numéro de carte bancaire, etc. Dans le système numérique vietnamien, la plupart des chiffres sont lus de manière unique mais il y a quelques cas particuliers de synonymes comme les nombres se terminant par les chiffres 4 et 5 qui peuvent être lus de plusieurs façons différentes. Afin de couvrir tous les cas, le corpus se compose de toutes les variantes (synonymes) de ces chiffres.

Un ensemble de plus de 50 mots clés de commande (*application words*) est défini dans le corpus *VNSpeechCorpus*. Chaque mot correspond à une action qui est utilisée dans plusieurs applications telles que service vocal téléphonique, interface homme-machine, ...

Après avoir collecté et traité des paragraphes de texte et des phrases, le corpus de phrases est divisé en deux parties : une partie commune à tous les locuteurs et une partie privée. La partie commune contient 33 dialogues courts et 37 paragraphes de texte qui sont lus par tous les locuteurs. La partie privée inclut environ 2 000 paragraphes de texte. Elle est divisée en 50 tranches de 40 paragraphes dont chaque tranche est lue indépendamment par chaque locuteur.

5.2. Collection d'énoncés pour l'enregistrement

Deux phases de collecte de données de texte ont été réalisées dans le cadre du projet de collaboration international CORUS. Dans la première phase, les données sont collectées par quelques experts afin d'assurer les conditions désirées [Tran 2003]. Dans la deuxième phase, les données sont récupérées et filtrées automatiquement à partir de corpus de textes issus du Web.

Les données textuelles recueillies sur le Web ne sont cependant pas sous une forme présentable à un locuteur pour être enregistrées. Il faut donc les traiter et les filtrer en appliquant la boîte à outils *CLIPS-Text-Tk*. Par exemple, les nombres (dates, numéros de téléphone, numéros de carte bancaire) ont été transcrits sous forme textuelle (*exemple :* "12/3/1998" a été transcrit en : *"ngày mười hai tháng ba năm một nghìn chín trăm chín mươi tám"*, le numéro de portable "0904266805" a été transcrit en : *"không chín không bốn hai sáu sáu*

tám không năm″) [Le 2003a].

Nous constatons que les données textuelles choisies couvrent différents domaines de la vie quotidienne et contiennent beaucoup de dialogues et paragraphes courts (environ 130 mots à 170 mots équivalents à environ 20-25 secondes de parole par paragraphe ou conversation). Des détails supplémentaires sur la génération d'énoncés se trouvent dans [Le 2003a].

5.3. Enregistrement du corpus *VNSpeechCorpus*

A terme, le corpus *VNSpeechCorpus* contiendra 50 locuteurs (25 hommes et 25 femmes), dans une tranche d'âge de 15 à 45 ans. Les locuteurs choisis sont issus de quatre grandes villes et provinces du Vietnam : Hanoï, Nghe An, Hà Tinh et Ho Chi Minh ville, qui représentent les 3 régions dialectales principales.

Le logiciel d'enregistrement et de gestion du corpus vocal que nous avons utilisé est le logiciel EMACOP-Unicode[1], spécialement conçu dans notre laboratoire [Vaufreydaz 1998], et que nous avons adapté aux caractères Unicode (voir chapitre 3).

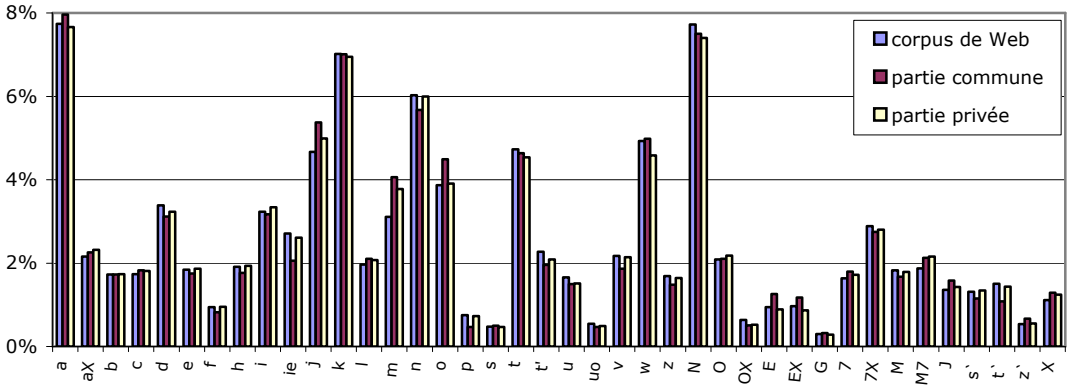
En 2005, 39 locuteurs dont 19 femmes et 20 hommes ont été enregistrés dans le studio du Centre MICA, Hanoi, Vietnam. Chaque locuteur a prononcé environ 60 minutes de parole, qui inclut 45-50 minutes communes (phonèmes, tons, chiffres, mots de commande et corpus de phrases et paragraphes communs) et 12-14 minutes privées (40 paragraphes courts).

5.4. Évaluation du corpus vocal

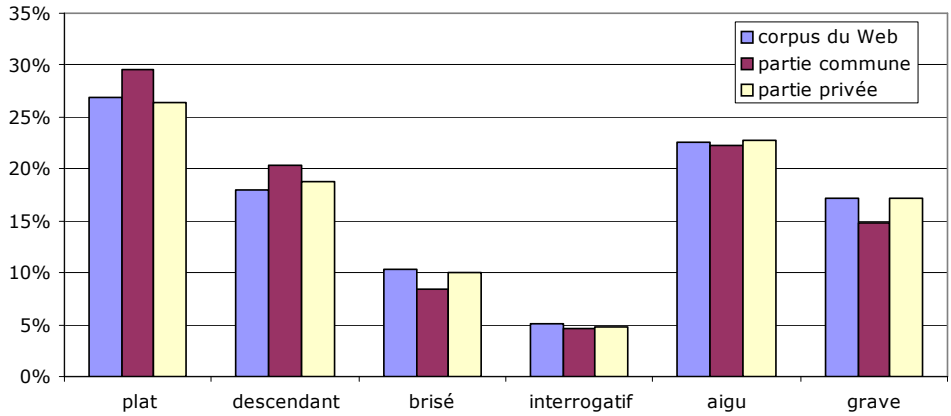
Le corpus sera entre autres utilisé, comme nous l'avons déjà précisé, pour l'entraînement d'un modèle acoustique pour faire de la reconnaissance automatique de la parole. Nous devons donc vérifier que la distribution des phonèmes de notre corpus n'est pas trop éloignée de celle de la langue vietnamienne. A notre connaissance, il n'existait pas encore de recherche sur la distribution phonétique de la langue vietnamienne, ainsi nous avons dû analyser nous-mêmes la distribution phonétique du vietnamien. Pour ce faire, nous utilisons un grand corpus de texte extrait à partir du Web qui couvre un grand nombre de domaines. Ce corpus est dont considéré comme une représentation de la langue vietnamienne.

Pour évaluer la distribution phonétique d'un corpus de texte, nous avons phonétisé toutes les phrases dans le corpus à l'aide du phonétiseur *VNPhoneAnalyzer* que nous avons construit. Les transcriptions de monophones (phonèmes), diphones, triphones, tons, parties initiales, parties finales des syllabes sont accumulées pour obtenir les fréquences d'apparition de chaque unité dans le corpus de texte. La figure 4.11 illustre une comparaison de la distribution phonétique des monophones (phonèmes) et la distribution des tons entre le grand corpus tiré du Web et le corpus vocal *VNSpeechCorpus* (partie *privée* et *commune*).

[1] Environnement Multimédia pour l'Acquisition et la gestion de Corpus Parole - Version Unicode multilingue



a) Distribution des phonèmes



b) Distribution des tons

Figure 4.11: Distribution phonétique dans le corpus VNSpeechCorpus par rapport à la distribution dans le grand corpus de texte récupéré sur le Web

Par ailleurs, nous avons obtenu des coefficients de corrélation entre les fréquences des différentes unités acoustiques du corpus *VNSpeechCorpus* et celles des unités acoustiques correspondante dans le grand corpus de texte issu du Web (tableau 4.11). Les coefficients de corrélation entre unités acoustiques de notre corpus vocal et du corpus de texte sont très proches de 1. Par conséquent, notre corpus peut être considéré comme équilibré phonétiquement.

Unité acoustique	Coefficient de corrélation	
	Partie privée	Partie commune
Monophone	0,99	0,99
Diphone	0,98	0,95
Triphone	0,98	0,94
Ton	0,99	0,98
Partie initiale de la syllabe	0,99	0,96
Partie finale de la syllabe	0,99	0,97

Tableau 4.11 : Coefficients de corrélation entre unités acoustiques du corpus vocal et du corpus de texte du Web

5.5. Répartition du corpus vocal obtenu

Au moment où nous avons réalisé nos expérimentations sur la portabilité et l'adaptation des modèles acoustiques multilingues pour le vietnamien, il y avait seulement 10 locuteurs ayant accompli l'enregistrement sur la première tranche du corpus *VNSpeechCorpus* dans le studio du Centre MICA [Le 2004]. Cela explique que dans les sections suivantes, nous utiliserons seulement 7 locuteurs pour l'apprentissage et l'adaptation des modèles acoustiques et 3 locuteurs pour l'évaluation des performances du système de reconnaissance du vietnamien.

Parmi les 7 locuteurs d'apprentissage et d'adaptation, nous n'utilisons que leurs énoncés correspondant aux paragraphes de texte, dont 50% des énoncés sont communs à tous les locuteurs et 50% des énoncés sont privés. Pour évaluer l'influence de la quantité de données d'adaptation sur la qualité des modèles acoustiques, nous divisons ce corpus d'apprentissage et d'adaptation en 3 sous-corpus :

- corpus-0,5h contient 30 minutes de données vocales enregistrées par les 2 premiers locuteurs ;
- corpus-1h contient 60 minutes de données vocales enregistrées par les 4 premiers locuteurs ;
- corpus-2,25h contient 135 minutes de données vocales enregistrées par les 7 locuteurs du corpus d'apprentissage.

Pour les 3 locuteurs de test, nous utilisons seulement leurs énoncés contenant des phrases de type « dialogue » soit 136 phrases par locuteur. Le tableau 4.12 présente la répartition du corpus d'apprentissage et du corpus de test.

Sous-corpus		Nombre de locuteurs	Durée totale du signal	Nombre de phrases (nombre de mots)
Corpus d'apprentissage	corpus-0,5h	2	30 min	330 (7933)
	corpus-1h	4	60 min	672 (16080)
	corpus-2,25h	7	135 min	1440 (34700)
Corpus de test	Phrases de dialogue	3	20 min	408 (4347)
	Nombres connectés	3	13,6 min	194 (2138)

Tableau 4.12 : Répartition du corpus d'apprentissage et corpus de test

Très récemment, nous avons aussi conduit une évaluation des performances de notre système sur un grand corpus de signaux en vietnamien ; nous utilisons le corpus de 39 locuteurs qui ont été enregistrés jusqu'alors. Sans compter les 3 locuteurs d'évaluation qui sont uniques pour toutes les expérimentations dans ce chapitre, nous avons alors dans ce dernier cas 36 locuteurs d'apprentissage et d'adaptation, soit 14 heures de signaux environ.

6. Modélisation acoustique à base de modèles multilingues

Dans cette section, nous présentons notre travail de recherche et des expérimentations sur la portabilité et l'adaptation des systèmes de reconnaissance automatique de la parole multilingue vers une langue peu dotée : le vietnamien.

6.1. Paramètres du système de reconnaissance automatique de la parole

Notre système de reconnaissance automatique de la parole est basé sur la boîte à outils Janus développée au laboratoire ISL-CMU [Finke 1997]. Toutes les unités acoustiques sont construites sur une topologie de HMMs dites de Bakis, c'est-à-dire un HMM gauche-droit d'ordre 1 à 3 états (sauf le silence, qui compte un seul état) où chaque état est une distribution multigaussienne. Chaque vecteur acoustique consiste en 13 premiers coefficients MFCCs, l'énergie, les dérivées première et seconde de ces coefficients et le taux de passage par zéro (ZCR) pour obtenir pour chaque fenêtre d'analyse de 20 ms un ensemble de 43 paramètres. Une transformation LDA (*Linear Discriminant Analysis*) est appliquée pour réduire l'espace de représentation à 24 coefficients.

Comme cela est dit précédemment, la syllabe joue un rôle très important dans la langue vietnamienne, notamment au niveau phonétique car les syllabes vietnamiennes sont souvent prononcées séparément dans une phrase. D'autre part, avec seulement 6 686 syllabes, nous pouvons couvrir presque tous les mots possibles de la langue vietnamienne. Ainsi, dans nos expérimentations, nous utilisons les *syllabes* plutôt que les *mots* comme unités de reconnaissance. Le système de reconnaissance du vietnamien est donc un système à base de syllabes (*Syllable-Based ASR system*). Par ailleurs, le vietnamien est une langue tonale avec 6 tons dont le ton est une caractéristique discriminante. Cependant, dans nos expérimentations, les phonèmes vietnamiens (monophones et polyphones) sont modélisés indépendamment du ton. Par conséquent, la différenciation de deux syllabes avec la même série de phonèmes est décidée par les modèles de langage. Des méthodes de modélisation et de reconnaissance du ton existent pour les langues tonales, telles que le mandarin [Chang 2000], le cantonais [Lee 2002], le vietnamien [Nguyen 2002, Nguyen-Duc 2006].

Par ailleurs, pour obtenir des modèles acoustiques pour le vietnamien, nous avons appliqué notre méthodologie à partir de deux modèles acoustiques en langue source différents :

- des modèles acoustiques appris sur le français. Ces modèles sont développés à partir du projet BRAF100 développé au laboratoire CLIPS-IMAG [Vaufreydaz 2000]. La couverture phonémique français / vietnamien est d'environ 63% ;

- des modèles acoustiques multilingues indépendants du contexte obtenus à partir de 7 langues (*MM7-CI*) du projet GlobalPhone développé au laboratoire ISL[1] [Schultz 2002] : chinois, croate, français, allemand, japonais, espagnol et turc. Des modèles acoustiques

[1] Interactive Systems Laboratories – Carnegie Mellon University (Etats Unis) & University of Karlsruhe (Allemagne)

multilingues dépendants du contexte ont été obtenus à partir de 6 langues (*MM6-CD*) du projet GlobalPhone : arabe, chinois, anglais, allemand, japonais et espagnol. La couverture phonémique multilingue (MM7) / vietnamien est d'environ 87%[1].

6.2. Systèmes expérimentaux

Pour estimer la performance de nos méthodes de portabilité et d'adaptation des modèles acoustiques, nous avons construit des systèmes expérimentaux différents :

- **système « baseline » :** nous utilisons la méthode traditionnelle de type « bootstrapping » pour développer des modèles acoustiques dépendants de la langue. Les modèles sont appris en utilisant la transformation LDA, l'initialisation des paramètres par l'algorithme kmeans, et l'apprentissage par l'algorithme Viterbi. Nous appelons ces systèmes : *VN-CI* (pour la modélisation indépendante du contexte) et *VN-CD2000* (pour la modélisation dépendante du contexte utilisant 2000 distributions) ;

- **systèmes « crosslingues »** : ce sont des systèmes dont les modèles acoustiques sont bootstrappés à partir de modèles issus d'une autre langue source (monolingue et multilingue). Dans notre expérimentation, nous utilisons deux systèmes en langue source différente : le système français (FR) et le système multilingue (MM). Ainsi les systèmes crosslingues s'ont appelés : *MM7/VN-CI* ou *FR/VN-CI* (pour la modélisation indépendante du contexte) et *MM6/VN-CD2000* ou *FR/VN-CD2000* (pour la modélisation dépendante du contexte en utilisant 2000 distributions) ;

- par ailleurs, nos méthodes de portabilité des modèles acoustiques reposent sur des méthodes d'estimation de distances et d'appariement entre phonèmes ou entre d'autres types d'unités acoustiques : méthode automatique (*data-driven*) ou méthode manuelle à base de connaissance phonétique (*knowledge-based*). Par exemple : *FR/VN-CI-Data* indique la modélisation « crosslingue » indépendante du contexte reposant sur une méthode automatique d'appariement, *MM6/VN-CD1000-Knowledge* s'indique la modélisation « crosslingue » dépendante du contexte reposant sur une méthode manuelle d'appariement.

6.3. Initialisation des modèles acoustiques crosslingues

Dans le chapitre 3 – section 5, nous avons présenté quelques méthodes de portabilité des modèles acoustiques vers une nouvelle langue peu dotée. Nous présentons dans cette section l'application de ces méthodes sur la langue vietnamienne. La langue source est les français ou un groupe de langues (multilingue).

[1] Ce travail a été réalisé lors de mon séjour à ISL-CMU et d'une collaboration avec Dr. Tanja Schultz.

6.3.1. Portabilité des modèles acoustiques indépendants du contexte

La portabilité des modèles acoustiques indépendants du contexte vers une nouvelle langue cible consiste à déterminer les couples de phonèmes les plus proches ou bien à construire un tableau de correspondances phonémiques (*phone mapping table*) source/cible. En effet, en appliquant la méthode de construction de tableaux de correspondances phonémiques présenté dans le chapitre 3, un tableau de correspondances phonémiques entre la langue cible (vietnamien) et la langue source (français, ou modèle multilingue issu de sept langues) est construit dans notre expérimentation (tableau 4.13).

Phonème vietnamien	Phonème français		Phonème issu d'un ensemble multilingue (*GlobalPhone*)	
	Obtenu manuellement	*Obtenu automatiquement*	*Obtenu manuellement*	*Obtenu automatiquement*
t	t	t	t	t
ɣ	g	g	ɣ	g
χ	k	k	χ	χ
ŋ	ŋ	ŋ	ŋ	ŋ
ʂ	s	s	ʂ	ʂ
w	w	w	w	au
e	e	e	e	e
uo	uœ	o	uɔ	u
ie	jø	i	iɛ	i
...	...	...	...	...

Tableau 4.13 : Exemple du tableau de correspondances phonémiques avec pour langue source le français et multilingue et pour langue cible le vietnamien

Après avoir obtenu le tableau de correspondances phonémiques source / cible, les modèles acoustiques indépendants du contexte (monophones) en langue source peuvent être dupliqués pour obtenir des modèles acoustiques en langue cible. L'avantage d'une telle approche est qu'elle ne nécessite pas ou peu de signaux d'apprentissage en langue cible puisque les modèles acoustiques du système de reconnaissance en langue cible sont, en fait, ceux d'une autre langue. Des tels modèles acoustiques sont ainsi appelés les *modèles acoustiques crosslingues*.

Ensuite, nous avons testé deux techniques d'obtention du tableau de correspondances phonémiques (*knowledge-based* et *data driven*) à partir des modèles acoustiques en langue source multilingues (MM7) et français (FR). Les taux d'exactitude en syllabes (*Syllable Accuracy* – SA) du système de reconnaissance automatique de la parole continue du vietnamien testé sur un corpus des phrases du type « dialogue » et sur un corpus de nombres connectés sont présentées dans la figure 4.12.

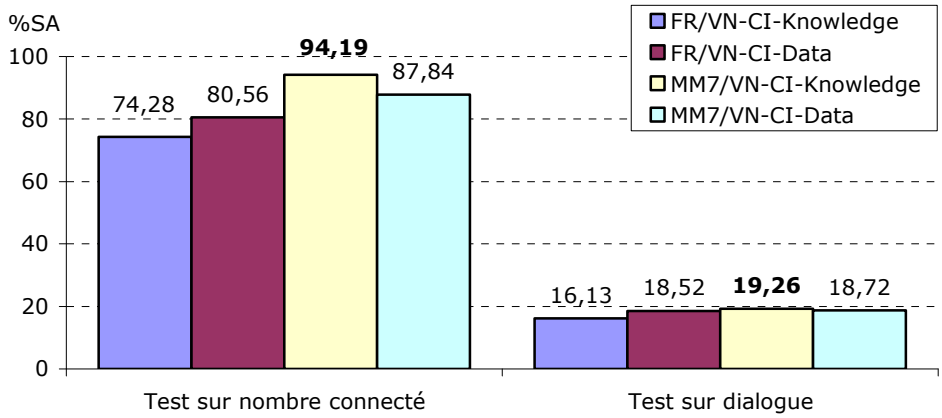


Figure 4.12 : Comparaison des performances des méthodes de portabilité des modèles acoustiques indépendantes du contexte
(langue source : français et multilingue ; langue cible : vietnamien)

Ces résultats montrent le potentiel de l'approche automatique pour la génération du tableau de correspondances phonémiques qui donne des performances équivalentes à celle obtenues avec la méthode manuelle. La méthode *data-driven* est meilleure que la méthode *knowledge-based* dans l'expérimentation monolingue (FR) mais elle est plus faible dans l'expérimentation multilingue (MM7).

Cependant, si les performances sur la reconnaissance de nombres connectés sont correctes, les performances restent inacceptables pour la reconnaissance de phrases. Précisons toutefois qu'à ce stade de l'expérimentation, aucun signal vietnamien n'a été utilisé pour apprendre les modèles acoustiques qui sont empruntés à d'autres langues.

Nous pouvons ajouter à cela quelques remarques :

- la qualité du tableau de correspondances phonétiques obtenu par la méthode automatique (*data-driven*) dépend de la performance du décodeur acoustico-phonétique et de la qualité des fichiers d'étiquettes phonétiques ;

- la couverture des phonèmes vietnamiens par l'ensemble des phonèmes français n'est pas optimale (63%) : il y a des phonèmes vietnamiens (par exemple, les diphtongues) qui n'existent pas dans la langue française. En conséquence, il est plus difficile de trouver un phonème en langue source qui est proche d'un phonème en langue cible dans le tableau API (méthode *knowledge-based*). Ainsi, la méthode *data-driven* est meilleure que la méthode *knowledge-based* dans ce cas. Par contre, pour les modèles multilingues qui présentent une meilleure couverture (87%), nous pouvons trouver plus facilement les phonèmes similaires dans le tableau API et la méthode knowledge-based est meilleure dans ce cas.

Nous constatons que le système multilingue MM7/VN-CI-Knowledge obtient le meilleur résultat dans cette expérimentation. Par conséquent, nous utiliserons ce système dans les expérimentations d'adaptation de la section 6.4.

De plus, nous pouvons utiliser les modèles acoustiques crosslingues MM7/VN-CI-

Knowledge comme les modèles acoustiques initiaux pour aligner temporellement automatiquement des données vocales en langue vietnamienne par l'algorithme *Viterbi*. Pratiquement, si aucun modèle acoustique n'existe au départ, nous pouvons utiliser les stratégies d'initialisation de modèles acoustiques : démarrage aléatoire (*random start*), démarrage uniforme (*flat start*), etc. Puis, les données vocales sont alignées temporellement à l'aide de ces modèles acoustiques initiaux. Les modèles acoustiques sont ensuite ré-entraînés à partir du corpus de signaux étiquetés et l'on réitère le cycle jusqu'à l'état stable du système est atteint. Cependant, les résultats présentés dans [Wheatley 1994, Schultz 1997] montrent que les modèles acoustiques initiaux crosslingues sont appris plus rapidement et mieux que les modèles acoustiques générés par les stratégies de démarrage aléatoire ou démarrage uniforme. Les étiquettes temporelles des données d'adaptation en vietnamien créées par les modèles acoustiques crosslingues seront donc utilisées dans les expérimentations d'adaptation suivantes.

6.3.2. Portabilité des modèles acoustiques dépendants du contexte

Dans la section précédente, nous avons montré le potentiel de l'emploi de modèles acoustiques multilingues à la place de modèles acoustiques monolingues pour avoir une meilleure couverture phonémique. Dans cette section, nous utilisons des modèles acoustiques multilingues dépendants du contexte (MM6-CD) pour construire des modèles acoustiques en langue vietnamienne.

Nous utilisons tout d'abord 2,25 heures de données vocales en vietnamien pour développer un arbre de décision (PT_T) pour 500, 1000 et 2000 distributions de sous-triphones. Nous notons que le terme de « sous-polyphone » (*sub-polyphone* en anglais) signifie un polyphone (triphone ou quintphone) qui est divisé en trois états : début, milieu et fin [Schultz 2001]. Par ailleurs, un arbre de décision multilingue (MM6) est déjà disponible avec 12000 distributions de sous-quinphones entraînés sur une base de données de signaux multilingues.

En appliquant la méthode de portabilité des modèles acoustiques dépendant du contexte présentée dans le chapitre 3 - section 5.4.2, les modèles acoustiques en vietnamien sont obtenus à partir des modèles acoustiques multilingues. Les résultats expérimentaux obtenus avec des modèles dépendants du contexte seront présentés dans la section suivante.

6.4. Adaptation des modèles acoustiques crosslingues

MLLR *(Maximum Likelihood Linear Regression*) et MAP *(Maximum A Posteriori)* sont deux méthodes d'adaptation bien connues dans la communauté de reconnaissance automatique de la parole. MLLR [Leggetter 1995] est une technique d'adaptation à base d'une transformation linéaire des paramètres des modèles acoustiques. L'adaptation MAP [Gauvain 1994] permet une réestimation des paramètres du HMM observés dans les données d'adaptation. Pour une petite quantité de données d'adaptation, la méthode MLLR est censée être meilleure. Mais pour une grande quantité de données d'adaptation, MAP a montré sa supériorité sur MLLR [Wang 2003]. Par conséquent, nous avons utilisé dans notre travail la technique d'adaptation MAP pour adapter des modèles acoustiques crosslingues obtenus dans la section précédente en utilisant des signaux d'adaptation en langue cible.

6.4.1. Adaptation des modèles indépendants du contexte (CI)

La performance de l'adaptation des modèles acoustiques crosslingues dépend de la quantité de données d'adaptation. Par ailleurs, les résultats dans [Wheatley 1994] montrent que le nombre de locuteurs utilisés est plus important que la quantité de signaux d'adaptation. Par conséquent, pour adapter les modèles acoustiques crosslingues, nous utilisons la méthode d'adaptation MAP sur 0,5 heure (2 locuteurs), 1 heure (4 locuteurs) et 2,25 heures (7 locuteurs) de données vocales en vietnamien, respectivement. La figure 4.13 présente les taux d'exactitude en syllabes des systèmes selon la quantité de signaux d'adaptation. Nous rappelons que VN-CI correspond à des modèles appris sur des données en vietnamien seules, tandis que MM7/VN-CI correspond à des modèles multilingues adaptés avec la même quantité de parole vietnamienne.

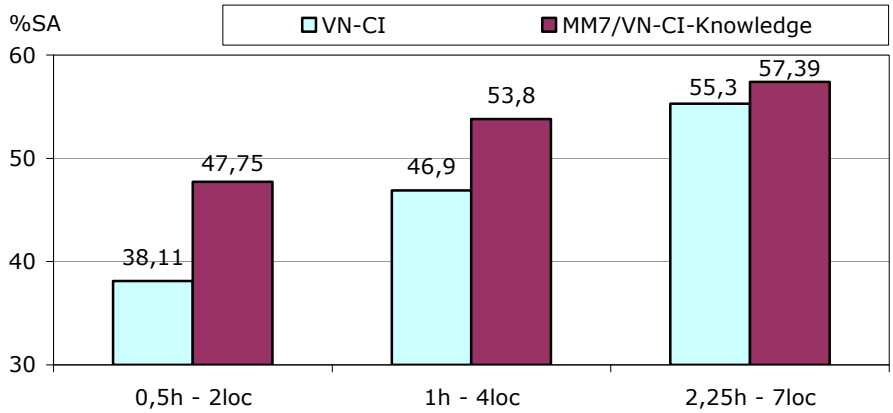


Figure 4.13 : Comparaison de taux d'exactitude en syllabes (SA) selon la quantité de signaux d'adaptation en vietnamien entre le système baseline et le système crosslingue

Nous constatons que le système crosslingue MM7/VN-CI-Knowledge obtient les meilleurs résultats selon la quantité de signaux d'adaptation. Quand la quantité de signaux d'adaptation est augmentée de 0,5 heures à 2,25 heures, le gain de taux d'exactitude entre les deux systèmes VN-CI et MM7/VN-CI s'est réduit (de 24,6% à 3,8%). Cela montre le potentiel de l'approche de modélisation acoustique crosslingue indépendante du contexte quand nous possédons seulement une quantité très réduite de signaux d'adaptation.

6.4.2. Adaptation des modèles dépendants du contexte (CD)

La figure 4.14 illustre le taux d'exactitude en syllabes (*Syllable Accuracy - SA*) du système baseline (VN-CD) appris sur le vietnamien et du système crosslingue (MM6-CD-Knowledge) adapté par la méthode MAP. Nous constatons que le système crosslingue utilisant 500, 1000 et 2000 distributions de « sous-triphones » améliore de 1,7%, 6,9% et 28,1% respectivement, le taux d'exactitude par rapport au système « baseline ». Quand le nombre de modèles augmente de 500 à 2000, le taux d'exactitude du système baseline VN-CD diminue proportionnellement, probablement en raison de la quantité insuffisante de données d'adaptation. Cependant, le système crosslingue reste peu sensible à ce problème.

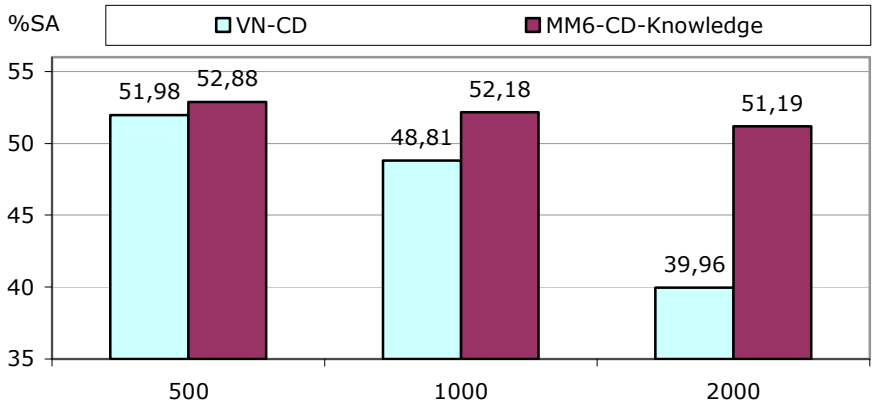


Figure 4.14 : Comparaison de taux d'exactitude en syllabes (SA) entre le système baseline et le système crosslingue selon le nombre de modèles de sous-triphone sur la modélisation dépendante du contexte (2,25 heures de signaux d'adaptation)

La figure 4.15 présente l'influence de la taille des données d'adaptation et le nombre de locuteurs d'adaptation en langue cible sur deux méthodes d'appariement d'unités source / cible : la méthode automatique (*data-driven*) et la méthode à base de connaissances phonémiques (*knowledge-based*). Nous constatons que la méthode automatique est légèrement meilleure que la méthode à base de connaissances phonémiques pour les modèles MM6/VN-CD500 seulement tandis que la méthode à base de connaissances est meilleure pour les modèles MM6/VN-CD1000 et MM6/VN-CD2000.

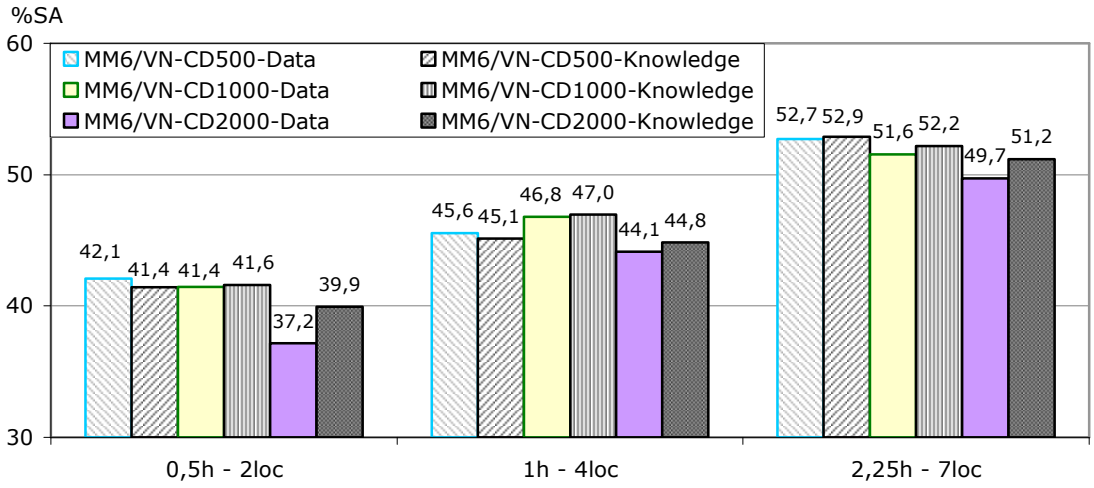


Figure 4.15 : Comparaison de taux d'exactitude en syllabes (SA) des méthodes d'appariement d'unités source / cible selon la quantité de signaux d'adaptation en vietnamien

6.4.3. Sélection d'une méthode de modélisation acoustique selon la quantité de données d'adaptation

La section précédente montre les avantages de la modélisation acoustique crosslingue lorsque peu de données en langue cible (2,25h) sont disponibles. Les deux méthodes de modélisation indépendante du contexte et dépendante du contexte sont meilleures que les

méthodes *baselines* correspondantes.

La figure 4.16 présente la performance des méthodes de modélisation acoustique selon la quantité de données d'adaptation. Le système MM7/VN-CI est meilleur sur 2,25h (7 locuteurs) de données mais le système MM6/VN-CD1000 est meilleur sur 14h (36 locuteurs) de données. Nous constatons que quand nous avons une quantité réduite de signaux de parole (par exemple 2-3 heures), la méthode de modélisation crosslingue indépendante du contexte est le premier choix. Cependant, quand la taille du corpus signaux devient suffisante (10-15 heures), la méthode de modélisation crosslingue dépendante du contexte est meilleure.

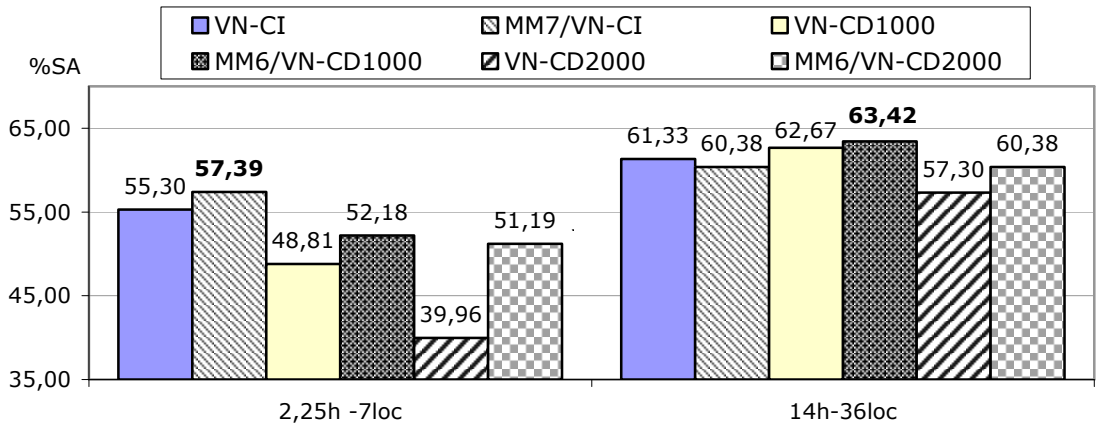


Figure 4.16 : Comparaison des méthodes de modélisation acoustique selon la quantité de données d'adaptation

7. Modélisation acoustique à base de graphèmes

Bien que nous ayons obtenu automatiquement un dictionnaire phonétique pour la construction d'un système de reconnaissance automatique de la parole en vietnamien, nous nous sommes intéressés à la modélisation acoustique à base de graphèmes en vietnamien pour les raisons suivantes :

- pour estimer l'efficacité de la modélisation à base de graphèmes pour une langue peu dotée, il est intéressant de la comparer à celle de modèles de référence à base de phonèmes. Dans la section précédente, une telle modélisation acoustique a été réalisée sur le vietnamien ;

- l'écriture vietnamienne est une écriture latinisée qui est une transcription du vietnamien parlé en signes alphabétiques empruntés aux langues d'origine latine. Par conséquent, la distance entre phonème et graphème est assez faible ;

- nous utilisons la langue vietnamienne pour évaluer la performance de la méthode d'initialisation de modèles acoustiques graphémiques proposée dans le chapitre 3.

En fait, la construction automatique d'un dictionnaire de prononciation à base de graphèmes

est très simple pour le vietnamien : la représentation d'une entrée lexicale (un mot) dans le vocabulaire est une suite de graphèmes dans le dictionnaire de prononciation. Le tableau 4.14 illustre un exemple du dictionnaire de prononciation vietnamien à base de graphèmes que nous avons généré dans notre travail.

Mot	Prononciation à base de graphèmes	Mot	Prononciation à base de graphèmes
a dua	a d u a	chuyên gia	c h u y ê n g i a
am tường	a m t ư ơ n g	dần	d â n
anh	a n h	gả	g a
ban hành	b a n h a n h	giã	g i a
ban trưa	b a n t r ư a	hồng quân	h ô n g q u â n
bút pháp	b u t p h a p	mạnh thường quân	m a n h t h ư ơ n g q u â n
bơi thuyền	b ơ i t h u y ê n	quạng	q u a n g
bằng	b ă n g	rập khuôn	r â p k h u ô n

Tableau 4.14 : Exemple du dictionnaire de prononciation vietnamien à base de graphèmes

Pour la modélisation indépendante du contexte à base de graphèmes, nous initialisons les modèles de départ par deux approches différentes d'initialisation de modèles acoustiques graphémiques :

- méthode « baseline » : initialisation uniforme (*flat start*) [Killer 2003a] ;
- méthode d'initialisation à base de détection de frontières des mots.

Après avoir initialisé des modèles acoustiques graphémiques, nous utilisons ces modèles pour aligner temporellement les signaux d'apprentissage. Ensuite, les modèles acoustiques sont ré-entraînés à partir de ce corpus de signaux étiquetés (apprentissage à base d'étiquettes) et l'on réitère le cycle. La procédure d'apprentissage est arrêtée quand le système atteint un état stable, normalement après 6-8 itérations de la procédure d'apprentissage.

Enfin, pour construire des modèles acoustiques dépendants du contexte à base de graphèmes, nous utilisons la méthode de « *singleton* » pour modéliser la dépendance au contexte en utilisant des « *polygraphèmes* ». L'idée de génération des questions linguistiques pour construire l'arbre de décision de polyphone est simplement de demander quel est le graphème dans le contexte gauche ou droit. Chaque question linguistique consiste en un seul graphème (d'où le nom de méthode « singleton »). Pour plus de détail, voir [Killer 2003a].

La figure 4.17 présente une comparaison de performances (taux d'exactitude en syllabes) des systèmes de reconnaissance automatique de la parole du vietnamien en utilisant deux méthodes de modélisation acoustique : à base de phonèmes et à base de graphèmes.

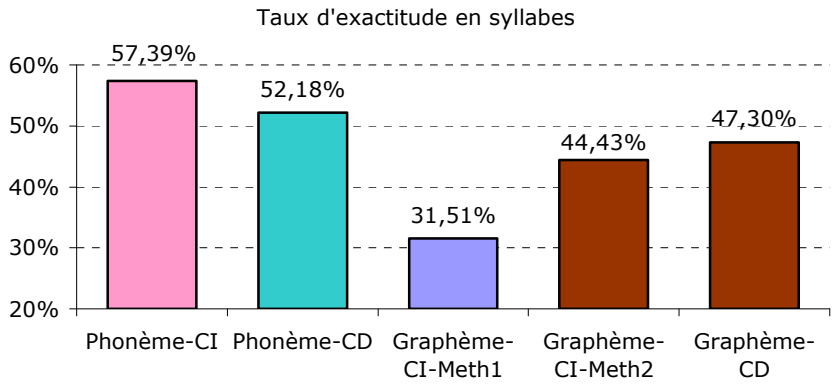


Figure 4.17 : Comparaison de performance des méthodes de modélisation acoustique sur 2,25 heures des signaux d'apprentissage
(CD : modélisation dépendante du contexte, CI : modélisation indépendante du contexte)

À partir des résultats obtenus, nous constatons que la méthode « *baseline* » d'initialisation de modèles graphémiques indépendants du contexte n'obtient qu'un taux d'exactitude en syllabes de 31,51% tandis que notre méthode d'initialisation des modèles graphémiques à partir des modèles « *mot/silence* » est plus efficace avec 44,43% de taux d'exactitude. Ainsi, une amélioration significative de 41% de taux d'exactitude absolus est obtenue. De plus, comme cela est dit dans le chapitre 3, la modélisation indépendante du contexte à base de graphèmes n'est pas très efficace, la modélisation dépendante du contexte obtient en revanche des résultats presque comparables aux approches à base de phonèmes (47,30% contre 52,18%).

8. Conclusions du chapitre

Dans ce chapitre, nous avons présenté nos travaux sur la construction d'un système de reconnaissance automatique de la parole pour la langue vietnamienne. Dans un premier temps, des ressources linguistiques importantes ont été recueillies en appliquant la méthodologie et les outils présentés dans les chapitres II et III. La quantité de pages Web collectées était de 2,5Go. Après filtrage, la quantité de données textuelles pouvant servir à l'apprentissage d'un modèle de langage statistique était d'environ 400 Mo (5 millions de phrases). A titre de comparaison, une année complète du journal « Le Monde » en français correspond à 120Mo en moyenne.

Ensuite, nous avons recueilli un vocabulaire de 6 686 syllabes et un vocabulaire de 40 000 mots à partir des ressources lexicales existantes. Un analyseur phonétique à base de règles a été développé pour obtenir automatiquement un dictionnaire de prononciation vietnamien.

Pour la modélisation statistique du langage, 4 méthodes de filtrage du corpus de texte sont présentées. Un modèle de langage a pu être construit très rapidement en redéveloppant seulement 15% des outils initiaux de notre boîte à outils *CLIPS-Text-Tk*. En plus, en filtrant les informations redondantes dans les documents récupérées à partir du Web, la taille de notre corpus de texte est réduite de 54% dans nos expériences. La valeur de perplexité des modèles de langage obtenus est nettement améliorée de 26%. On constate que notre filtrage de l'information redondante est efficace, voire indispensable lorsqu'on utilise des données issues du Web.

Pour la modélisation acoustique, nous avons utilisé nos méthodes de portabilité et d'adaptation rapide des modèles acoustiques multilingues vers une langue cible peu dotée. Les résultats obtenus montrent le potentiel des méthodes de portabilité que nous avons proposées, notamment dans le contexte de langues peu dotées ne possédant pas ou peu de ressources acoustiques.

Enfin, nous avons terminé ce chapitre sur la langue vietnamienne par une modélisation acoustique à base de graphèmes qui montre l'intérêt d'une initialisation de modèles graphémiques à partir de modèles « *mot/silence* ».

Chapitre 5

Application au khmer

Dans ce chapitre, nous présenterons nos travaux sur la construction rapide d'un système de reconnaissance automatique de la parole pour le khmer. Ils constituent une validation de notre méthodologie sur une nouvelle langue, et sur une période courte de développement. Tout d'abord, pour réaliser un système de reconnaissance automatique de la parole, une connaissance minimale des caractéristiques linguistiques et acoustiques de la langue est nécessaire. Ensuite, nous abordons nos travaux concernant la collecte de ressources textuelles (vocabulaires, corpus textuels) et acoustiques (corpus de parole, dictionnaire de prononciation) pour le khmer. La modélisation du langage et la modélisation acoustique à base de graphèmes pour le khmer seront présentées. Enfin, nous terminerons ce chapitre par des résultats d'expérimentations et les comparaisons des performances du système obtenu.

1. Linguistique, phonologie et phonétique du khmer

Cette section présente les grandes lignes concernant la linguistique, la phonologie et la phonétique du khmer. Ce travail s'est fait dans le cadre du projet TALK [Talk 2005] et par une collaboration avec un groupe linguistique de la langue khmer dirigé par M. Jean-Michel PHILIPI, un expert linguistique français.

1.1. Généralités

La figure 5.1 présente la distribution des langues au Cambodge. La langue khmère est la langue principale et la langue officielle du Cambodge. Elle est parlée par environ 12,1 millions de personnes au Cambodge et environ 13,3 millions de personnes dans le monde[1].

La langue khmère appartient à la famille linguistique môn-khmère[2] qui regroupe 147 langues en Asie du Sud-est, et est une des deux familles du groupe linguistique austro-asiatique. De plus, la langue khmère a été largement influencée par le sanskrit[3], le pali[4] et le français.

Contrairement au chinois, au thaï et au vietnamien, le khmer est une langue atonale avec un haut pourcentage de mots bisyllabiques. Par contre, elle possède un des plus riches systèmes

[1] http://www.ethnologue.com/show_language.asp?code=khm
[2] http://www.ethnologue.com/show_family.asp?subid=90153
3 Le Sanskrit est une langue de la famille Indo-Européenne, et elle est une langue officielle de l'Inde. http://www.nationmaster.com/encyclopedia/Sanskrit
[4] Pali : Langage de Buddha : http://www.nationmaster.com/encyclopedia/Pali

vocaliques au monde.

Figure 5.1 : Distribution linguistique au Cambodge[1]

1.2. Système d'écriture du khmer

Différemment de l'écriture chinoise (sinogramme) pour laquelle chaque lettre représente un morphème, l'écriture script khmère appelée lettres khmères (អក្ខរក្រមខេមរភាសា /ʔaʔsar kmae/ [Huffman 1970]) provient d'un alphabet indien du sud de l'Inde (comme le thaï, le laotien, le birman). Le système d'écriture du khmer est alphasyllabique (ou alphabet syllabique[2]) et consiste en symboles (les caractères khmers) pour les consonnes et les voyelles. Chaque consonne appartient à l'une des deux séries de consonnes. La voyelle produite dépend de la combinaison faite avec la consonne initiale. Ainsi la plupart des voyelles ont deux prononciations possibles. De plus, la langue khmère a aussi des diacritiques qui changent une série de consonnes ou changent la prononciation de la voyelle. Par conséquent, au niveau de l'écriture, pour adapter les fontes informatiques, il a fallu gérer un ordonnancement, sur plusieurs niveaux d'écriture.

La langue khmère a bénéficié d'un grand nombre d'emprunts au sanskrit, au pali et au français ainsi que, dans les milieux urbanisés, au chinois. La plus grande partie du vocabulaire administratif, militaire et littéraire est empruntée au sanskrit. Avec l'introduction du bouddhisme au début de 15ème siècle, le pali devient une source d'emprunts lexicaux très importante. Plus récemment, du fait de l'occupation française en Indochine, certains mots ont été empruntés du français, mais orthographiés en khmer [Talk 2005].

[1] http://www.ethnologue.com/show_map.asp?name=KH
[2] http://www.omniglot.com/writing/syllabic.htm

Actuellement, l'écriture du khmer se présente sans séparation entre les mots ; on place fréquemment un espace entre des groupes de mots (comme une virgule ou un point-virgule en français), mais cette insertion repose sur des critères flous qui sont soumis à de grandes variations suivant les auteurs. Il existe aussi un signe de ponctuation qui correspond quasiment au point du français, mais ce que ce signe délimite ne correspond pas non plus à des critères nettement établis : il peut être placé, dans certains cas, à la fin d'une phrase mais, dans d'autres cas, à la fin d'un paragraphe groupant plusieurs phrases. Ceci pose des problèmes en traitement automatique qui devront être résolus**.** A titre d'exemple, la figure 5.2 illustre des phrases khmères et la segmentation « manuelle » de ces phrases en phrases et en mots. Nous notons que la notation « _ » est utilisée comme un délimitateur des mots dans les phrases segmentées.

Phrases khmères originale :

នៅស្រុកខ្មែ វាងាយស្រួលសំរាប់នារីជាងបុរសក្នុងការរកការងារធ្វើ តែការងារខ្លះ មានតែបុរសទៃដលអាចធ្វើបាន ។

Phrases khmères segmentées :

Phrase 1 : នៅ_ស្រុក_ខ្មែ វា_ងាយ_ស្រួល_សំរាប់_នារី_ជាង_បុរស_ក្នុង_ការ_រក_ការងារ_ធ្វើ

Phrase 2 : តែ_ការងារ_ខ្លះ មានតែ_បុរស_ទេ_ទៃដល_អាច_ធ្វើ_បាន ។

Traduction en français :

Au Cambodge, il est plus facile pour un femme de trouver un travail que pour un homme. Mais certains travaux ne peuvent etre effectues que par des hommes.

Figure 5.2 : Exemple de segmentation en phrases et en mots

1.3. Alphabet et phonologie du khmer

L'alphabet khmer possède 33 consonnes, 32 consonnes souscrites, 21 voyelles dépendantes et 14 voyelles indépendantes, sans compter les consonnes empruntées au thaï et au français, les chiffres, les ligatures, les diacritiques et la ponctuation.

1.3.1. Consonnes

Le tableau 5.1 présente l'alphabet de 33 consonnes et 32 consonnes souscrites khmères avec les phonèmes correspondants dans le tableau de l'Alphabet Phonétique International (API). Nous notons que la consonne souscrite est une consonne qui est écrite au-dessous d'une consonne initiale de la syllabe.

Les consonnes khmères sont classées en deux séries : la série A (3ème colonne) correspond aux consonnes prononcées avec la voyelle /aː/ et la série O (4ème colonne) correspond aux consonnes prononcées avec la voyelle /ɔː/ [Huffman 1970].

Ecriture khmère		API		Ecriture khmère		API	
Consonne	Consonne souscrite	Série A	Série O	Consonne	Consonne souscrite	Série A	Série O
ក	◌្ក	k		ទ	◌្ទ		t
ខ	◌្ខ	k^h		ធ	◌្ធ		t^h
គ	◌្គ		k	ន	◌្ន		n
ឃ	◌្ឃ		k^h	ប	◌្ប	b	
ង	◌្ង		ŋ	ផ	◌្ផ	p^h	
ច	◌្ច	c		ព	◌្ព		p
ឆ	◌្ឆ	c^h		ភ	◌្ភ		p^h
ជ	◌្ជ		c	ម	◌្ម		m
ឈ	◌្ឈ		c^h	យ	◌្យ		j
ញ	◌្ញ		ɲ	រ	◌្រ		r
ដ	◌្ដ	d		ល	◌្ល		l
ឋ	◌្ឋ	t^h		វ	◌្វ		v
ឌ	◌្ឌ		d	ស	◌្ស	s	
ឍ	◌្ឍ		t^h	ហ	◌្ហ	h	
ណ	◌្ណ	n		ឡ		l	
ត	◌្ត	t		អ	◌្អ	ʔ	
ថ	◌្ថ	t^h					

Tableau 5.1 : Consonnes khmères

1.3.2. Voyelles dépendantes et indépendantes

Les voyelles de l'écriture khmères sont classées en deux groupes : les voyelles dépendantes et les voyelles indépendantes.

a) Voyelles dépendantes

Dans une syllabe écrite, les voyelles khmères peuvent apparaître avant, après, au-dessus ou au-dessous d'une consonne initiale. Elles sont divisées en deux séries différentes. Le tableau 5.2 présente la liste des voyelles khmères suivies par leurs transcriptions phonétiques en API.

Ecriture khmère	API		Ecriture khmère	API	
Voyelle	Série A	Série O	Voyelle	Série A	Série O
◌	aː	ɔː	◌ៀ	iɜ	
◌ា	aː	iɜ	◌ោ	ei	eː
◌ិ	e	i	◌ែ	aɛ	ɛː
◌ី	ɜj	iː	◌ៃ	aj	ɨj
◌ឹ	ɜ	ɨ	◌ាោ	aɔ	oː
◌ឺ	ɜɨ	ɨː	◌ាៅ	aw	ɨw
◌ុ	o	u	◌ុំ	om	um
◌ូ	ou	uː	◌ំ	ɑm	um
◌ួ	uɜ		◌ាំ	am	oɜm
◌ើ	aɜ	ɐː	◌ះ	ah	eɜh
◌ឿ	ɨɜ				

Tableau 5.2 : Voyelles dépendantes khmères

Au niveau de la phonétique, il y a 29 phonèmes vocaliques : 10 phonèmes longs, 7 phonèmes brefs et 12 diphtongues. Dans le tableau 5.2, les phonèmes longs ont un signe / ː / à la fin. Le lecteur trouvera une analyse phonétique détaillée dans [Sam 2004] et [Talk 2005].

b) Voyelles indépendantes

Phonétiquement, une voyelle indépendante (ou voyelle complète) est une fusion d'une consonne initiale et d'une voyelle [Huffman 1970]. Par exemple, la voyelle indépendante ឣឿ est équivalente au អើ (= consonne អ + voyelle ◌ើ), la voyelle indépendante ឤ est équivalente au លឺ (= consonne ល + voyelle ◌ឺ).

Ecriture khmère	API	Ecriture khmère	API
ឥ	ʔɜj	ឦ	lɨ
ឣឿ	ʔɜj	ឤ	lɨː
ឧ	ʔu	ឯ	ʔae
ឩ	ʔou	ឰ	ʔaj
ឪ	ʔɜw	ឱ	ʔao
ឫ	rɨ	ឲ្យ	ʔaoj
ឬ	rɨː	ឳ	ʔaw

Tableau 5.3 : Voyelles indépendantes khmères

Le tableau 5.3 présente la liste des voyelles indépendantes khmères. Dans cette liste, il n'y a que 3 consonnes initiales qui peuvent être combinés avec les voyelles : /ʔ/, /r/ et /l/.

1.3.3. Chiffres khmers

Comme le thaï, les chiffres proviennent aussi de l'alphabet indien du sud de l'Inde. Actuellement, malgré le fait que les chiffres arabes, utilisés dans les écritures latines, sont très connus au Cambodge grâce à l'influence du français, les chiffres khmers sont principalement utilisés. Le tableau 5.4 présente les chiffres khmers avec leur prononciation.

Chiffre arabe	0	1	2	3	4	5	6	7	8	9
Chiffre khmer	០	១	២	៣	៤	៥	៦	៧	៨	៩
API	soun	muɜj	piː	bɜj	buɜn	pram	pram muɜj	pram piː	pram bɜj	pram buɜn

Tableau 5.4 : Chiffres khmers

Pour les nombres, on peut combiner simplement ces chiffres khmers dans l'écriture, mais on doit ajouter les prononciations pour la dizaine, la centaine, le millier, le million et aussi pour le milliard correspondants. Par exemple, le nombre 6752 s'écrit par ៦៧៥២ et se prononce / pram-muɜj poɜn pram-piː roː haː sep piː / en API. Pour plus de détails, voir [Huffman 1970].

1.3.4. Ponctuation, diacritiques et ligatures khmères

Comme nous l'avons dit, le système d'écriture khmère est un système non-segmenté. Ainsi il n'y pas de signes distinctifs entre deux mots et entre deux phrases. Cependant, il existe encore des signes de ponctuations qui sont illustrés dans le tableau 5.5.

Nom	Ecriture	Situation d'utilisation
espace		comme : « , », « ; », « . » en français
khan	។	terminaison d'une phrase ou fin d'un paragraphe
baariyaosaan	៕	terminaison d'un chapitre
deux points	៖	comme « : » en français
laq	។ល។	comme « et cetera » en français
…	...	…

Tableau 5.5 : Exemple de ponctuation khmère

D'autre part, l'écriture khmère utilise aussi quelques ponctuations latines comme celles du français : le point « . »[1] , la virgule « , », le tiret, le point d'interrogation « ? », le point d'exclamation « ! », les points de suspension « … », les parenthèses « { }‘’() », les

[1] Ce n'est par un point à la fin d'une phrase comme celui du français, mais à la fin d'un caractère *khan* ។

guillemets « "" », le point-virgule « ; ». L'existence de ces ponctuations latines dans l'écriture khmère est un résultat de l'influence du français.

Il existe, par ailleurs, des ligatures khmères qui combinent des consonnes et des voyelles. De plus, on trouve encore des diacritiques qui ont pour rôle, par exemple, de modificateurs de la prononciation des caractères, tels que :

◌ះ, ◌ៈ, ◌៉, ◌៊, ◌់, ◌៌, ◌៎, ◌៏, ◌័, ◌៑

Le lecteur trouvera une explication détaillée de celle-ci dans [Huffman 1970].

1.4. Structure syllabique et structure du mot

La langue khmère se compose de monosyllabes et polysyllabes. Les polysyllabes sont courantes surtout dans leur forme bi- et tri-syllabique et consistent surtout en des termes savants comme illustré dans le tableau 5.6.

Type de syllabe	Mots Khmers	En français	API
1-syllabe	កា	quart	kaː
2-syllabes	សាលា	école	salaː
3-syllabes	កសិកម្ម	agriculture	kaʔseʔkam
4-syllabes	សុខាភិបាល	médecin	sɔʔkʰapʰibal
5-syllabes	ភោជនីយដ្ឋាន	restaurant	pʰocɜnijɜtʰaːn

Tableau 5.6 : Exemples de syllabes khmères

La langue khmère possède un grand nombre du polysyllabes, mais tend au monosyllabisme dans sa variante parlée. Certaines structures syllabiques initiales sont systématiquement réduites à une consonne suivie d'un appendice vocalique de type voyelle centrale [ɜ]. Ainsi le mot [bɑnlaɛ] (légume) sera prononcé dans la conversation courante : [pɜlaɛ], le tout ne formant qu'une syllabe.

En effet, la structure syllabique générale du khmer se retrouve sous la forme suivante avec **C** consonne et **V** voyelle [Huffman 1970]:

$$\mathbf{C_1\,(C_2)\,(C_3)\,V\,(C_4)}$$

Nous notons que la consonne initiale et la voyelle du noyau sont obligatoires et les autres consonnes sont facultatives. Cependant, pour les voyelles brèves (courtes), la consonne finale est obligatoire. En effet, la consonne C_3 est rarement utilisée.

2. Recueil des ressources linguistiques

2.1. Vocabulaire

À partir du projet KhmerOS[1] (*Khmer Software Initiative*), un dictionnaire traditionnel « Chuon Nat » a été récupéré. Ensuite, pour générer un vocabulaire de mots en khmers, nous avons filtré ce dictionnaire traditionnel et un vocabulaire de 16 000 mots a été ainsi obtenu dans notre travail.

Nous avons effectué une analyse de la couverture lexicale de mots dans un corpus de texte en khmer collecté à partir du Web. Aucun filtrage de phrases n'est appliqué sur ce corpus de texte. Alors, un total de 17800 formes lexicales est trouvé sur le corpus de texte. Nous trouvons que les 8000 mots les plus fréquents du corpus de texte (~50% des mots) couvrent déjà 99,7% des mots du corpus de texte collecté sur le Web (figure 5.3).

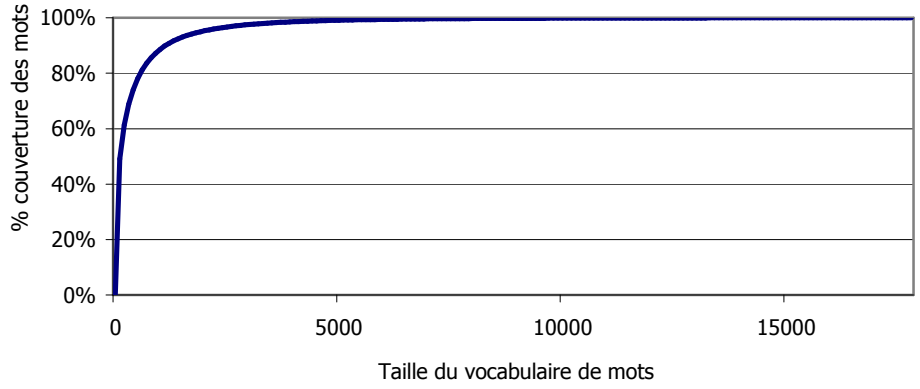


Figure 5.3 : Couverture lexicale sur le corpus tiré du Web selon la taille du vocabulaire de mots du khmer

2.2. Recueil d'un corpus de texte général à partir de l'Internet

Le recueil d'un corpus de documents html en langue khmère à partir de l'Internet est une tâche très difficile car le nombre de sites Web écrits en khmer et la bande passante d'Internet dans ce pays sont très faibles. En effet, à l'aide d'étudiants cambodgiens, nous avons trouvé quelques sites web publiées par le gouvernement cambodgien, par des organisations ou des entreprises. Nous avons d'abord remarqué que beaucoup de sites web hébergés au Cambodge sont écrits en anglais ou en français. Il y a cependant quelques sites écrits en langue khmère. Avec ceux-ci, il y a encore des difficultés de récupération automatique : sites écrits en flash[2], pages encodées par un système d'encodage spécifique ou privé[3] que nous n'avons pas réussi à convertir en un encodage standard (Unicode). Nous avons trouvé cependant un site d'informations en khmer[4]. La quantité de pages Web collectées à partir de ce site est de 80Mo (environ 6000 pages), ce qui reste faible par rapport aux 2,5Go de pages web collectées pour le

[1] http://www.khmeros.info/
[2] http://www.everyday.com.kh
[3] http://www.seasite.niu.edu/khmer/
[4] Community Information Web Portal Cambodia : www.cambodiacic.org

vietnamien et par rapport aux 40Go pour le français (corpus WebFR4) [Vaufreydaz 2002].

En utilisant la méthode de récupération des documents à base de moteurs de recherche, nous obtenons seulement 113 documents pertinents en khmer à partir des 17 sites Web dans lesquelles nous choisissons 3 sites les plus riches en ressources. Puis, nous utilisons l'outil *wget*[1] pour collecter tous les documents de ces 3 sites. Un total de 13 131 pages html a obtenu très rapidement (pendant environ 2 heures seulement), soit 174 Mo de données bruitées. Cette méthode est intéressante car nous pouvons contrôler la performance et la vitesse de récupération. Par contre, nous constatons que les documents proviennent d'une même source de site Web ont l'avantage d'être homogènes en format et en qualité. Ainsi le traitement de ces documents devient plus facile dans les étapes suivantes.

En appliquant la boîte à outils de traitement de corpus de texte *CLIPS-Text-Tk* mentionnée dans le chapitre 2, nous pouvons obtenir rapidement un corpus de texte traité en khmer. Cependant, une procédure d'adaptation de la boîte à outils *CLIPS-Text-Tk* vers le khmer est nécessaire : Nous héritons de tous les modules indépendants de la langue et adaptons rapidement les modules dépendants de la langue khmère tels que : la conversion des encodages vers Unicode, la séparation en phrases, la transcription des chiffres et des nombres khmers. Cette procédure d'adaptation est réalisée en 2 semaines seulement. Il est évident que cela économisera le temps de développement.

Par ailleurs, la procédure de segmentation d'une phrase en mots ou en syllabes est très difficile pour la langue khmère qui possède une écriture non-segmentée : les mots sont écrits sans espace et il n'existe pas de ponctuation claire entre les phrases (figure 5.4). Par conséquent, le traitement d'un corpus de texte en khmer nécessite de résoudre ces problèmes.

Phrase originale : ការហើមសាច់នៅកន្លែងរលាក

Phrase segmentée : ការ_ហើម_សាច់_នៅ_កន្លែង_រលាក

Figure 5.4 : Exemple d'une phrase khmère non-segmentée et segmentée en mots

Comme cela est abordé dans le chapitre 2, pour segmenter une phrase khmère en syllabes, un segmenteur syllabique est construit en employant un algorithme de programmation dynamique, à l'aide d'un modèle syllabique, qui segmente une phase de texte en optimisant le critère de « plus petit nombre de syllabes » (*Maximal Matching*). Nous avons évalué la performance du segmenteur syllabique sur un corpus de 47 phrases khmères non-segmentées, soit 621 syllabes. Pour la référence, le corpus de texte est segmenté manuellement en syllabes par un étudiant cambodgien. Le taux de syllabes segmentées correctement est de *81,5%*.

Par ailleurs, pour segmenter en mots, nous essayons plusieurs méthodes de segmentation et nous trouvons que la méthode à base d'un modèle de langage unigramme est la meilleure. En effet, cette méthode améliore de *0,71%* le taux de mots segmentés corrects et de *3,96%* le taux

[1] http://www.gnu.org/software/wget/wget.html

de phrases segmentées correctes par rapport aux méthodes traditionnelles à base de vocabulaire.

Après filtrage et traitement, le corpus de texte en khmer pouvant servir à l'apprentissage d'un modèle de langage statistique était d'environ 97 Mo, soit 1,1 millions de phrases ou 8 millions de mots segmentés.

2.3. Modélisation statistique du langage

2.3.1. Filtrage du corpus de texte d'apprentissage

Concernant les modèles de langage du khmer, parce que la performance du segmenteur de mots est meilleure que celle du segmenteur syllabique, nous avons choisi un système de reconnaissance de mots où les cooccurrences modélisées sont des suites de mots.

Ensuite, nous utilisons le corpus de texte segmenté obtenu dans la section précédente pour construire et entraîner les modèles statistiques du langage du khmer. De plus, nous essayons des méthodes différentes de filtrage des phrases dans un corpus de texte comme celles présentées dans le chapitre 3 en utilisant le vocabulaire Chuon Nat de 16 000 mots :

- **filtrage 1 :** prendre toutes les phrases (ne pas appliquer de filtrage) ;
- **filtrage 2 :** prendre les phrases ayant au moins *N* mots et dont tous les mots appartiennent au vocabulaire ;
- **filtrage 3 :** prendre une séquence consécutive (un bloc) d'au moins *M* mots appartenant au vocabulaire ;
- **filtrage 4 :** utiliser une méthode hybride qui consiste à prendre les phrases entières ayant au moins *N* mots appartenant au vocabulaire (filtrage 2) et appliquer le filtrage par blocs minimaux de taille *M* (filtrage 3) sur les phrases rejetées.

Les tailles du corpus de texte obtenu par ces méthodes de filtrage avec tailles du vocabulaire correspondant (le nombre de mots différents) sont présentées dans le tableau 5.7.

Filtrages	Filtrage 1	Filtrage 2 (*N*=1)	Filtrage 3		Filtrage 4 (*N*=1)	
			M=3	*M*=5	*M*=3	*M*=5
Nombre de mots	8 003 256	3 464 336	6 919 185	5 417 977	7 170 983	6 209 583
Nombre de mots différents	7 795	7 109	7 428	7 107	7 665	7 553

Tableau 5.7 : Taille du corpus de texte selon la méthode de filtrage

Nous constatons que le nombre de mots différents, qui représente le vocabulaire réellement rencontré dans les corpus de texte filtrés, est faible également (de 7100 à 7800) par rapport la taille du vocabulaire Chuon Nat (16000 mots).

2.3.2. Apprentissage des modèles de langage

Pour apprendre les modèles de langage, nous utilisons la boîte à outils SRILM [Stolcke 2002] en utilisant la méthode de Good-Turing pour le lissage avec le repli de Katz [Katz 1987]. Le tableau 5.8 présente la taille des modèles de langage (2-grammes et 3-grammes) selon les méthodes de filtrages utilisées.

Filtrages		Filtrage 1	Filtrage 2 (*N*=1)	Filtrage 3		Filtrage 4 (*N*=1)	
				M=3	*M*=5	*M*=3	*M*=5
ML	2-grammes	409 313	270 805	398 052	362 590	402 756	381 311
	3-grammes	484 965	277 998	471 190	417 682	479 941	451 180

Tableau 5.8 : Nombre de n-grammes des modèles de langage

Nous constatons qu'à cause de la taille limitée du corpus de texte d'apprentissage (plus de 8 millions de mots seulement), le nombre de 3-grammes est faible par rapport au nombre de 2-grammes, notamment pour la 2ème méthode de filtrage (prendre seulement des phrases entières dont tous les mots sont connus du vocabulaire).

2.3.3. Evaluation de la perplexité

Comme nous l'avons dit dans le chapitre d'état de l'art, la perplexité est une mesure très répandue pour l'évaluation des modèles de langage. Dans notre travail, nous avons évalué la perplexité de nos modèles de langage sur un corpus de test de 200 phrases, soit 2499 mots avec un taux de mots hors-vocabulaire de 0,2%.

Le tableau 5.9 présente les valeurs de perplexité des modèles de langage obtenus. Nous notons que les phrases de test et les phrases du corpus d'apprentissage, bien que différentes, sont extraites à partir de la même source (un site Web). Cela explique que la valeur de perplexité évaluée sur les modèles de trigrammes soit très basse malgré la petite taille du corpus d'apprentissage.

Filtrages		Filtrage 1	Filtrage 2 (*N*=1)	Filtrage 3 (*N*=1)		Filtrage 4 (*N*=1)	
				M=3	*M*=5	*M*=3	*M*=5
ML	2-grammes	127,3	**116,6**	126,2	132,5	125,1	123,1
	3-grammes	88,0	**83,92**	87,7	94,1	86,9	86,1

Tableau 5.9 : Valeurs de perplexité calculées sur le corpus de test

À partir des valeurs de perplexités, nous trouvons que la méthode de filtrage 2 (prendre seulement des phrases entières dont tous les mots sont connus du vocabulaire) obtient la meilleure valeur de perplexité bien que les faibles différences ne soient pas très représentatives. Nous utiliserons toutefois les modèles de langage construits à partir de corpus issus de la méthode de filtrage 2 pour évaluer notre système de reconnaissance final en khmer.

3. Acquisition d'un corpus de parole en khmer

Pour développer et évaluer notre système de reconnaissance vocale, nous avons d'abord enregistré un corpus vocal de plusieurs locuteurs selon le prototype d'acquisition d'un corpus de parole déjà présenté dans le chapitre 3. Dans cette section, nous présentons la constitution, l'enregistrement et la répartition du corpus ITC-10, un corpus vocal enregistré par 10 locuteurs khmers.

3.1. Obtention d'énoncés pour l'enregistrement

Pour enregistrer un corpus vocal, nous devons tout d'abord définir un ensemble de phrases à faire prononcer par les locuteurs. Cependant, le processus de sélection des phrases prononcées en grande quantité n'est pas aisé car il doit être réalisé automatiquement. Le corpus de phrases est extrait à partir du corpus de texte général en khmer récupéré dans la section précédente. Ensuite, nous avons utilisé le vocabulaire Chuon Nat de 16 000 mots pour récupérer toutes les phrases contenant plus de 5 mots du corpus d'apprentissage récupéré du Web. Le vocabulaire a été obtenu à partir du dictionnaire khmer Chuon Nat[1].

Enfin, le corpus textuel, entièrement vérifié à l'ITC par S. Sam, un enseignant-chercheur de l'ITC, est constitué pour l'instant d'un total de 2600 phrases qui sont prononcées par 10 locuteurs phnompenhnois avec environ la moitié de phrases communes à tous les locuteurs (130 phrases environ) et une autre moitié différente pour chaque locuteur (129 phrases par locuteur en moyenne).

3.2. Enregistrement du corpus vocal ITC-10

Les enregistrements ont été réalisés dans un studio isolé du département de Génie Informatique et Communication, de l'ITC, à l'aide d'un microphone-casque. Ce microphone est d'excellente qualité avec une réponse en fréquence plate et optimisée pour la parole. De plus, l'utilisation d'un microphone-casque permet de régler la position du microphone par rapport à la bouche, d'une manière quasi identique pour tous les locuteurs (pour plus de détail, voir [Talk 2005]).

Le logiciel d'enregistrement et de gestion du corpus vocal que nous avons utilisé est le logiciel EMACOP-Unicode[2], spécialement conçu dans notre laboratoire [Vaufreydaz 1998], qui est adapté aux caractères Unicode (voir chapitre 3).

[1] http://www.khmeros.info/
[2] Environnement Multimédia pour l'Acquisition et la gestion de Corpus Parole - Version Unicode multilingue

L'enregistrement de 10 locuteurs du corpus ITC-10 a été réalisée, soit un total de 3 heures de données vocales environ (voir tableau 5.10).

Composition du corpus ITC-10	Nombre de locuteurs	Durée totale du signal	Nombre de phrases (nombre de mots)
Moitié des phrases communes	10	86,8 min	1306 (16133)
Moitié des phrases différentes	10	92.3 min	1293 (16479)
Total	10	179 min	2599 (32612)

Tableau 5.10 : Statistique du corpus ITC-10

3.3. Répartition du corpus vocal obtenu

Nous divisons le corpus vocal entier en 2 sous-corpus : corpus d'apprentissage et corpus de test. Le corpus de test comprend 200 phrases prononcées par tous les 10 locuteurs, soit 20 phrases par locuteur, extraites à partir de la moitié des phrases différentes du corpus ITC-10. Le corpus d'apprentissage contient le reste des phrases différentes et toute la moitié de phrases communes du corpus ITC-10. Le tableau 5.11 présente la répartition du corpus en données d'apprentissage et de test.

Sous-corpus	Nombre de locuteurs	Durée totale du signal	Nombre de phrases (nombre de mots)	Nombre de mots différents du corpus
Corpus de test	10	14 min	200 (2499)	690
Corpus d'apprentissage	10	166 min	2399 (30113)	1875
Total	10	179 min	2599 (32612)	1971

Tableau 5.11 : Répartition du corpus d'apprentissage et corpus de test

Nous notons que les 10 locuteurs du corpus d'apprentissage sont apparus dans le corpus de test. Par conséquent, les évaluations du système de reconnaissance vocale en khmer appris à partir de corpus ITC-10 sont relativement dépendantes du locuteur.

4. Modélisation acoustique

Comme cela est dit dans le chapitre 3, un dictionnaire de prononciation est une ressource essentielle aux tâches de synthèse et de reconnaissance automatique de la parole. Cependant, au moment où nous construisions le système de reconnaissance automatique de la parole en khmer, aucun dictionnaire phonétique n'existait dans la communauté de traitement de la langue khmère. L'Institut de Technologie du Cambodge a commencé des recherches sur la langue khmère [Sam 2004, TALK 2005] mais l'information phonétique sur la langue khmère dont nous disposions à ce moment ne nous a pas permis de construire un dictionnaire phonétique.

Ainsi, pour le khmer, nous décidons d'utiliser l'approche de modélisation acoustique graphémique (modélisation acoustique à base de graphèmes ou *grapheme based modeling* en anglais). Cette section aborde plus en détail notre travail de construction d'un dictionnaire de prononciation du khmer et la modélisation acoustique à base de graphèmes.

4.1. Construction d'un dictionnaire de prononciation à base de graphèmes

La construction automatique d'un dictionnaire de prononciation à base de graphèmes est très simple pour les écritures latines (anglais, français, allemand, espagnol, ...) : la représentation d'une entrée lexicale (un mot) dans le vocabulaire est une suite de graphèmes dans le dictionnaire de prononciation. Cependant, pour l'écriture khmère et les autres écritures non-latines, la construction d'un dictionnaire à base de graphèmes est un peu plus délicate. Elle consiste d'abord à convertir les caractères en une forme plus lisible pour le développeur et dans un jeu de caractères interne plus simple pour l'ordinateur (latinisation ou romanisation[1]).

4.1.1. Romanisation

Le tableau 5.12 illustre un extrait du tableau de caractère Unicode et la liste des noms de caractère du Standard Unicode[2] (version 4.1) pour le khmer.

@@ 1780	**Khmer 17FF**	17B8	KHMER VOWEL SIGN II
@	**Consonants**	17B9	KHMER VOWEL SIGN Y
1780	KHMER LETTER KA	17BA	KHMER VOWEL SIGN YY
1781	KHMER LETTER KHA	...	
1782	KHMER LETTER KO	**@**	**Numeric symbols for divination lore**
1783	KHMER LETTER KHO	17F0	KHMER SYMBOL LEK ATTAK SON
1784	KHMER LETTER NGO	17F1	KHMER SYMBOL LEK ATTAK MUOY
...		17F2	KHMER SYMBOL LEK ATTAK PII
@	**Independent vowels**	17F3	KHMER SYMBOL LEK ATTAK BEI
17A4	KHMER INDEPENDENT VOWEL QAA	17F4	KHMER SYMBOL LEK ATTAK BUON
17A5	KHMER INDEPENDENT VOWEL QI	17F5	KHMER SYMBOL LEK ATTAK PRAM
17A6	KHMER INDEPENDENT VOWEL QII	17F6	KHMER SYMBOL LEK ATTAK PRAM-MUOY
17A7	KHMER INDEPENDENT VOWEL QU	17F7	KHMER SYMBOL LEK ATTAK PRAM-PII
17A8	KHMER INDEPENDENT VOWEL QUK	17F8	KHMER SYMBOL LEK ATTAK PRAM-BEI
...		17F9	KHMER SYMBOL LEK ATTAK PRAM-BUON
@	**Dependent vowel signs**	...	
17B6	KHMER VOWEL SIGN AA		
17B7	KHMER VOWEL SIGN I		

Tableau 5.12 : Tableau de l'Unicode pour le khmer

[1] La romanisation (ou latinisation) est la translittération ou la transcription d'une écriture non latine vers une écriture latine.
[2] http://www.unicode.org/charts/PDF/U1780.pdf

Pour romaniser les caractères khmers à partir de leurs noms représentés dans le tableau de caractères Unicode, nous réduisons le nom de chaque caractère comme cela est présenté dans le tableau 5.13. La liste complète de la romanisation de l'écriture khmère se trouve en annexe B. Ainsi, un total de 66 caractères khmers est romanisé dans notre expérimentation.

Caractère Unicode khmer	Caractère de romanisation	Nom anglais du caractère khmer
ក	Ka	KHMER LETTER KA
ខ	KHa	KHMER LETTER KHA
គ	Ko	KHMER LETTER KO
ឃ	KHo	KHMER LETTER KHO
…	…	…
ឥ	QI	KHMER INDEPENDENT VOWEL QI
ឦ	QII	KHMER INDEPENDENT VOWEL QII
…	…	…
ឱ	QOO1	KHMER INDEPENDENT VOWEL QOO TYPE 1
ឲ	QOO2	KHMER INDEPENDENT VOWEL QOO TYPE 2
◌ា	AA	KHMER VOWEL SIGN AA
◌ិ	I	KHMER VOWEL SIGN I
◌ៀ	YA	KHMER VOWEL SIGN YA
◌ំ	AE	KHMER VOWEL SIGN AE
…	…	…
◌ំ	NIKAHIT	KHMER SIGN NIKAHIT
◌ះ	REAHMUK	KHMER SIGN REAHMUK
◌ः	YUUKALEAPINTU	KHMER SIGN YUUKALEAPINTU
…	…	…

Tableau 5.13 : Romanisation des caractères khmers

4.1.2. Génération du dictionnaire

En s'appuyant sur le tableau de romanisation créé dans la section précédente, nous pouvons générer un dictionnaire de prononciation du khmer à base de graphèmes. En effet, pour chaque entrée lexicale (un mot) du vocabulaire, nous romanisons tous les caractères du mot consécutivement. La représentation de ce mot est alors une suite de graphèmes dans le dictionnaire de prononciation.

Le tableau 5.14 illustre un extrait du dictionnaire de prononciation à base de graphèmes que nous avons créé. Dans notre travail, une série de caractères latins qui romanise un caractère khmer est appelé un *graphème romanisé* ou simplement un *graphème*. Nous notons qu'un graphème dans les écritures latines est souvent représenté par un seul caractère latin ce qui n'est pas le cas ici.

Mot khmer	Prononciation à base de graphèmes
SIL	SIL
ក	Ka
កក	Ka Ka
កកិចកកុច	Ka Ka I Ca Ka Ka U Ca
កកិត	Ka Ka I Ta
កកិល	Ka Ka I Lo
គោងក្រពាត់	Ta OO NGo Ka Ro Po AA Ta
គោងទាម	Ta OO NGo To AA Mo
គោតត្វង	Ta OO Ta Ta UU Ngo
បិណ្ឌបាតចារិកវត្ត	Ba I NNo Do Ba AA Ta Ca AA Ro I Ka Vo Ta Ta
បិណ្ឌបាតទាន	Ba I NNo Do Ba AA Ta To AA No
...	...

Tableau 5.14 : Dictionnaire de prononciation en khmer à base de graphèmes

D'autre part, en khmer, chaque chiffre n'existe que sous forme d'un caractère. Il n'existe donc pas de transcription des chiffres khmers comme c'est le cas pour les écritures latines (par exemple, 1 est transcrit par UN, 9 est transcrit par NEUF, …). Par conséquent, chaque chiffre khmer est représenté par un mot dans le dictionnaire de prononciation (tableau 5.15). Les prononciations pour la dizaine, la centaine, le millier, le million et le milliard, … des nombres khmer sont aussi ajoutées (se référer à la section 1.3.3)

Chiffre khmer	Valeur	Prononciation à base de graphèmes
០	0	SON
១	1	MUOY
២	2	PII
៣	3	BEI
៤	4	BUON
៥	5	PRAM
៦	6	PRAM MUOY
៧	7	PRAM PII
៨	8	PRAM BEI
៩	9	PRAM BUON

Tableau 5.15 : Prononciation des chiffres khmers

5. Modélisation acoustique à base de graphèmes

5.1. Modélisation indépendante du contexte

Pour la modélisation acoustique indépendante du contexte à base de graphèmes, nous devons tout d'abord déterminer combien de modèles nous avons (nombre de graphèmes) pour le

khmer. Dans notre expérimentation, trois types de modèles sont utilisés pour la modélisation acoustique indépendante du contexte :

1. un modèle de silence qui est un HMM à 1 état ;
2. 86 modèles de graphèmes correspondant aux 86 graphèmes khmers représentés dans le dictionnaire de prononciation. Chaque modèle est un HMM de 3 états, qui modélisent le début, le milieu et la fin du graphème.

Après avoir déterminé les modèles acoustiques, nous initialisons les modèles au départ par l'approche d'initialisation de modèles acoustiques graphémiques que nous proposons dans la chapitre 3. Pour une phrase khmère en entrée, nous pouvons, en utilisant les modèles « mot/silence » construits précédemment pour le vietnamien, décoder la frontière des mots par l'algorithme de Viterbi. Ensuite, chaque morceau de signal correspondant à un mot est segmenté *uniformément* suivant le nombre de graphèmes qui composent la transcription de ce mot. Ces segments de données vocales ainsi obtenus sont alors utilisés pour apprendre les modèles graphémiques correspondant.

Après avoir appris des modèles acoustiques initiaux, nous utilisons ces modèles pour réaligner temporellement (étiqueter) les signaux d'apprentissage. Ensuite, les modèles acoustiques sont ré-entraînés à partir de ce corpus de signaux étiquetés (apprentissage à base d'étiquettes) et l'on réitère le cycle. La procédure d'apprentissage est arrêtée quand le système atteint un état stable (normalement après 6-8 itérations de la procédure d'apprentissage).

5.2. Modélisation dépendante du contexte

Comme cela est présenté dans la section sur la phonologie khmère, la langue khmère est une langue alphasyllabique. Ainsi la prononciation des phonèmes dans une syllabe et la prononciation de la syllabe dépendent de la combinaison des caractères (par exemple, la consonne initiale et la voyelle de noyau), des signes et des diacritiques. Par conséquent, la relation graphème-phonème n'est pas aussi triviale que pour d'autres langues comme le vietnamien ou l'allemand par exemple [Huffman 1970]. En résumé**,** la prononciation d'un caractère khmer est souvent dépendante du contexte. Cette caractéristique particulière du khmer est très importante pour nous car dans notre travail, nous utilisons le graphème au lieu du phonème. Ainsi, la modélisation acoustique dépendante du contexte à base de graphèmes devrait aussi être considérée.

L'une des difficultés de la modélisation acoustique dépendante du contexte est la génération des questions linguistiques afin de regrouper les unités similaires dans un modèle acoustique en utilisant un arbre de décision (*decision tree-based clustering*). Dans nos expérimentations, nous utilisons deux méthodes de génération d'arbre de décision :

- Méthode du « singleton » : chaque question linguistique consiste en un seul graphème ;
- Méthode à base de relation « graphème – phonème ».

Un exemple des questions linguistiques générées par la méthode du « singleton » est illustré dans la figure 5.5. Par exemple, les graphèmes *To* dont le contexte droit est le graphème OE peuvent être regroupés dans un nœud de l'arbre de décision ; le graphème QUUV (une voyelle indépendante) est regroupé dans un nœud indépendamment de son contexte, …

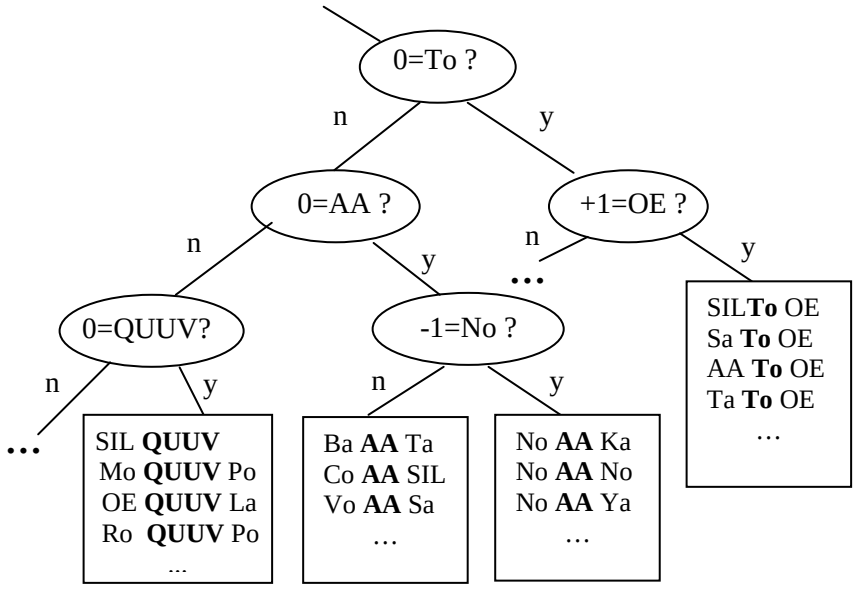


Figure 5.5 : Arbre de décision avec la méthode du singleton

Le tableau 5.16 présente les questions linguistiques que nous générons par la méthode à base de relation « graphème – phonème ». Nous notons que pour le khmer, la relation graphème-phonème n'est pas directe (un graphème correspond à un ou plusieurs phonèmes et inversement).

Nom de la question	Graphèmes regroupés
SILENCES	SIL
CONSONNES	Ka KHa Ko KHo NGo Ca CHa Co CHo NYo Da TTHa Do TTHo NNo Ta THa To THo No Ba PHa Po PHo Mo Yo Ro Lo Vo Sa Ha La Qa
VOYELLES_IND	QI QII QU QUU QUUV RY LY LYY QE QAI QOO1 QAU
VOWELLES_DEP	AA I II Y YY U UU UA OE YA IE E AE AI OO AU
DIACRITIQUES	NIKAHIT REAHMUK YUUKALEAPINTU
CONS_TYPE_A	Ka KHa Ca CHa Da TTHa Ta THa Ba PHa Sa Ha La Qa
CONS_TYPE_O	Ko KHo NGo Co CHo NYo Do TTHo NNo To THo No Po PHo Mo Yo Ro Lo Vo
VELARS	Ka KHa Ko KHo Ngo
PALATALS	Ca Co CHa CHo Nyo
RETROFLEXES	Da Do TTHa TTHo Nno
…	…

Tableau 5.16 : Ensemble des questions linguistiques à base de relation « graphème - phonème »

A titre d'exemple, les graphèmes *To* dont le contexte droit est une voyelle peuvent être regroupés dans un nœud de l'arbre de décision (figure 5.6).

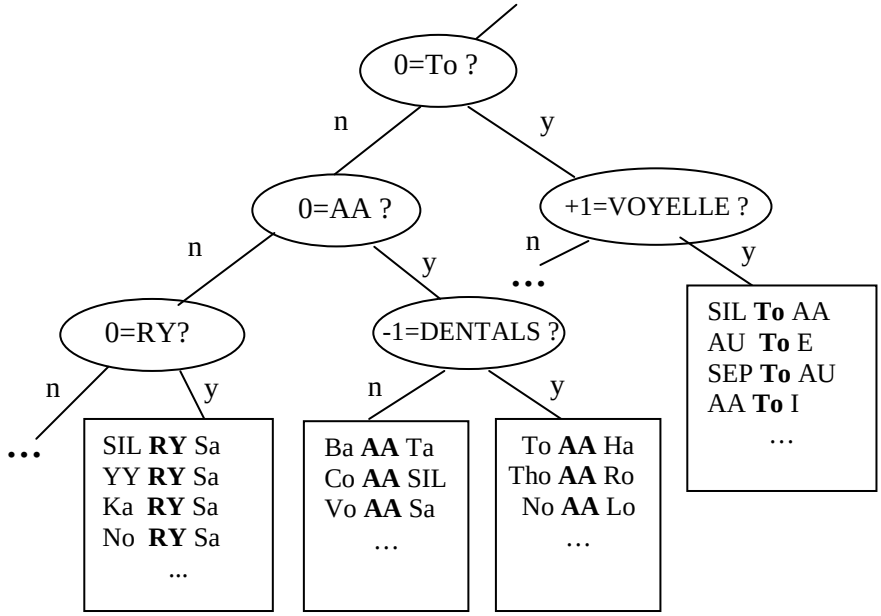


Figure 5.6 : Arbre de décision avec la méthode utilisant des connaissances phonétiques

6. Résultats d'expérimentation du système

Pour évaluer la performance du système de reconnaissance automatique de la parole continue à grand vocabulaire du khmer, nous utilisons le corpus de test du corpus ITC-10 que nous avons construit et présenté dans la section 3 ci-dessus. Le corpus de test comprend 200 phrases prononcées par les 10 locuteurs, soit 20 phrases par locuteurs. Nous notons que les 10 locuteurs du corpus d'apprentissage sont tous apparus dans le corpus de test. Par conséquent, les évaluations du système de reconnaissance vocale en khmer appris à partir du corpus ITC-10 sont en mode *dépendant du locuteur*.

Le tableau 5.17 présente le taux d'exactitude en mots (*Word Accuracy – WA*) du système selon les modèles acoustiques indépendants du contexte et l'historique (N) pris en compte dans les modèles de langage. Nous constatons qu'à cause de la taille limitée du corpus de texte d'apprentissage (plus de 8 millions de mots), le nombre de 3-grammes est faible par rapport au nombre 2-grammes (voir le tableau 5.8). Cependant, la performance du système utilisant les modèles 3-grammes est toujours la meilleure.

Modèles de langage	2-grammes	3-grammes
Taux d'exactitude	71,31%	**73,63 %**

Tableau 5.17 : Taux d'exactitude en mots du système selon le modèle de langage

Nous présentons dans la figure 5.7 l'évolution de la performance du système pendant 8 cycles de « bootstrapping » des modèles acoustiques indépendants du contexte. Après avoir construit les modèles acoustiques initiaux, le taux d'exactitude en mots est égal à 50,46%. Après 5 cycles de « bootstrapping » (1 cycle de bootstrapping = 1 étiquetage temporel + 1 apprentissage à base d'étiquettes), la performance du système est de 73,63% et le système atteint un état stable.

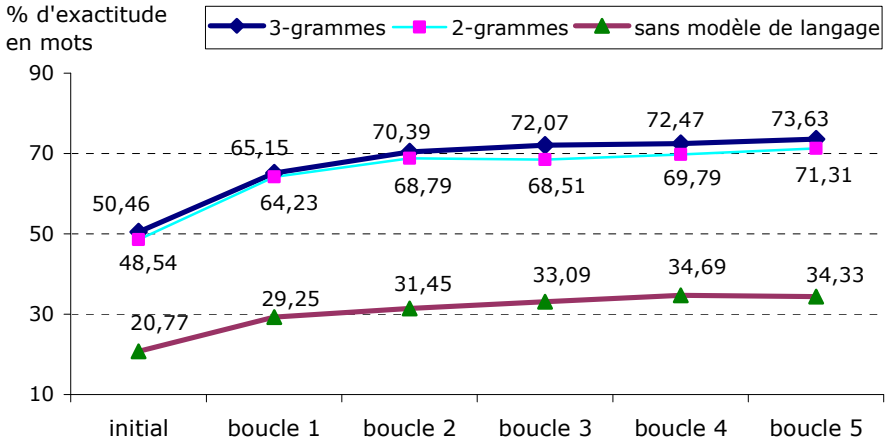


Figure 5.7 : Evolution du système pendant 5 cycles de « bootstrapping »

La figure 5.8 présente les taux d'exactitude en mots évalué selon deux méthodes de modélisation dépendante du contexte : singleton et graphème-phonème (utilisant des connaissances phonétiques). Pour chaque méthode, le nombre de modèles acoustiques (modèles de sous-graphème) choisis est 500, 1000 et 1500. Malgré la faible quantité de données d'apprentissage, nous trouvons que la modélisation dépendante du contexte améliore les performances. Les meilleurs modèles acoustiques sont ceux créés par la méthode dépendante du contexte « graphème-phonème » avec 1000 modèles.

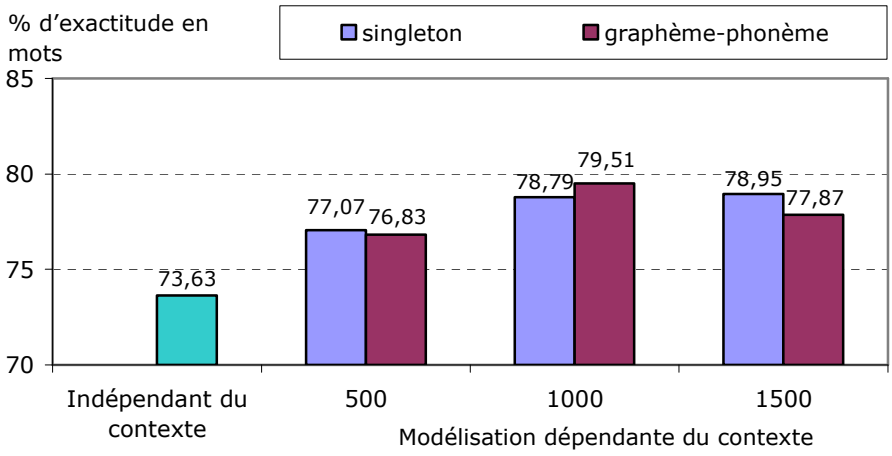


Figure 5.8 : Comparaison des méthodes de modélisation acoustique

7. Conclusions du chapitre

Nous avons présenté dans ce chapitre nos travaux sur la validation de notre méthodologie de construction rapide d'un système de reconnaissance automatique de la parole pour une langue peu dotée : le khmer.

D'abord, des ressources linguistiques relativement importantes ont été recueillies rapidement en appliquant les méthodes et les outils de récupération et de traitement présentés dans les chapitres II. Un vocabulaire de 16 000 mots khmers est collecté à partir du dictionnaire traditionnel Chuon Nat. Puis, en utilisant notre méthode de récupération des documents pour des langues peu dotées, un corpus de documents html de 174Mo en khmer a été collecté dans un premier temps. Après filtrage et traitement, le corpus de texte obtenu était d'environ 97 Mo, soit 1,1 millions de phrases ou 8 millions de mots segmentés. Pour la modélisation statistique du langage, 4 méthodes de filtrage du corpus de texte ont été comparées et le filtrage 2 (prendre seulement des phrases entières dont tous les mots sont connus du vocabulaire) est le meilleur dans notre expérimentation. Le corpus de texte issu par ce filtrage a été utilisé pour apprendre des modèles de langage.

Ensuite, pour la modélisation acoustique, un corpus vocal contenant 10 locuteurs a été enregistré, soit un total de 3 heures de données vocales environ. Par ailleurs, au moment où nous avons construit le système de reconnaissance automatique de la parole en khmer, aucun dictionnaire phonétique n'existait dans la communauté de traitement de la langue khmère. Ainsi, nous avons choisis le graphème comme unité de modélisation lexicale et acoustique. Un dictionnaire de prononciation à base de graphèmes a été généré par une procédure de romanisation qui convertit les caractères khmers en une forme lisible par l'ordinateur. Bien que nous ayons moins de 3 heures de signaux d'apprentissage, ce qui est insuffisant pour la modélisation acoustique dépendante du contexte, les résultats d'expérimentations ont montré que la modélisation dépendante du contexte est meilleure que la modélisation indépendante du contexte et une amélioration de 22,3% sur le taux d'erreur absolu est obtenu dans notre expérimentation.

Enfin, un premier système de reconnaissance a été achevé en 4 mois et un taux de reconnaissance de mots d'environ 80% (sur une tâche de parole lue cependant) en mode dépendant du locuteur a été obtenu ce qui montre l'efficacité de notre méthodologie et des outils développés.

Conclusions et perspectives

Conclusions

Premier médium de communication entre les hommes, la parole est, de ce fait, le signal d'information le plus communément transmis. Cependant, dans la plupart des langues peu dotées, les services liés au traitement de l'oral sont inexistants. Nous avons ainsi abordé la problématique du développement rapide de technologies vocales pour des langues peu dotées, en nous intéressant plus particulièrement à la reconnaissance automatique de la parole. Il est aussi très important de noter que les recherches sur les technologies vocales (synthèse et reconnaissance) peuvent permettre d'améliorer les ressources linguistiques nécessaires à d'autres services. Par exemple, les travaux sur la conversion graphème-phonème nécessitent des analyseurs qui peuvent avoir des applications pour des tâches plus « textuelles » comme la correction automatique. En outre, la reconnaissance et la synthèse vocale nécessitent l'emploi de dictionnaires de prononciation (ou phonétiques) qui sont des ressources très intéressantes dans l'optique générale de l'informatisation d'une langue (enrichissement et amélioration de dictionnaires existants par exemple).

Les méthodes présentées dans ce manuscrit sont consacrées à la construction rapide d'un système de reconnaissance automatique de la parole pour des langues peu dotées. Les principales contributions de ce travail de thèse sont résumées ci-dessous. Pour valider nos méthodes, nous avons choisi deux langues peu dotées : le vietnamien, parlé par 67,4 millions de personnes et le khmer, parlé par 13,3 millions de personnes.

Tout d'abord, le recueil des ressources lexicales (vocabulaire, dictionnaire de prononciation) pour des langues peu dotées a été présenté. En ce qui concerne le recueil d'un vocabulaire, puisque la technique de construction automatique d'un vocabulaire à partir d'un corpus de texte s'avère difficile à cause de la faible qualité de la source de données textuelles et la difficulté de segmentation en mots pour des langues non-segmentées, notre approche consiste à récupérer un vocabulaire à partir de ressources lexicales existantes. À son tour, ce vocabulaire sert à segmenter des phrases dans un corpus de texte. Le vocabulaire peut être enrichi et limité en fonction du nombre d'occurrences de mots dans le corpus de texte segmenté. En appliquant cette méthode, un vocabulaire de 6 800 syllabes et 20 000 mots en vietnamien et un vocabulaire de 16 000 mots en khmer sont recueillis dans notre travail. Pour la construction d'un dictionnaire de prononciation dans une langue peu dotée, nous avons abordé deux approches automatiques différentes (à base de règles et à base de graphèmes) dont l'approche à base de règles a été appliquée en vietnamien tandis que l'approche à base de graphèmes a été utilisée pour le khmer. La sélection de méthodes pertinentes pour chaque langue peu dotée dépend fortement de la connaissance phonétique de la langue.

Pour construire un système de reconnaissance automatique de la parole, une grande quantité de données textuelles et signaux sont nécessaires pour entraîner les modèles sous-jacents et

tester les performances des systèmes. En ce qui concerne la collecte facilitée des signaux de parole en langue peu dotée, nous avons adapté la plateforme du logiciel EMACOP en Unicode pour manipuler respectivement les caractères dans plusieurs langues. En effet, environ 30 heures de signaux (dont 14 heures de signaux d'apprentissage du système de reconnaissance vocale) en vietnamien et 3 heures de signaux en khmer ont été enregistrés par le logiciel EMACOP Unicode. Pour le recueil d'un grand corpus de texte, nous avons proposé une méthode générique de récupération et de traitement de corpus de textes. Une boîte à outils générique « multilingue » nommé *CLIPS-Text-Tk* contenant des outils de traitement recyclables et adaptables a été développée. Pour une nouvelle langue peu dotée, nous pouvons hériter de tous les outils indépendants de la langue et adapter rapidement les outils dépendants de la langue. La difficulté de récupération d'un corpus de documents et les problèmes spécifiques de traitement de documents textuels pour des langues peu dotées tels que la normalisation de documents, la segmentation de texte, le filtrage de phrases à base d'un vocabulaire, ont été abordés dans notre travail.

Par ailleurs, nous avons constaté que les informations redondantes existant dans les documents récupérés à partir d'une même source de l'Internet influencent la taille du corpus de texte obtenu et la qualité du modèle de langage. En effet, avec la suppression des informations redondantes, une réduction d'environ 50% sur la taille du corpus d'apprentissage et d'environ 22% sur la perplexité des modèles de langage du vietnamien ont été obtenues. Après avoir traité le corpus de texte, la quantité de données textuelles pouvant servir à l'apprentissage d'un modèle de langage statistique était d'environ 400 Mo (soit 5 millions de phrases) pour le vietnamien et 97 Mo (soit 1,1 millions de phrases) pour le khmer. Pour entraîner les modèles de langage, ces corpus de textes peuvent être filtrés, en fonction d'un vocabulaire, par une des méthodes de filtrage proposées.

En ce qui concerne la modélisation acoustique, nous avons montré l'intérêt des approches de modélisation acoustique crosslingue indépendantes et dépendantes du contexte. Pour construire le tableau de correspondance phonémique source/cible, des mesures de similarité entre des unités acoustique-phonétiques différentes (phonèmes, polyphones, groupes de polyohones, etc) ont été proposées. Par la suite, en fonction du tableau de correspondance phonémique source/cible obtenu avec ces mesures, les modèles acoustiques (indépendants ou dépendants du contexte) initiaux de la langue cible sont dérivés en dupliquant les modèles acoustiques correspondants de la langue source ; ces modèles initiaux sont ensuite adaptés avec une quantité réduite de signaux en langue cible. Les résultats expérimentaux sur le vietnamien ont, en effet, montré qu'avec quelques heures de signaux d'adaptation en langue cible, la méthode de modélisation crosslingue *indépendante* du contexte est la plus efficace pour bootstrapper un système dans une langue cible. Cependant, quand nous avons plus de signaux d'adaptation, la méthode de modélisation crosslingue *dépendante* du contexte est meilleure.

Nous avons testé également l'approche de modélisation acoustique à base de graphèmes. Pour les langues peu dotées qui ne possèdent pas encore un dictionnaire phonétique telle que le khmer, cette approche est une bonne solution. Nous avons présenté une méthode d'initialisation des modèles acoustiques graphémiques par une prédétection de frontières de mots. En appliquant cette méthode d'initialisation sur le vietnamien, une amélioration significative de

41% de taux d'exactitude absolu a été obtenue par rapport à la méthode d'initialisation uniforme classique. De plus, les résultats expérimentaux sur la modélisation acoustique graphémique dépendante du contexte ont montré le potentiel de l'approche à base de graphèmes lorsque aucun dictionnaire de prononciation n'est disponible.

Perspectives

Pour la poursuite de ce travail, nous envisageons tout d'abord d'améliorer les techniques de portabilité et d'adaptation des modèles acoustiques vers une nouvelle langue. Dans le chapitre 3, nous avons proposé une méthode d'estimation de similarité entre deux polyphones. Cette similarité a été estimée en fonction des contextes gauche et droit des polyphones (au niveau du phonème). Cependant, dans chaque modèle polyphonique, l'influence du contexte est différente suivant les états du HMM considérés. Par exemple, le contexte gauche influence fortement l'état au début du modèle polyphonique mais il influence plus faiblement l'état à la fin du modèle. Cette remarque a été appliquée dans les plusieurs techniques de modélisation acoustique dépendante du contexte comme la technique de *tree-based state-tying* [Young 1994]. Ainsi, l'estimation de similarité entre deux polyphones peut être considérée au niveau des états du HMM. Par ailleurs, la méthode d'estimation de similarité des groupes de polyphones peut aussi être améliorée par une étude comparative entre la mesure en cours et d'autres mesures de distance entre deux groupes, par exemple la distance entre centroïdes des groupes, la minimisation de distance entre les deux éléments les plus éloignés des deux groupes, etc .

De plus, l'application des techniques d'adaptation au locuteur à l'adaptation des modèles acoustiques crosslingues est également à considérer pour améliorer la performance des systèmes obtenus. Par exemple, une expérience comparative entre des méthodes d'adaptation connues telles que MLLR, MAP, MLLR+MAP,… nous permettrait d'optimiser les performances de nos systèmes selon la quantité de données d'adaptation en langue cible.

Pour la reconnaissance automatique de la parole en vietnamien, la modélisation acoustique présentée dans ce manuscrit était indépendante du ton. En effet, cette modélisation doit être améliorée dans le futur car les tons vietnamiens jouent un rôle phonétique important et ils influencent l'évolution temporelle de la syllabe vietnamienne. Par ailleurs, le débat n'est pas clos sur la sélection de l'unité de reconnaissance la plus convenable (mot ou syllabe) pour la RAP grand vocabulaire en vietnamien. Dans notre expérimentation, nous avions essayé, dans un premier temps, d'implémenter un système de reconnaissance vocale à base de mots afin de le comparer avec le système à base de syllabes. Cependant, à cause de la relativement faible quantité du corpus de texte utilisé (400 Mo), les tests préliminaires ont montré que le système syllabique était meilleur que le système de mots. Cependant, nous espérons que quand la taille du corpus de texte d'apprentissage sera augmentée, la performance du système de mots augmentera proportionnellement.

En ce qui concerne la reconnaissance automatique de la parole en khmer, bien que la performance du système à base de graphèmes construit dans notre travail soit acceptable, nous espérons qu'avec le recueil d'un dictionnaire phonétique de bonne qualité, la performance du

système sera augmentée. En plus, nous continuons à collecter des documents de texte et à enregistrer un grand corpus de signaux en khmer pour améliorer la qualité des modèles acoustiques et de langage.

Les perspectives à plus long terme concernent la validation de nos méthodes de construction rapide d'un système de reconnaissance vocale pour d'autres langues, par exemple les langues d'Europe de l'Est peu dotées, les langues africaines peu dotées. D'autre part, une classification des langues peu dotées en groupes de langues selon des problèmes spécifiques à travers des langues nous permettra de proposer des méthodes génériques appropriées pour chaque groupe de langues. Par exemple, plusieurs systèmes d'écriture de la région d'Asie du Sud-Est ont les mêmes problèmes de traitement de textes tels que la segmentation de textes, la romanisation des caractères ; d'autre part, les langues tonales comme le vietnamien, le thaï ont le même problème de modélisation des tons, ...

Nous envisageons par ailleurs d'élargir nos travaux sur les langues peu dotées au domaine de la traduction automatique de la parole (*Speech-to-Speech Translation*). Des premières pistes sur ce thème ont été présentées dans [Besacier 2006]. Dans le domaine de la traduction, les méthodes statistiques existantes, nécessitant de grandes quantités de corpus bilingues alignés, risquent de vite montrer leurs limites avec des langues mal dotées ; des travaux de recherche utilisant des approches différentes sont donc nécessaires pour atteindre des performances acceptables sur un grand nombre de langues.

Annexe A

Evaluation automatique de la similarité entre phonèmes

1. *PESCORING* – Outil d'évaluation automatique de distance entre phonèmes source/cible

Pour évaluer le taux de reconnaissance de mots du système, nous utilisons souvent l'outil SCLITE *(speech recognition scoring package)* développé à l'Institut National des Standards et des Technologies – Etats Unis (NIST). SCLITE[1] utilise un algorithme de programmation dynamique qui essaie de minimiser la fonction de distance Levenshtein entre deux chaînes de texte. Cependant, en alignant les hypothèses et les références au niveau phonémique par l'outil SCLITE, nous avons parfois de mauvais alignements.

Nous avons donc développé un outil d'estimation automatique de similarité phonémique nommé PESCORING pour :

- générer automatiquement une matrice de confusion de phonèmes source/cible ;

- construire d'un tableau de correspondance phonémique source/cible ;

- évaluer le taux de reconnaissance phonémique (Phone Error Rate - PER) du système de décodage acoustico-phonétique.

Nous rappelons que pour évaluer automatiquement la fonction de distance entre deux phonèmes source/cible, une quantité limitée de données vocales étiquetées en langue cible est nécessaire. Ces étiquettes temporelles doivent être disponibles dans un fichier de description type SAM [Tomlison 1991].

D'autre part, nous utilisons un décodeur acoustique-phonétique en langue source pour décoder tous les signaux en langue cible [Beyerlein 1999, Stuker 2002, Le 2005]. La série des phonèmes d'hypothèses avec ses étiquettes temporelles est sauvegardée également dans un fichier de description SAM.

Ensuite, les fichiers SAM d'étiquette de référence et les fichiers SAM d'étiquette d'hypothèse correspondant sont chargés consécutivement et alignés temporellement, avec une

[1] http://www.nist.gov/speech/tools/index.htm

comparaison « trame par trame » (figure A.1) pour accumuler les co-occurrences entre un phonème en langue source et un phonème en langue cible.

trame	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
référence	sil		ɕ	i		h	ɔ	j		a		j	sil		v		ɣ		j	
hypothèse	sil			s	i	ɔ		i	a			i	sil		v			a		sil

Figure A.1 : Alignement temporel de phonèmes en langues source/cible

Cette accumulation de co-occurrences crée une matrice *C* dont la ligne $i^{ème}$ correspond à un phonème en langue cible, la colonne $j^{ème}$ correspond à un phonème en langue source et $C_{i,j}$ est le nombre de trames pour lesquelles le phonème t_i en langue cible est reconnu (ou confondu) par le phonème s_j en langue source. Si on divise la valeur de $C_{i,j}$ par le nombre total d'occurrences du phonème t_i reconnu par tous les phonèmes en langue source, on obtient *A*, une matrice de confusion source/cible.

Soit *m* le nombre de phonèmes en langue source et *n* le nombre de phonèmes en langue cible, le nombre total d'occurrences du phonème t_i reconnu par tous les phonèmes en langue source (la valeur de normalisation) est calculée comme suit :

$$N_i = \sum_{j=1}^{m} C_{i,j} \qquad \textit{(A-1)}$$

La valeur de confusion entre le phonème s_i en langue source et le phonème t_j en langue cible est déterminée par :

$$A_{i,j} = \frac{C_{i,j}}{N_i} \qquad \textit{(A-2)}$$

En fonction de la valeur de confusion source/cible, plusieurs types de format de sortie de l'outil PESCORING sont proposés : la matrice de confusion, le tableau de correspondances phonémiques source/cible (*phone mapping table*), le taux de phonème correct (PER), … PESCORING est un outil « open source »[1] sous la licence GPL[2].

[1] http://www-clips.imag.fr/geod/User/viet-bac.le/outils/
[2] http://www.gnu.org/copyleft/gpl.html

2. Exemple de génération de tableaux de correspondance des phonèmes source/cible par la méthode automatique

En utilisant un décodeur acoustico-phonétique (DAP) français et un DAP multilingue pour décoder un corpus de signaux en langue cible (vietnamien), un tableau de correspondance des phonèmes/source cible (français/vietnamien et multilingue/vietnamien) est généré automatiquement à l'aide de l'outil PESCORING (tableau A.1).

Phonème du vietnamien	Décodeur français FR		Décodeur multilingue MM7	
	Phonème FR	Confusion	Phonème MM7	Confusion
7	in	0,20	E	0.09
7X	a	0,18	ai	0.10
E	E	0,49	e	0,17
EX	a	0,25	ai	0,23
G	g	0,21	g	0,14
M	e	0,13	M	0,21
M7	R	0,17	E	0,07
NG	n	0,10	NG	0,39
NJ	NJ	0,17	NJ	0,19
O	an	0,26	au	0,11
OX	an	0,29	A	0,11
SIL	SIL	0,77	SIL	0,67
X	R	0,36	X	0,37
a	a	0,34	ai	0,13
aX	a	0,30	a	0,10
b	b	0,27	b	0,59
c	t	0,37	tS	0,33
d	d	0,28	d	0,27
e	e	0,39	e	0,18
f	f	0,53	f	0,61
h	R	0,25	X	0,20
i	i	0,50	i	0,30

Phonème du vietnamien	Décodeur français FR		Décodeur multilingue MM7	
	Phonème FR	Confusion	Phonème MM7	Confusion
ie	e	0,40	i	0,19
j	e	0,27	uei	0,07
k	k	0,43	k	0,48
l	l	0,40	l	0,46
m	m	0,26	m	0,46
n	n	0,22	n	0,38
o	o	0,31	o	0,17
p	a	0,20	a	0,21
s	s	0,73	ss	0,22
ss	s	0,72	ss	0,34
t	t	0,37	t	0,37
th	t	0,32	th	0,28
tr	t	0,37	tS	0,29
u	o	0,36	u	0,17
uo	o	0,60	u	0,25
v	v	0,38	v	0,19
w	o	0,16	au	0,10
z	z	0,48	z	0,30
zr	z	0,48	z	0,36

Tableau A.1 : Tableau de correspondances des phonèmes FR/VN et MM7/VN

Annexe B

Liste des caractères khmers romanisés pour nos travaux

Caractère Unicode khmer	Caractère de romanisation	Nom anglais du caractère khmer
ក	Ka	KHMER LETTER KA
ខ	KHa	KHMER LETTER KHA
គ	Ko	KHMER LETTER KO
ឃ	KHo	KHMER LETTER KHO
ង	NGo	KHMER LETTER NGO
ច	Ca	KHMER LETTER CA
ឆ	CHa	KHMER LETTER CHA
ជ	Co	KHMER LETTER CO
ឈ	CHo	KHMER LETTER CHO
ញ	NYo	KHMER LETTER NYO
ដ	Da	KHMER LETTER DA
ឋ	TTHa	KHMER LETTER TTHA
ឌ	Do	KHMER LETTER DO
ឍ	TTHo	KHMER LETTER TTHO
ណ	NNo	KHMER LETTER NNO
ត	Ta	KHMER LETTER TA
ថ	THa	KHMER LETTER THA
ទ	To	KHMER LETTER TO
ធ	THo	KHMER LETTER THO
ន	No	KHMER LETTER NO
ប	Ba	KHMER LETTER BA
ផ	PHa	KHMER LETTER PHA
ព	Po	KHMER LETTER PO
ភ	PHo	KHMER LETTER PHO
ម	Mo	KHMER LETTER MO
យ	Yo	KHMER LETTER YO
រ	Ro	KHMER LETTER RO

Caractère Unicode khmer	Caractère de romanisation	Nom anglais du caractère khmer
ល	Lo	KHMER LETTER LO
វ	Vo	KHMER LETTER VO
ឝ	SHa	KHMER LETTER SHA
ឞ	SSo	KHMER LETTER SSO
ស	Sa	KHMER LETTER SA
ហ	Ha	KHMER LETTER HA
ឡ	La	KHMER LETTER LA
អ	Qa	KHMER LETTER QA
ឣ	QAQ	KHMER INDEPENDENT VOWEL QAQ
ឤ	QAA	KHMER INDEPENDENT VOWEL QAA
ឥ	QI	KHMER INDEPENDENT VOWEL QI
ឦ	QII	KHMER INDEPENDENT VOWEL QII
ឧ	QU	KHMER INDEPENDENT VOWEL QU
ឨ	QUK	KHMER INDEPENDENT VOWEL QUK
ឩ	QUU	KHMER INDEPENDENT VOWEL QUU
ឪ	QUUV	KHMER INDEPENDENT VOWEL QUUV
ឫ	RY	KHMER INDEPENDENT VOWEL RY
ឬ	RYY	KHMER INDEPENDENT VOWEL RYY
ឭ	LY	KHMER INDEPENDENT VOWEL LY
ឮ	LYY	KHMER INDEPENDENT VOWEL LYY
ឯ	QE	KHMER INDEPENDENT VOWEL QE
ឰ	QAI	KHMER INDEPENDENT VOWEL QAI
ឱ	QOO1	KHMER INDEPENDENT VOWEL QOO TYPE ONE
ឲ	QOO2	KHMER INDEPENDENT VOWEL QOO TYPE TWO
ឳ	QAU	KHMER INDEPENDENT VOWEL QAU
◌ា	AA	KHMER VOWEL SIGN AA
◌ិ	I	KHMER VOWEL SIGN I
◌ី	II	KHMER VOWEL SIGN II
◌ឹ	Y	KHMER VOWEL SIGN Y
◌ឺ	YY	KHMER VOWEL SIGN YY
◌ុ	U	KHMER VOWEL SIGN U
◌ូ	UU	KHMER VOWEL SIGN UU
◌ួ	UA	KHMER VOWEL SIGN UA
◌ើ	OE	KHMER VOWEL SIGN OE
◌ឿ	YA	KHMER VOWEL SIGN YA
◌ៀ	IE	KHMER VOWEL SIGN IE

Caractère Unicode khmer	Caractère de romanisation	Nom anglais du caractère khmer
◌ေ	E	KHMER VOWEL SIGN E
◌ែ	AE	KHMER VOWEL SIGN AE
◌ៃ	AI	KHMER VOWEL SIGN AI
◌ោ	OO	KHMER VOWEL SIGN OO
◌ៅ	AU	KHMER VOWEL SIGN AU
◌ំ	NIKAHIT	KHMER SIGN NIKAHIT
◌ះ	REAHMUK	KHMER SIGN REAHMUK
◌ः	YUUKALEAPINTU	KHMER SIGN YUUKALEAPINTU
។ល។	BEYYAL	KHMER SIGN BEYYAL
៛	RIEL	KHMER CURRENCY SYMBOL RIEL
០	SON	KHMER DIGIT ZERO
១	MUOY	KHMER DIGIT ONE
២	PII	KHMER DIGIT TWO
៣	BEI	KHMER DIGIT THREE
៤	BUON	KHMER DIGIT FOUR
៥	PRAM	KHMER DIGIT FIVE
៦	PRAM MUOY	KHMER DIGIT SIX
៧	PRAM PII	KHMER DIGIT SEVEN
៨	PRAM BEI	KHMER DIGIT EIGHT
៩	PRAM BUON	KHMER DIGIT NINE

Tableau C.1 : Liste des caractères khmers romanisés

Annexe C

Liste de mes publications personnelles

[Le 2006] V-B. Le, L. Besacier, T. Schultz, *Acoustic-Phonetic Unit Similarities for Context Dependent Acoustic Model Portability*, 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06), Toulouse, France, 15-19 May 2006.

[Besacier 2006] L. Besacier, V-B. Le, C. Boitet, V. Berment, *ASR and translation for under-resourced languages*, 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06), Toulouse, France, 15-19 May 2006.

[Besacier 2005] L. Besacier, V-B. Le, E. Castelli, S. Sam, L. Protin, *Reconnaissance Automatique de la Parole pour des Langues peu Dotées : Application au Vietnamien et au Khmer*, Atelier « TALN et langues peu dotées », TALN'05, vol 2, pp. 207-217, Dourdan, France, 6-10 juin 2005.

[Le 2005a] V-B. Le, L. Besacier, *First steps in fast acoustic modeling for a new target language: application to Vietnamese*, 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05), vol. 1, pp. 821-824, Philadelphia, USA, 19-23 March 2005.

[Le 2005b] V-B. Le, D-D. Tran, L. Besacier, E. Castelli, J-F. Serignat, *First steps in building a large vocabulary continuous speech recognition system for Vietnamese*, 3rd International Conference in Computer Science (RIVF'05), pp. 330-333, Can Tho, Vietnam, 21-24 February 2005.

[Villaseñor-Pineda 2005] L. Villaseñor-Pineda, V-B. Le, M. Montes-y-Gómez, M. Pérez-Coutiño, *Toward Acoustic Models for Languages with Limited Linguistic Resources*, CICLing'05, Lecture Notes in Computer Science, ISBN: 3-540-24523-5, vol. 3406/2005, pp. 433-436, Springer-Verlag, February 2005.

[Tran 2004a] D-D. Tran, E. Castelli, V-B. Le, V-L. Trinh, L. Besacier, *Vietnamese Speech Corpus and Applications*, 6th PAN-Asiatic International Symposium on Languages and Linguistics, Hanoi, Vietnam, November 2004.

[Tran 2004b] D-D. Tran, V-B. Le, V-L. Trinh, E. Castelli, L. Besacier, *Spoken and Written Language Resources Construction for Vietnamese*, 2nd National Conference in Information Technology, Da Nang, Vietnam, 13-14 August 2004.

[Le 2004] V-B. Le, D-D. Tran, E. Castelli, L. Besacier, J-F. Serignat, *Spoken and written language resources for Vietnamese*, 4th International Conference on Language Resources And

Evaluation (LREC'04), pp. 599-602, Lisbon, Portugal, 26-28 May 2004.

[Tran 2004c] D-D. Tran, V-B. Le, E. Castelli, V-L. Trinh, *Building a large Vietnamese Speech Database*, Vietnamese Journal of Science and Technology, vol. 46+47/2004, pp. 13-17, Vietnam, February 2004.

[Le 2003a] V-B. Le, B. Bigi, L. Besacier, E. Castelli, *Using the Web for fast language model construction in minority languages*, 8th European Conference on Speech Communication and Technology (Eurospeech'03), pp. 3117-3120, Geneva, Switzerland, 1-4 September 2003.

[Le 2003b] V-B. Le, *Construire rapidement les modèles de langage pour des langues minoritaires*, Rencontres Jeunes Chercheurs en parole (RJC'03), Grenoble, France, Septembre 2003.

Annexe D

Articles joints

1. Modélisation du langage pour des langues peu dotées

V-B. Le, B. Bigi, L. Besacier, E. Castelli, *Using the Web for fast language model construction in minority languages*, 8th European Conference on Speech Communication and Technology (Eurospeech'03), pp. 3117-3120, Geneva, Switzerland, 1-4 September 2003.

2. Recueil de ressources écrites et orales pour une langue mal dotée

V-B. Le, D-D. Tran, E. Castelli, L. Besacier, J-F. Serignat, *Spoken and written language resources for Vietnamese*, 4th International Conference on Language Resources And Evaluation (LREC'04), pp. 599-602, Lisbon, Portugal, 26-28 May 2004.

3. Modélisation acoustique crosslingue indépendante du contexte

V-B. Le, L. Besacier, *First steps in fast acoustic modeling for a new target language: application to Vietnamese*, 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05), vol. 1, pp. 821-824, Philadelphia, USA, 19-23 March 2005.

4. Evaluation de similarités des unités acoustique-phonétiques - Application à la modélisation acoustique crosslingue dépendante du contexte.

V-B. Le, L. Besacier, T. Schultz, *Acoustic-Phonetic Unit Similarities for Context Dependent Acoustic Model Portability*, 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06), Toulouse, France, 15-19 May 2006.

Using the Web for fast language model construction in minority languages

Viet Bac LE, Brigitte BIGI*, Laurent BESACIER*, Eric CASTELLI***

* CLIPS-IMAG Laboratory, UMR CNRS 5524
** BP 53, 38041 Grenoble Cedex 9, FRANCE

email: viet-bac.le@imag.fr

Abstract

The design and construction of a language model for minority languages is a hard task. By minority language, we mean a language with small available resources, especially for the statistical learning problem. In this paper, a new methodology for fast language model construction in minority languages is proposed. It is based on the use of Web resources to collect and make efficient textual corpora. By using efficient filtering techniques, this methodology allows a quick and efficient construction of a language model with a small cost in term of computational and human resources.
Our primary experiments have shown excellent performance of the Web language models vs. newspaper language models using the proposed filtering methods on a majority language (French). Following the same way for a minority language (Vietnamese), a valuable language model was constructed in 3 month with only 15% new development to convert some filtering tools.

1. Introduction

There are more than 6000 languages in the world but only a small number possess the resources required for implementation of Human Language Technologies (HLT). Thus, HLT are mostly concerned by languages which have large resources available or which suddenly became of interest because of the economic or political scene. On the contrary, languages from developing countries or minorities were less treated in the past years. One way of ameliorating this "linguistic divide" is through starting research on portability of HLT for multilingual applications. This question has been increasingly discussed in the recent years. The SALTMIL[1] (Speech and Language Technology for Minority Languages), which is a Special Interest Group of ISCA, was created to promote research and development in the field of speech and language technology for lesser-used languages, particularly those of Europe. However, in SALTMIL, "minority language" mostly means "language spoken by a minority of people". We rather focus, in our work, on languages which have a "minority of resources usable in HLT". These languages are mostly those from developing countries, but can be spoken by a large population. In this paper, we will notably deal with Vietnamese, which is spoken by about 70 millions of persons, but for which very few usable electronic resources are available.

Among HLT, we are interested, in this paper, in Automatic Speech Recognition (ASR). We are currently investigating new techniques and tools for a fast portability of speech recognition systems to new languages. This topic has already been tackled in [1] and [2] but mostly for languages which already have large corpora available. Conversely, we particularly address languages, like Vietnamese, for which few signal and text resources are available. This activity includes different aspects:

- Portability of acoustic models: this can be achieved, for examples by using tools for performing a fast collection of speech signals [3] or by using Language Adaptive Acoustic Modeling [4].
- Language modeling for new languages: we propose to use web-based techniques [5] which have already shown ability to collect large amount of text corpora. For languages in which no usable text corpora exist, this is moreover the only viable approach to collect text data.
- Dictionaries: collaborative approaches like in [6] could be also proposed for ASR.

This paper addresses particularly fast language model construction for ASR. The proposed method uses the web to collect large amount of data. In section 2, we first describe our text data collection tools and the filtering techniques associated which were first developed for French language modeling. Then, in section 3, we describe the modifications implied to adapt our collecting and filtering tools to Vietnamese. Section 4 is dedicated to experiments performed to validate our methodology; for comparison purpose, perplexity figures are simultaneously given for French and Vietnamese. Finally, section 5 concludes this work and gives some perspectives.

2. Language modeling using Web resources

Language Modeling (LM) is one of the most important modules in a large vocabulary speech recognition system. Statistical language models (SLM), which describe probabilistically the constraints on word order found in language, are traditionally used. However, it is difficult to construct a SLM because we must have a large enough corpus which models all possible user input. A large corpus tends to have more contexts for each word, and thus tends to produce more accurate and robust SLMs. N-grams based model is a useful one for solving this problem. In this model, an estimate of the likelihood of a word is made solely on the identity of the N-1 preceding words in the utterance. For more details, see [7].

With the development of the Internet and its services, the WWW is the greatest information space distributed over the world, in many languages and on many topics. Web resources can be a very interesting source for spoken language modeling if we process it in an appropriate way. There are many solutions for a SLM construction using the Web. In

[1] http://www.cstr.ed.ac.uk/~briony/SALTMIL/

particular, in the domain of information retrieval, we found some "web search query" based approaches [8, 9]. In our case, these solutions can not be applied at the moment because we have no tool for automatically generating the queries.

In this section, we will describe some techniques for language model construction. First, by using a web-robot (or web-spider), we can collect and store web pages in the given language. And then, we filter and analyse them for building a text corpus. Finally, all N-grams models are estimated from this text corpus.

2.1. Web pages collecting

Documents were gathered from Internet by some web robots (among them, one was developed in our lab[2]). From some starting points on the Web, the robots can reach and find all the text documents and web pages which have a direct or indirect link with these starting points. However we must manage the Web sites (Internet domain names) accessed by the robots because we want to collect the pages and the documents in a given domain and in a given language only.

2.2. Data preparation

Some filtering techniques are needed to construct the text corpus from HTML pages. We must extract the text parts from the HTML pages and insert some document separators. The tokens <s> and </s> signal respectively the begin and end of a sentence. Web texts contain also a variety of "non-standard" token types such as digit sequences, words, acronyms and letter sequences in all capitals, mixed case words, abbreviations, roman numerals, URL's and e-mail addresses... These non-standard types cause problems for training language models. Normalizing or rewriting such text using ordinary words is a first important issue.

Then, for language modeling application, by using a compound word lexicon, we also compute compound words that are treated in the language model as one word. There are two benefits in this method: there is no biased usage of the word penalty of the recognizer and we increase the context taken into account in the language model [5]. We then regroup the common words into classes for introducing classes in the LM. The choice of the classes depends on the task-specific application. For example: country name, city name, days of the week, month... Finally, we also transcribe numbers in context (date, money, etc) to textual form (number-to-text).

2.3. Sentence filtering

There are many different solutions to extract the relevant sentences from a text corpus. Classically, we can keep all the sentences exclusively made with words of the task-specific vocabulary preliminary defined. The other method proposed in [10] is a text filtering algorithm based on character perplexity. However, it needs a "Standard Language Model" for reference and there is not any such reference model available in minority languages. We can also use the "minimal blocks" filtering method proposed in [3]. A minimal block of order n is a sequence of at least n consecutive words from the document with all words of the block in the given vocabulary.

Table 1: *List of the fixed and variable modules to adapt tools from French to Vietnamese*

Fixed modules	Variable modules
data collecting html2text token normalizing sentence splitting word splitting common word grouping data filtering	character converting case changing number2text lexicon constructing

We note that there is not a lot of research work which compares the performances of these methods. So, in the section 4, we propose, combine and compare several filtering methods applied in our experiments.

3. Language modeling using Web resources

3.1. Methodology

As we show in the previous section, using the web to construct the LMs implies to develop tools for text extraction and text selection. This development can be carried out specifically for each language but this work is laborious and time consuming. In the context of genericity, producing reusable components for language-and-task-specific development is an important goal. The aim is to create a set of tools that would represent the common text normalization for many languages. Consequently, we decided to construct a lot of small tools for French (the "source" language) and to estimate time consuming to adapt tools from French to Vietnamese (the minority target language).

First, because we want to construct LMs in multiple languages, we must choose a unique character set for encoding all the documents and for covering all languages possible. Universal Character Set (UCS) which is a part of Unicode international standard[3] provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. We have chosen Unicode standard for encoding all characters of our corpora. But there are hundreds of different character sets for encoding a character of the Web documents. We note that the French LM tools we inherited are single byte (ASCII) based. Indeed, we construct a tool to convert a character in several character sets to the Unicode (UTF-8 encoding system).

Secondly, we decide to split the original tool for source language (French) to a set of modules (see table 1). And then, we have determined what are:

- *Fixed modules*: these are the modules which do not depend on the language.
- *Variable modules*: these are the modules which depend on each language.

This splitting and determination work is really important. For a new language modeling, we will inherit all the *fixed modules* and rapidly adapt the *variable modules* to that language. It will economize the time consuming to build a complete language model. We propose these tools available on demand for any person who is interested in.

[2] http://slmg-index.imag.fr

[3] http://unicode.org/

3.2. Application to Vietnamese language modeling

3.2.1. Dictionaries for ASR

To build a language model and filter out the documents, it is necessary to have a vocabulary containing words to include in the language model. This vocabulary can come from a variety of resources in the Internet. We can use a bilingual or multilingual dictionary for generating this vocabulary.

In fact, there are many methods to construct a dictionary. In the context of the Papillon[4] project, the construction of a lexical base for a new language may take several different ways depending on where the author has to start: collaborative approaches [6], dictionary recycling [11]... This project aims at creating a multilingual lexical database covering among others English, French, Japanese, Malay, Lao, Thai and Vietnamese.

From this Papillon project, we got a dictionary for Vietnamese language (French-Vietnamese and Vietnamese-French). Then, we filtered this dictionary to have a list of more than 40,000 unique words in Vietnamese: compound words, borrowed words and isolated words. By taking only the most frequent words, we can discount this size of vocabulary to 20,000 words. These were the highest frequency words which occur in the documents of our training corpus.

3.2.2. Data collecting

Text corpus for language modeling cannot be collected easily in the minority languages for some reasons:

- There are less pages and websites than in the majority languages.
- The debit of communication is often very low (several kilobits per second).

Consequently, we can not crawl all of the websites but we must focus on some which have more pages and higher debit than the others. So, a non negligible time was used to find out the websites to collect.

There are about 2500 Vietnamese websites in Vietnam which publish: daily news, information, entertainment, ecommerce, forum... The daily news web pages introduced a constraint in the data collection, since we had to regularly access the same sites to get an acceptable amount of data. This is the major difference with web data collection for a majority language like French or English where there are enough web pages that can be collected at a given time.

3.2.3. Text-corpus filtering techniques

Our first positive result was the porting of our LM tools to a new language. Indeed, we must have only a short time to modify and to adapt these variable modules to a new language. We have built a language model for Vietnamese in only three months with this methodology. A comparison of this minority LM (for Vietnamese) with a majority one (for French) is proposed in the next section.

4. Experiments

4.1. Training corpora

The French data collection (called WebFR4) is a very large corpus containing a few less than 6 millions web pages representing 44 GB. This corpus was gathered in December 2000 and the collect was restricted to the *.fr* domain. The set of exploitable resources (after data preparation) is made of 12 GB, i.e. 184,738,292 sentences. This set of text is a very huge corpus and it is difficult to use the entire corpus to learn the model. In our experiments, we choose to use only the first 700MB of this prepared data corresponding to a size comparable to what was obtained for Vietnamese.

To compare web-based language models with conventional language models, we also used a corpus extracted from the newspaper *Le Monde* from year 1997 to 2001. This corpus is made of 716 MB, i.e. 4,323,629 sentences. This newspaper corpus is used by the majority of French ASR systems.

The Vietnamese data collection is a corpus representing more than 2.5 GB of web pages. After data preparation, the text corpus is made of 858 MB, i.e. 10,020,267 sentences.

4.2. Test corpora

The French test corpus is made up of two dialogs extracted from the NESPOLE! project[5] database[5]. They are related to a client/agent discussion for organizing holidays in Italy. Only the 216 client turns were kept for our experiments. The French vocabulary (20,000 words) is made of this task specific words plus the most frequent French words of WebFR4 corpus.

The Vietnamese test corpus is a translation of the French corpus. The Vietnamese vocabulary (20,000 words) is obtained with the methodology described in 3.2.1.

4.3. Filtering and LM construction

In these experiments, we tried some solutions to filter the training corpora. In all cases, we selected sentences without size restriction. To filter, we tested the following solutions:

1. *all-sentences*: take all the text corpus (without sentence filtering).
2. *block-based*: take only blocks which have at least 5 in-vocabulary words by block.
3. *sentence-based*: take all sentences containing only invocabulary words (no unknown words).
4. *hybrid*: take all sentences containing only in-vocabulary words (3) and apply minimal blocks filtering (2) on the rejected sentences.

To learn our language models, we use the SRILM toolkit [12] with a Good-Turing discounting and Katz backoff for smoothing method. It is very important to note that with this toolkit, the unknown words are removed in our case, since we are in the framework of closed-vocabulary models.

4.4. Results

The perplexities of the language models with these data filtering solutions are given in table 2.

The last two filtering methods have the best perplexities in our dialogue test because test corpus contains many short sentences.

Table 2 also shows that the perplexities of the language models according to corpus collected from Web are better than from journalistic source in our context of dialogue test. That means that the Web is a very rich source for spoken

[4] http://bushido.imag.fr/papillon/

[5] http://nespole.itc.it/

language modeling and that it can be successfully applied to model minority languages like Vietnamese even if the correspondence between perplexities of Vietnamese and French language models is not very significant here because each language have a particular characteristic.

Table 2: *Perplexities of the language models*

Exp.	FR: Newspaper		FR: Web		VN: Web	
	Size (MB)	PPL	Size (MB)	PPL	Size (MB)	PPL
all	716	673	686	539	858	260
block	642	796	366	637	667	359
sent.	92	**513**	156	580	370	**252**
hybrid	644	687	411	**509**	729	259

4.5. Redundancy

We also noticed that the Vietnamese Web resources contain some redundant information (menus, references, advertisements, announcements...) which is repeated in different pages. This is due to the day by day collecting of daily news which may have a direct influence on the performance of the language modeling.

Therefore, we tried to evaluate this redundancy in the part of the Vietnamese corpus from Vietnam News Daily[6] website (called *VnExpress*). So, we applied a redundancy filtering method before html2text module. The new perplexity figures with and without this redundancy filtering are given in table 3.

Table 3: *Influence of the redundant information*

Exp.	VN: Web original filter		VN: Web redun. filter	
	Size (MB)	Size (MB)	Size (MB)	PPL
all	868	260	402	201
block	667	359	357	282
sent.	370	**252**	226	**195**
hybrid	729	259	373	199

By filtering the redundant information contained in the web pages collected from the same site, the training corpus size is reduced by 54% in our experiments. On the other hand, the perplexity value is significantly improved by 26%.

5. Conclusions and perspectives

An effective methodology for fast language model construction in minority languages is introduced in this paper. It consists in collecting Web sites and filtering the web pages using some generic tools. This methodology has been tested and validated using the Vietnamese minority language. In a first step, we have built the set of tools for the French majority language and validated the Web-based language model comparing to the classical newspaper-based language model. In a second step, we have adapted our tools to Vietnamese and we have defined which modules are fixed and which are specific of the target language.

By collecting regularly two daily news web sites, a language model for Vietnamese was obtained in only three months. In our experiments, we have also presented an evaluation of perplexities for some proposed filtering methods in majority language (French) and in minority language (Vietnamese).

As future subjects, we will focus on the task-dependency language modeling (for example using a Web search engine) and we will improve these data filtering methods. Fast construction of acoustic model for minority language is also a very important and challenging part of our future work.

6. References

[1] J. Kunzmann, K. Choukri, E. Jahnke, A. Kiessling, K. Knill, L. Lamel, T. Schultz, and S. Yamamoto, "Portability of automatic speech recognition technology to new languages: Multilinguality issues and speech/text resources," in ASRU, Madonna di Campiglio, Italy, 2001.

[2] L. Lamel, "Some issues in speech recognizer portability," in Workshop on portability issues in human language technologies (LREC), 2002.

[3] D. Vaufreydaz, C. Bergamini, J. F. Serignat, L. Besacier, and M. Akbar, "A new methodology for speech corpora definition from internet documents," in LREC, vol. I, Athens, Greece, 2000, pp. 423–426.

[4] T. Schultz and A.Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," Speech Communication, vol. 35, no. 1-2, pp. 31–51, 2001.

[5] D. Vaufreydaz, L. Besacier, C. Bergamini, and R. Lamy, "From generic to task-oriented speech recognition: French experience in the nespole! european project," in ITRW Workshop on Adaptation Methods for Speech Recognition, Sophia-Antipolis, France, 2001.

[6] V. Berment, "Several technical issues for building new lexical bases," in Workshop Papillon, Tokyo, Japon, 2002.

[7] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer Speech and Language*, vol. 10, pp. 187–228, 1996.

[8] R. Ghani, R. Jones, and D. Mladenic, "Building minority language corpora by learning to generate web search queries, Tech. Rep. CMU-CALD-01-100, 2001.

[9] G. A. Monroe, J. C. French, and A. L. Powell, "Obtaining language models of web collections using query-based sampling techniques," in HICSS, 2002.

[10] R. Nisimura et al., "Automatic ngram language model creation from web resources," in Eurospeech, 2001.

[11] H. Doan-Nguyen, "Techniques génériques d'accumulations d'ensembles lexicaux structurés à partir de ressources dictionnairiques informatisées multilingues hétérogènes," Ph.D. dissertation, INPG, Grenoble (France), 1998.

[12] A. Stolcke, "Srilm - an extensible language modeling toolkit," in International Conference Spoken Language Processing, Denver, Colorado, 2002.

[6] http://vnexpress.net/

SPOKEN AND WRITTEN LANGUAGE RESOURCES FOR VIETNAMESE

Viet-Bac Le[*], Do-Dat Tran[*, **], Eric Castelli[**], Laurent Besacier[*], Jean-François Serignat[*]

[*] CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, FRANCE

[**] International Research Center MICA
1 Dai Co Viet, Hanoi, VIETNAM

Email: (Viet-Bac.Le, Do-Dat.Tran, Eric.Castelli, Laurent.Besacier, Jean-Francois.Serignat)@imag.fr

ABSTRACT

This paper presents an overview of our activities for spoken and written language resources for Vietnamese implemented at CLIPS-IMAG Laboratory and International Research Center MICA. A new methodology for fast text corpora acquisition for minority languages which has been applied to Vietnamese is proposed. The first results of a process of building a large Vietnamese speech database (VNSpeechCorpus) and a phonetic dictionary, which is used for automatic alignment process, are also presented.

1. INTRODUCTION

There are more than 6000 languages in the world but only a small number possess the resources required for implementation of Human Language Technologies (HLT). Thus, HLT are mostly concerned by languages which have large resources available or which suddenly became of interest because of the economic or political scene. On the contrary, languages from developing countries or minorities were less treated in the past years. One way of ameliorating this "linguistic divide" is through starting research on portability of HLT for multilingual applications. This question has been increasingly discussed in the recent years, for instance in the SALTMIL[1] (Speech and Language Technology for Minority Languages) group. However, in SALTMIL, "minority language" mostly means "language spoken by a minority of people". We rather focus, in our work, on languages which have a "minority of resources usable in HLT". These languages are mostly from developing countries, but can be spoken by a large population. In this paper, we will notably deal with Vietnamese, which is spoken by about 70 millions of persons, but for which very few usable electronic resources are available.

Among HLT, we are interested in Automatic Speech Recognition (ASR). We are currently investigating new techniques and tools for a fast portability of speech recognition systems to new languages like Vietnamese, for which few signal and text resources are available. This activity includes different aspects:

- Portability of acoustic models: this can be achieved, for example by using tools for performing a fast collection of speech signals (Vaufreydaz et al., 2000) or by using Language Adaptive Acoustic Modeling (Schultz & Waibel, 2001).
- Language modeling for new languages: we propose to use web-based techniques which have already shown ability to collect large amount of text corpora. For languages in which no usable text corpora exist, this is moreover the only viable approach to collect text data (Le et al., 2003).
- Dictionaries: collaborative approaches like in (Berment, 2002) could be also proposed for ASR.

This paper presents an overview of our activities for spoken and written language resources for Vietnamese. Firstly, a new methodology for fast text corpora acquisition for minority languages is proposed. With more than 800MB of text size, our Vietnamese text corpus will be used for many different applications in language processing domain: language modeling, spoken corpora definition, information retrieval...

Secondly, we describe the characteristics of a large Vietnamese language speech database (called VNSpeechCorpus) which will contain about 100 hours recorded in both quiet and office environment from 50 native speakers. VNSpeechCorpus will be available on CD-ROM and could be distributed by the ELRA-ELDA association. Automatic alignments could be provided too.

Finally, a phonetic dictionary was obtained by using our *VNPhoneAnalyzer*. It is based on the phone concatenation of the initial part, final part and tone of a syllable. The symbolic representation for the various sounds and a description of articulatory features is provided by the International Phonetic Alphabet – IPA (IPA, 1999).

2. TEXT CORPORA ACQUISITION

In this section, we will describe some techniques for fast text corpora acquisition and evaluation. First, by using a web-robot (or web-spider), we can collect and store web pages in the given language. And then, these web pages were filtered and analyzed for building a text corpus. Finally, a language model was estimated from this text corpus.

2.1. Web pages collection and data preparation

Documents were gathered from Internet by some web robots (among them, one was developed in our lab[2]). From some starting points on the Web, the robots can reach and find all the text documents and web pages which have a direct or indirect link with these starting points. However we must manage the Web sites (Internet domain names) accessed by the robots because we want to collect the pages and the documents in a given domain and in a given language only.

[1] http://www.cstr.ed.ac.uk/~briony/SALTMIL

[2] http://slmg-index.imag.fr

Some filtering techniques are needed to construct the text corpus from HTML pages. Firstly, we also noticed that the Vietnamese Web resources contain some redundant information (menus, references, advertisements, announcements ...) which is repeated in different pages. This is due to the day by day collecting of daily news which may have a direct influence on the quality of the text corpus and also on the performance of the language modeling. By filtering the redundant information contained in the web pages collected from the same site, the corpus size is reduced by 54% in our experiments. On the other hand, the corpus perplexity evaluation value is significantly improved by 26%.

Secondly, the text parts from the rest of these HTML pages were extracted and some document separators were inserted. The tokens <s> and </s> signal respectively the begin and the end of a sentence. Web texts contain also a variety of "non-standard" token types such as digit sequences, words, acronyms and letter sequences in all capitals, mixed case words, abbreviations, roman numerals, URL's and e-mail addresses... Normalizing or rewriting such text using ordinary words is a first important issue.

Thirdly, because the collected documents were encoded in many encoding system (TCVN3, VNI, UCS, UTF-8 …), we must choose a unique character set for encoding all the documents. Universal Character Set (UCS) which is a part of Unicode international standard[1] provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language. We have chosen Unicode standard for encoding all characters of our corpora and we constructed a tool to convert a character in several character sets to the Unicode (UTF-8 encoding system).

Finally, there are many different solutions to extract the relevant sentences from a text corpus. Classically, we can keep all the sentences exclusively made with words of the task-specific vocabulary preliminary defined.

2.2. Experiments and evaluation

Text corpus cannot be collected easily in the minority languages for some reasons:

- There are less websites than in the majority languages.
- The transmission rate is often very low.

Consequently, we can not crawl all of the websites but we must focus on some which have more pages and higher debit than the others. So, a non negligible time was used to find out the websites to collect.

There are about 2500 Vietnamese websites in Vietnam which publish: daily news, information, entertainment, e-commerce, forum... The daily news web pages introduced a constraint in the data collection, since we had to regularly access the same sites to get an acceptable amount of data. This is the major difference with web data collection for a majority language like French or English where there are enough web pages that can be collected at a given time.

[1] http://unicode.org

The Vietnamese data collection is composed of more than 2.5 GB of web pages. After data preparation, the text corpus is made of 868 MB, i.e. 10,020,267 sentences.

We construct a Vietnamese language model for estimating this text corpus. In these experiments, we tried 4 different solutions to filter the text corpora: *all-sentences* (without sentence filtering)*; block-*, *sentence-based* and *hybrid* (take blocks and sentences containing only in-vocabulary words). In all cases, we selected sentences without size restriction (for more detail, see Le et al., 2003). The perplexity measures for Vietnamese are given and compared to a same size as French text corpus in table 1.

	VN: Web original filter		**VN: Web** redun. filter		**FR: Web**	
Expe.	**Size** (MB)	**PPL**	**Size** (MB)	**PPL**	**Size** (MB)	**PPL**
all	868	260	402	201	686	539
block	667	359	357	282	366	637
sent.	370	**252**	226	**195**	156	580
hybrid	729	259	373	199	411	**509**

Table 1: Perplexity of the language models

Table 1 also shows that our Web-based methods can be successfully applied to majority and minority languages like French and Vietnamese even if the correspondence between perplexities of Vietnamese and French language models is not very significant here because each language have a particular characteristic.

3. VIETNAMESE LANGUAGE SPEECH DATABASE (VNSPEECHCORPUS)

3.1. Text Corpus

Two phases of collecting text data were implemented in our project. In the first phase, data is collected by some experts in order to ensure the desired requirements. And in the second phase, data is extracted automatically with one desired distribution of acoustic units from the web corpora obtained in part 2.

Beside database of phonemes, digits, application words, other data including sentences and paragraphs were collected from different resources such as stories, books, and web documents... The selected data covers different fields and contains many dialogs and short paragraphs. This initial data then was manipulated and was divided into smaller paragraphs and conversations (about 4-6 lines/ paragraph or conversation) that help speaker to utter or read easily.

3.2. Corpus Organization

The VNSpeechCorpus contains 5 different kinds of data:

- Phonemes.
- Tones.
- Digits and string of digits.
- Application words.
- Sentences and paragraphs.

The phonemes are read by all speakers. The vowels can be read independently except two vowels *ă* /ă/ and *â* /ɤ̆/, because their sounds are only represented completely in words in which they appear, such as *ng<u>ắ</u>n (short)*, *t<u>â</u>n*

(new)… The consonants are combined with vowel *ơ* /ɤ/ and falling tone for pronouncing.

Vietnamese is a tonal language with 6 tones (Doan, 1997), for example: *ba* (three), *bà* (grandmother), *bá* (king), *bả* (bane), *bã* (waste), *bạ* (any). The speakers are asked to read the words with different tones. These words have almost the same initial and final part, but they have different tones.

The digit corpus consists of isolated digits, connected digits and natural numbers. In Vietnamese digital system, most of the digits and numbers are read or uttered with the unique sound. However, there are some synonyms, especially, the numbers ended by digit 4 and 5; they could be read in different ways. In order to cover all cases, the corpus consists of all of the variants (synonyms) of theses digits and numbers.

A set of more than 50 application words is defined in the corpus. Each word corresponds to an action which is useful in several applications such as telephone services, measurement, human-machine interface …

After selecting and processing selected paragraphs and conversations, our sentences corpus is divided into two parts, a common part and a private part. The common part contains 33 conversations and 37 paragraphs. They were read by all speakers. The private part includes about 2,000 short paragraphs, each speaker was asked to read 40 paragraphs.

3.3. Distribution of Acoustic units

To evaluate our corpus, we used several modules to analyze the distributions of acoustic units including mono-words, base syllables, Initial-Final parts, phonemes and tones in the corpus and compare their distribution with the distribution obtained on a larger corpus which was collected in section 1 (Web corpus). We consider the acoustic units distribution obtained on this large web corpus, as the reference distribution of what can occur on Vietnamese language. Vietnamese is a monosyllabic and tonal language. Besides analyzing the distributions of phonemes (mono-phone, diphone and triphone) like other language (figure 1), we carried out an analysis of the distributions of tones (figure 2), initial and final parts.

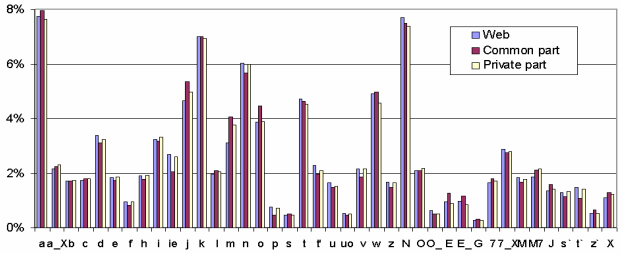


Figure 1: Distribution of mono-phones in common part, private part and Web corpora

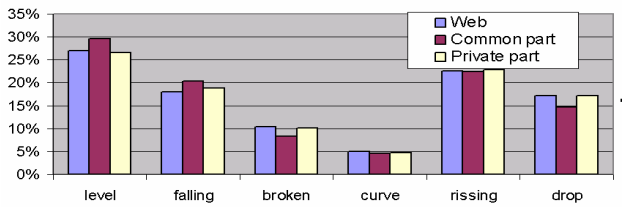


Figure 2: Distribution of six tones in common part, private part and Web corpora

In addition, we calculated the correlation coefficients between the distributions of the common part and the private part with the web reference corpora. Table 2 shows that the correlation coefficients are near to 1. So we can conclude that our corpus is acceptable and correctly balanced in terms of acoustic units and tones.

Acoustic Units	Correlation Coefficients	
	Private part	Common part
Mono-phone	0.9962	0.9885
Di-phone	0.9821	0.9458
Tri-phone	0.9811	0.9420
Tone	0.9984	0.9885
Initial Part	0.9904	0.9670
Final Part	0.9910	0.9706

Table 2: Correlation coefficients of acoustic units between common part, private part and Web data

3.4. Speaker selection and recording

Our speakers are the employees of the International Research Center MICA, teachers and students of Hanoi University of Technology and their friends. They are from four big cities and provinces, Hanoi, Nghe An, Ha Tinh, HCM city, which represent 3 major dialect regions: the South, the North, and the Middle. The age of the speakers ranges from 15 to 45 years old, among the 50 speakers, 25 are females and 25 are males.

For the acquisition, and managing of speech signals during recording, we use the EMACOP system, developed at CLIPS. EMACOP is a Multimedia Environment for Acquiring and Managing Speech Corpora, running under Windows 9x and Windows NT. EMACOP meets SAM specifications on input and output (Vaufreydaz, 2000).

In our project, recordings will take place with a SENNHEISER HMD 410-6 head microphone and a microphone pre-amplifier Soundcraft Spirit Folio FX8. The sampling frequency is 16 kHz.

At this time, 15 speakers have been recorded in the studio of the International Research Center MICA, Hanoi University of Technology, Vietnam. Each speaker has been asked for recording about 60 minutes, which includes 45 common minutes of phonemes, tones, digits and strings of digits, application words and common sentences and paragraphs corpus, and 15 private minutes of about 40 short paragraphs.

4. DESIGNING A VOCABULARY AND A PHONETIC DICTIONARY

4.1. Vocabulary

In order to build the language model for ASR, it is necessary to have a vocabulary. This list of words will be also useful to filter out the text documents before training a language model. We can use a bilingual or a multilingual dictionary for generating this vocabulary.

In fact, there are many methods to construct a bi-lingual or multi-lingual dictionary. In the context of the Papillon[1]

[1] http://bushido.imag.fr/papillon

project, the construction of a lexical base for a new language may take several different ways depending on where the author has to start. The Papillon project aims at creating a multilingual lexical database covering among others English, French, Japanese, Malay, Lao, Thai and Vietnamese.

From this Papillon project, we got a dictionary for Vietnamese language (French-Vietnamese and Vietnamese-French). Then, we filtered this dictionary to have a list of more than 40,000 unique words in Vietnamese: compound words, borrowed words and isolated words. By taking only the most frequent words, we can discount this size of vocabulary to 20,000 words. These were the highest frequency words which occur in the documents of our Web text corpus.

4.2. Phonetic dictionary

Phonetic dictionary or pronunciation dictionary is a key part for acoustic modeling in ASR. However, there is not any official pronunciation dictionary in Vietnam which satisfies our requirement. Therefore, we decide to construct a dictionary which is not used only for our works but also for other requirements in spoken language processing.

As referred above, Vietnamese language is a mono-syllabic and tonal language with 6 tones. A syllable in full structure (a tonal syllable or an isolated word) has five parts: *initial sound* (consonant), *medial sound* (semi-vowel), *nucleus sound* (vowel or diphthong), *final sound* (consonant or semi-vowel) and tone (see figure 3). Except the initial consonant (called INITIAL part), the rest of the syllable is called a FINAL part.

<table>
<tr><td colspan="4">Tonal syllable (6,492)</td></tr>
<tr><td colspan="3">Base syllable (2,376)</td><td rowspan="2">Tone (6)</td></tr>
<tr><td rowspan="2">INITIAL (22)</td><td colspan="2">FINAL (155)</td></tr>
<tr><td>Medial(1)</td><td>Nucleus(16)</td><td>Ending(8)</td></tr>
</table>

Figure 3: The phonological hierarchy of Vietnamese syllables with the total number of each phonetic unit

Since Vietnamese is a monosyllabic language (each syllable is one isolated word), we decide to extract a vocabulary of only about 6,500 isolated-words from 40,000 words vocabulary obtained in the previous section for building a pronunciation dictionary.

From this isolated-words vocabulary, we have firstly extracted all **22 initial parts**, **155 final parts** and **6 tones**. And then the phonological transcripts of these parts (IPA Symbols) were manually built. They were called *IPA Reference Table for Sub-word Units (**IPATU**)*.

Secondly, we have constructed an automatic syllable-based phonological analyzer (called ***VNPhoneAnalyzer***). This analyzer uses a phone concatenation algorithm described below.

Phone concatenation algorithm:

1. Separate a syllable into initial part, final part and tone.
2. Transcribe the initial part, final part and tone by their correspondent phonological representations looked up in the IPATU Table.
3. Concatenate all of these phonological transcripts into a complete phonological transcript for a syllable.

The *VNPhoneAnalyzer* could output many transcription forms: IPA-Unicode Symbolic, IPA number, SAMPA[1]…

Finally, a pronunciation dictionary for Vietnamese was also built by applying the *VNPhoneAnalyzer* for the isolated-word vocabulary and tested under the helping of Linguistic Institute of Vietnam.

5. CONCLUSIONS AND PERSPECTIVES

We have introduced the process of building the written and spoken resources for Vietnamese language at CLIPS-IMAG Laboratory (France) and International Research Center MICA (Vietnam). VNSpeechCorpus has been carried out in order to provide a great quantity of usable data for training and testing ASR systems. It could also be used for speech synthesis based on the acoustic units that are smaller than syllable in Vietnamese language. From the results, we can conclude that, with a suitable adaptation we can apply the methods which were developed for majority languages to minority languages. In our project, the Web-based methods built in CLIPS-IMAG which is useful for French and English is adapted for Vietnamese language. At present, the automatic alignment and recording process are being done for VNSpeechCorpus. In the future, VNSpeechCorpus will be available on CD-ROM. Furthermore, our methods will be used for other minority languages such as Lao, or Cambodian... for evaluating more precisely the efficacy of our methodology.

REFERENCES

Berment, V. (2002). Several technical issues for building new lexical bases. In Workshop Papillon, Tokyo, Japan, 2002.

Doan, T.T. (1997). Ngữ âm Tiếng Việt (Vietnamese Phonetics). Vietnam National University Publishing House, 1997.

IPA (1999). Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet. Cambridge University Press, 1999.

Le, V.B., Bigi, B., Besacier, L., Castelli, E. (2003). Using the Web for fast language model construction in minority languages. In Eurospeech 2003, pp. 3117-3120, Geneva, 1-4 Sept, 2003.

Schultz, T., Waibel, A. (2001). Language independent and language adaptive acoustic modeling for speech recognition. In Speech Communication, vol. 35, no. 1-2, pp. 31–51, 2001.

Vaufreydaz, D., Bergamini, C., Serignat, J.F., Besacier, L., Akbar, M. (2000). A new methodology for speech corpora definition from internet documents. In LREC, vol. I, pp. 423–426, Athens, Greece, 2000.

[1] http://www.phon.ucl.ac.uk/home/sampa

FIRST STEPS IN FAST ACOUSTIC MODELING FOR A NEW TARGET LANGUAGE: APPLICATION TO VIETNAMESE

Viet Bac Le, Laurent Besacier

CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, FRANCE
{viet-bac.le, laurent.besacier}@imag.fr

ABSTRACT

This paper presents our first steps in fast acoustic modeling for a new target language. Both knowledge-based and data-driven methods were used to obtain phone mapping tables between a source language (french) and a target language (vietnamese). While acoustic models borrowed directly from the source language did not perform very well, we have shown that using a small amount of adaptation data in the target language (one or two hours) lead to very acceptable ASR performance. Our best continuous vietnamese recognition system, adapted with only two hours of vietnamese data, obtains a word accuracy of 63.9% on one hour of vietnamese speech dialog for instance.

1. INTRODUCTION

There are more than 6000 languages in the world but only a small number possess the resources required for implementation of Human Language Technologies (HLT). Thus, HLT are mostly concerned by languages which have large resources available or which suddenly became of interest because of the economic or political scene. On the contrary, languages from developing countries or minorities were less treated in the past years. One way of ameliorating this "linguistic divide" is through starting research on portability of HLT for multilingual applications. This question has been increasingly discussed in the recent years, for instance in the SALTMIL[1] (Speech and Language Technology for Minority Languages) group. However, in SALTMIL, "minority language" mostly means "language spoken by a minority of people". We rather focus, in our work, on languages which have a "minority of resources usable in HLT". These languages are mostly from developing countries, but can be spoken by a large population. In this paper, we will notably deal with Vietnamese, which is spoken by about 70 millions of persons, but for which very few usable electronic resources are available.

Among HLT, we are interested in Automatic Speech Recognition (ASR). We are currently investigating new techniques and tools for a fast portability of speech recognition systems to new languages like Vietnamese, for which few signal and text resources are available. This activity includes different aspects:

- Portability of acoustic models: this can be achieved, for example by using tools for performing a fast collection of speech signals [1] or by using Language Adaptive Acoustic Modeling [2],
- Language modeling for new languages: we have already proposed to use web-based techniques which have shown ability to collect large amount of text corpora [3]. For languages in which no usable text corpora exist, this is moreover the only viable approach to collect text data,
- Dictionaries: collaborative approaches like in [4] could be also proposed for ASR.

This paper addresses particularly our first steps in fast acoustic modeling for a new target language. At first, we had no speech data at all in the target language (vietnamese). Then, after having collected some vietnamese speech data, no phonetic alignment was available and so an automatic labeling process was firstly employed to phonetically align these acoustic data. Acoustic models for this labeling process were borrowed from French (source language) in which a huge amount of acoustic data was already available.

Concerning cross-lingual acoustic modeling, two approaches are proposed in section 2. The first approach is based on a pronunciation modeling with the source language phoneme set and the second is a model mapping approach. The main difference between both approaches lies in the phoneme set used in acoustic modeling. While the first approach uses a french phoneme set, the second uses a vietnamese phoneme set. Since the first acoustic models trained without any target data did not lead to acceptable performance, an adaptation process is also presented to improve the recognition rate by using a small amount of speech data in target language. Experimental framework and results are presented in section 3 and 4 respectively. Finally, section 5 concludes with work and gives some perspectives.

2. CROSS-LINGUAL ACOUSTIC MODELING

2.1. Phone mapping table generation

Some methods of phone mapping can be used to evaluate acoustic similarities across languages. The core of these methods is the phone mapping table that describes the similarity of sounds between two different languages. Both *knowledge-based* and *data-driven* methods [5] are used in our work to manually or automatically obtain these phone mapping tables. Table 1 shows an example of some phone units in Vietnamese mapped from a french phone set by both knowledge-based and data-driven phone mapping methods.

[1] http://www.cstr.ed.ac.uk/~briony/SALTMIL

Vietnamese phoneme	French phoneme	
	Knowledge-based	Data-driven
t	t	t
G	g	g
X	k	R
NG	NG	n
w	w	o
u_o	u	o
...	...	...

Table 1: Sample of a Vietnamese/French phone mapping table

a) Manual Phone Mapping Table Generation (knowledge-based method)

The phone mapping table can be obtained by using acoustic-phonetic knowledge to categorize phonetic units with similar features of the individual languages. In a knowledge-based method, we find the IPA counterpart of target phonemes among phonemes in source language. This kind of method can be used if no data at all is available in the target language but a good knowledge of the target language is needed.

b) Automatic Phone Mapping Table Generation using Confusion Matrix (data-driven method)

By using small amounts of acoustic data in the target language, the phone mapping table can be automatically created with data-driven methods. In our work, a confusion matrix is calculated by applying a source language phoneme recognizer on a target language speech corpus already phonetized with the target language acoustic units.

Firstly, as in [6], a phoneme recognizer in the source language is applied on the development data set in target language to decode the phonetic representation of each utterance. A phonetic transcription in target language phone sets of these utterances must be already available. Then, to align phonetic hypotheses with their phonetic transcription references, we project them in a time scale (see figure 1).

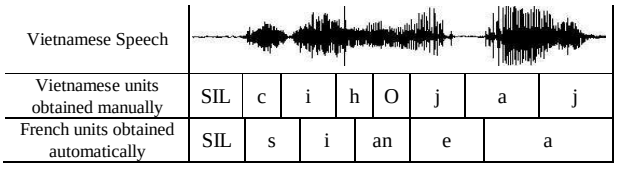


Figure 1: Time-based source/target phoneme scoring

The phonetic hypotheses and the references are thus compared frame by frame to count the co-occurrences between a phoneme in source language and a phoneme in target language. These counts make up a matrix where each entry gives the number of times a reference phoneme in the target language has been confused with a phoneme in source language. To obtain the final *confusion matrix*, each entry is normalized by dividing it through the total of occurrences of all corresponding phonemes in source language. Finally, the phone mappings are derived from this confusion matrix by selecting each phoneme in target language with the correspondence phoneme in source language which has the highest confusion value.

Furthermore, we can also calculate the Phoneme Correct Rate (PCR) of a phonetic recognizer based on this confusion matrix. Phonetic references and hypotheses in a same phone set are aligned to count the co-occurrences $C_{i,j}$ between phoneme i and phoneme j (at the frame level) in a set of n phonemes. Then the PCR is calculated as shown on the following equations.

Normalization of i[th] phoneme: $N_i = \sum_{j=1}^{n} C_{i,j}$

Phone Correct Rate: $PCR = \sum_{i=1}^{n} \frac{C_{i,i}}{N_i}$

In the experiments of section 4, we will report PCR to evaluate the quality of our acoustic models, but also Phone Accuracy (PA) rate which is also used in the literature and which corresponds to the count of the correct hypothesis phonemes found on a signal compared to an aligned reference (same scoring as the word accuracy using sclite tool for instance).

2.2. Pronunciation modeling with the source language phoneme set

The purpose of the pronunciation modeling approach is to describe the pronunciation of a word or a dictionary entry in the target language in terms of the symbols associated with the acoustic units of the source language(s). That means that target language words are described in term of source language units.

Cross-lingual pronunciation modeling techniques were previously proposed in [6, 7, 8] to phonetically transcribe a word in target language in terms of the symbols used in source languages acoustic models. To automatically associate each of the target language words with a sequence of phonemes in the source language(s), a phonetic recognition system in the source language(s) was applied on the words to be phonetized. A pronunciation model of each word in the training data was obtained via a N-best hypotheses list. One drawback of these approaches is that one have to apply a phoneme recognizer on each target language word to be phonetized.

In our work, an important difference to note is that we already have a pronunciation dictionnary for the target language where each word is phonetized in target language phone units, because a vietnamese phonetizer was already designed in a previous work [9]. Then, the pronunciation modeling process is based on the transformation of each entry of the target language pronunciation dictionary into dictionary entries described with the source language acoustic units. The process used to transform a pronunciation dictionary from a target language phoneme set to a source language phoneme set is achieved using the following steps:

1. Use data-driven phone mapping techniques to build the confusion matrix. Each phoneme in target language can be mapped into N-best phonemes in source language depending on their confusion values.
2. Transform the target language pronunciation dictionary by replacing its phone units by their corresponding units in source language. Thus, each target language dictionary entry may be transformed to one or more entries in source language (table 2).

Vietnamese word	Vietnamese pronunciation	French pronunciation
gánh(1)	g EX J	g a J
gánh(2)	-	g in J
gánh(3)	-	g in n
hiên(1)	h ie n	R e n
hiên(2)	-	R i e n

Table 2: Pronunciation modeling using phone mapping

After obtaining a pronunciation dictionary in terms of the symbols associated with the acoustic units of the source language, we can directly use acoustic models in source language to decode the speech of an utterance in the target language. Moreover, for this approach it is interesting to use several source languages instead of one to better cover the phone inventory of the target language. Thus, this approach can be extended and used with multilingual acoustic models [2].

2.3. Acoustic model mapping

In that case, the difference with section 2.2 approach is that the final phoneme set used is the target language one. A phone mapping table is first created by using both knowledge-based and data-driven phone mapping methods.

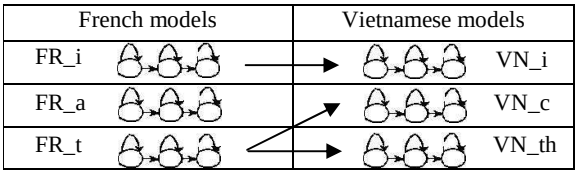


Figure 2: Vietnamese model mapped from French model

Cross-lingual acoustic models are then built by borrowing source language acoustic models (see figure 2). The obtained cross-lingual models can be used to decode the acoustic representation of each utterance by using a pronunciation dictionary in target language. This approach can be extended to use sub-phone units mapping [2, 5] where HMM states of a phoneme in target language can be borrowed from HMM states of some different phonemes in source language.

2.4. Model adaptation

While both the knowledge-based and data-driven phone mappings can be used without modification of the original source language acoustic models, HMM adaptation using MLLR or MAP [5] techniques can also be used to improve the systems using a small amount of target language adaptation data if available.

In our work, after the cross-lingual acoustic modeling step, limited data from the target language was phonetically aligned using the initial acoustic models. Then, these models were adapted with this limited data using MAP adaptation.

However, after adapting the cross-lingual models, a first evaluation showed high error rates for some phonemes. A reason can be that for these error prone phonemes, there was an important difference between the source data (used to build the initial model) and the target data (used for adaptation).

Thus, we decided to propose a *model collective adaptation* method to improve these phoneme error rates. The following algorithm is used:

1. By evaluating the individual phone error rates on a development set, we eliminate cross-lingual acoustic models which have a high error rate (using a threshold).

2. We reinitialize these eliminated models by running a k-mean clustering process on the training data of the target language. That means that only target language data is used to train these few models.

3. We perform some iterations of training on the data set of target language.

The performances of cross-lingual acoustic modeling and adaptation processes are shown in the section 4.

3. EXPERIMENTAL FRAMEWORK

3.1. Vietnamese pronunciation dictionary creation

Vietnamese language is a monosyllabic and tonal language with 6 tones (see figure 3). Except the initial consonant (called INITIAL part), the rest of the syllable is called a FINAL part.

<table>
<tr><td colspan="4">Tonal syllable (6,492)</td></tr>
<tr><td rowspan="2">INITIAL(22)</td><td colspan="3">Base syllable (2,376)</td><td rowspan="2">Tone (6)</td></tr>
<tr><td colspan="3">FINAL (155)</td></tr>
<tr><td></td><td>Medial(1)</td><td>Nucleus(16)</td><td>Ending(8)</td><td></td></tr>
</table>

Figure 3: The phonological hierarchy of Vietnamese syllables with the total number of each phonetic unit

A vocabulary of 6,492 isolated-words was firstly extracted from 40,000 full-words vocabulary. Then a pronunciation dictionary for Vietnamese was built by applying our *VNPhoneAnalyzer* [9] on this isolated-word vocabulary. The pronunciation dictionnary was finally verified with the help of the Linguistic Institute in Vietnam.

3.2. Text corpus and language modeling

A new methodology for fast text corpora acquisition for minority languages was already proposed and used in [3]. Documents were gathered from Internet by some web robots. Then, these web pages were filtered and analyzed for building a text corpus. The Vietnamese data collection is composed of more than 2.5 GB of web pages. After data preparation, the text corpus is made of 868 MB, i.e. 10,020,267 sentences.

By using the vocabulary of 6,492 isolated-words, a Vietnamese language model was constructed and estimated from this text corpus. The perplexity value evaluated on our future test corpus is 108.5.

3.3. Speech corpora

To calculate the confusion matrices (for data-driven approach), to train the acoustic models in baseline and adapted systems and to test the performance of ASR systems, a vietnamese speech corpus was needed. *VNSpeechCorpus* [9], which have been built in CLIPS-IMAG and MICA[2] Laboratories, was used. The speech is recorded in both quiet and office environment. The speakers are from 3 major dialect regions of Vietnam: the South, the North, and the Middle. Each speaker has been asked for recording about 55-60 minutes, which includes 25 minutes of phonemes, tones, digits, application words, 6-7 minutes of short dialogue and 25 minutes of about 40 common and private text-based paragraphs.

At this time, 15 speakers have been recorded in quiet environment. We use the text-based paragraphs subset as development and adaptation corpus (about 3 hours). The dialogue subset is used as test corpus (about 1 hour).

[2] www.mica.edu.vn

The BREF80[3] and BRAF100 speech corpora [1] were used to train the French ASR system. The BREF80 corpus, designed at LIMSI, contains about 10 hours of speech data of 80 speakers. The BRAF100 corpus, which was recorded in CLIPS-IMAG laboratory by 100 speakers, contains about 25 hours of speech data. Both Vietnamese and French speech data used a sampling frequency of 16 KHz and a sampling rate of 16 bits.

4. EXPERIMENTATIONS

4.1. ASR System

All recognition experimentations use the JANUS Speech Recognition Toolkit (JRTK) [10] developed by the ISL Laboratories. The model topology is 3 states left-to-right, 64 Gaussian mixtures. The pre-processing of the system consists of extracting a feature vector every 10 ms. The feature vector of 43 dimensions contains zero-crossing, 13 MFCC, energy and their first and second derivatives. A LDA transformation is used to reduce the feature vector size to 24. For the moment, we deal with context-independent acoustic models only for Vietnamese. In the experiments described in this paper, the phones are modeled independently of the tones. The decision between two different words corresponding to a same phone sequence but to different tones is made by the language model.

4.2. Experimental results

In order to test both the phoneme recognizer and the whole ASR system, we systematically report phoneme correct rate (PCR) and word accuracy (WA) obtained on our vietnamese test data (about 1 hour). The phoneme accuracy (PA) is also given in table 3 and 4 but it is obviously very correlated with the PCR.

Table 3 shows the performance obtained without any adaptation data. In model mapping approach, data-driven method produces slightly better results in comparison to knowledge-based (KB) method. However, the difference is very small and in any case, the performance is not really acceptable.

Approaches	%PA	%PCR	%W.A
DictMap-DataDriven	-	-	16.54
ModelMap-KB	**27.18**	24.51	16.13
ModelMap-DataDriven	26.04	**25.89**	**18.52**

Table 3: Phone recognition and ASR performances without adaptation data

Table 4 shows the performance obtained with adaptation data aligned with acoustic models obtained with the *model mapping* approach with either knowledge-based or data-driven method. The performance of our *model collective adaptation* process is also reported. Adaptation data comes from a part of the *VNSpeechCorpus* which contains about 3 hours of speech data. We divide this corpus into 3 adaptation subsets (each contains about 1 hours of data). It is important to note that in these 3 adaptation subsets, the speakers are the same as the ones in the test data. So, in addition to language adaptation, we have to admit that we also do speaker adaptation at the same time. This fact can explain the big difference in performance observed between "no adaptation data" and "one hour adaptation data". The adaptation process is then performed in 3 cycles over 3 subsets. It is interesting to note that with only one hour of signal we already reach acceptable ASR performance: 62.6% with the *model collective* adaptation which gives the best result. Quite surprisingly, the improvement of the results is then not very important when using 2 and 3 hours of adaptation data.

Models	Adapt 1h			Adapt 2h			Adapt 3h		
	PA	PCR	WA	PA	PCR	WA	PA	PCR	WA
K-Based	58.0	57.1	60.4	59.2	58.9	63.6	58.9	58.8	62.2
D-Driven	58.7	57.4	61.6	59.4	58.8	63.8	**59.8**	**59.4**	63.2
Collective	**59.3**	**58.9**	**62,6**	**59.9**	**58.9**	**63.9**	59.6	58.9	**63.4**

Table 4: Phone recognition and ASR performances with adaptation data

5. CONCLUSIONS AND PERSPECTIVES

This paper presented our first steps in fast acoustic modeling for a new target language (vietnamese). Both knowledge-based and data-driven methods were used to obtain phone mapping tables between a source and a target language. While acoustic models borrowed directly from the source language did not perform very well, we have shown that using a small amount of adaptation data in the target language (one or two hours) lead to very acceptable ASR performance. Our best vietnamese recognition system, adapted with only two hours of vietnamese data, obtains a word accuracy of 63.9% on a speech dialog test set for instance. These cross-lingual acoustic modeling and adaptation techniques are also currently applied on Mexican and Khmer languages.

6. REFERENCES

[1] D. Vaufreydaz, C. Bergamini, J.F. Serignat, L. Besacier, M. Akbar, "A New Methodology For Speech Corpora Definition From Internet Documents", *LREC 2000*, Athens, Juin 2000.

[2] T. Schultz, A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition", *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, 2001.

[3] V.B. Le, B. Bigi, L. Besacier, E. Castelli, "Using the Web for fast language model construction in minority languages", *Eurospeech 2003*, pp. 3117-3120, Geneva, September 2003.

[4] V. Berment, "Several technical issues for building new lexical bases", *PAPILLON Workshop*, Tokyo, 2002.

[5] P. Beyerlein et al., "Towards language independent acoustic modeling", *ASRU'99*, Keystone, Colorado, 1999.

[6] S. Stüker, "Automatic Generation of Pronunciation Dictionaries For New, Unseen Languages by Voting Among Phoneme Recognizers in Nine Different Languages", *Master thesis*, Carnegie Mellon University, April, 2002.

[7] T.Martin, T.Svendsen, S. Sridharan, "Cross-Lingual Pronunciation Modelling for Indonesian Speech Recognition", *Eurospeech 2003*, pp. 3125-3128, Geneva, September 2003.

[8] R. Bayeh, S.Lin, G.Chollet, C.Mokbel, "Towards multilingual speech recognition using data driven source/target acoustical units association", *ICASSP 2004*, vol. I, pp. 521-524, Montreal, Canada, May 2004.

[9] V.B. Le, D.D. Tran, E. Castelli, L. Besacier, J-F. Serignat , "Spoken and written language resources for Vietnamese", *LREC 2004*, Lisbon, May 2004.

[10] M.Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal, "The Karlsruhe-Verbmobil Speech Recognition Engine", *ICASSP'97*, Munich, 1997.

[3] http://www.elda.fr/catalogue/en/speech/S0006.html

ACOUSTIC-PHONETIC UNIT SIMILARITIES FOR CONTEXT DEPENDENT ACOUSTIC MODEL PORTABILITY

Viet Bac Le, Laurent Besacier*, Tanja Schultz***

* CLIPS-IMAG Laboratory, UMR CNRS 5524
BP 53, 38041 Grenoble Cedex 9, FRANCE

** Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, PA, USA

email: viet-bac.le@imag.fr

ABSTRACT

This paper addresses particularly the use of acoustic-phonetic unit similarities for portability of context dependent acoustic models to new languages. Since the IPA-based method is limited to a source/target phoneme mapping table construction, an estimation method of the similarity between two phonemes is proposed in this paper. Based on these phoneme similarities, some estimation methods for polyphone similarity and clustered polyphonic model similarity are investigated. For a new language, first a polyphonic decision tree is built with a small amount of speech data. Then, clustered models in the target language are duplicated from the nearest clustered models in the source language and adapted with limited data to the target language. Results obtained from the experiments demonstrate the feasibility of these methods.

1. INTRODUCTION

Nowadays, computers are heavily used to communicate via text and speech. Text processing tools, electronic dictionaries, and even more advanced systems like text-to-speech or dictation are readily available for several languages. However, the implementation of Human Language Technologies (HLT) requires significant resources, which have only been accumulated for a very small number of the 6900 languages in the world. Among HLT, we are particularly interested in Automatic Speech Recognition (ASR). Therefore, we are interested in new techniques and tools for rapid portability of speech recognition systems when only limited resources are available. Resource sparse languages are typically spoken in developing countries, but can nevertheless have many speakers. In this paper, we investigate Vietnamese, which is spoken by about 70 million people, but for which only very few usable electronic resources are available.

In crosslingual acoustic modeling, previous approaches have been limited to context independent models [1, 2, 3]. Monophonic acoustic models in target language were initialized using seed models from source language. Then, these initial models could be rebuilt or adapted using training data from the target language.

Since the recognition performance is increased significantly in wider contexts, the crosslingual context dependent acoustic modeling portability and adaptation can be investigated. J. Köhler [4] used HMM distances to calculate the similarity between two monophonic models. This method can be extended to context dependent models. A triphone similarity estimation method based on phoneme distances was first proposed by B. Imperl [5] and used an agglomerative clustering process to define a multilingual set of triphones. One problem in portability of context dependent acoustic models is that the context mismatch across languages increases dramatically for wider contexts. T. Schultz [6] proposed PDTS (Polyphone Decision Tree Specialization) to overcome this problem. In PDTS, the clustered multilingual polyphone decision tree is adapted to the target language by restarting the decision tree growing process according to the limited adaptation data in the target language. While PDTS is purely data-driven method, the intention of this paper is to explore a knowledge-based approach.

In this work, we investigate a new method for this crosslingual transfer process. We do not use the existing decision tree in source language but build a new decision tree just with a small amount of data from the target language. Then, based on the acoustic-phonetic unit similarities, some crosslingual transfer and adaptation processes are applied.

In this paper, we start in section 2 by proposing different acoustic-phonetic unit similarities estimation methods. In section 3 these similarities are applied to port context independent and dependent acoustic models across languages. The experimental framework and results are presented in section 4. Section 5 concludes the work and gives some future perspectives.

2. ACOUSTIC-PHONETIC UNIT SIMILARITIES

The research in crosslingual acoustic modeling is based on the assumption that the articulatory representations of phonemes are so similar across languages that phonemes can be considered as units which are independent from the underlying language [6]. Based on this assumption, we proposed in this section some methods for estimating the similarities of some phonetic units (phoneme, polyphone, clustered polyphone) which will be further used in crosslingual context dependent acoustic modeling.

2.1. Phoneme Similarity

In our work, both *data-driven* and *knowledge-based* methods are applied and proposed to automatically or manually obtain the phoneme similarities across languages.

2.1.1 Data-driven methods

The acoustic similarity between two phonemes can be obtained automatically by calculating the distance between two acoustic models (HMM distance [4], Kullback-Leibler distance, Bhattacharyya distance, Euclidean distance [7] or by calculating a confusion matrix [1, 2]). A confusion matrix is calculated by applying a source language phoneme recognizer to a small amount of target language acoustic data, which was already phonetized with the target language acoustic units. Note that in the basic phoneme recognizer we use, all phonemes have the same probability to appear. Then, each entry of the confusion matrix is

normalized by dividing it through the number of occurrences of all corresponding phonemes in the source language [3].

Normally, the confusion matrix represents the likelihood of the confusion between two phonemes. Thus, we can use these phoneme confusions to evaluate phoneme similarities. Let M, N be numbers of phonemes in source and target language. Let A(M,N) be the confusion matrix. The similarity $d(s_i, t_j)$ between phoneme t_j in the target language and phoneme s_i in the source language is calculated as:

$$d(s_i, t_j) = A_{i,j} \quad (1)$$

where $A_{i,j} \in [0,1]$, i=1..M, j=1..N.

2.1.2. Proposed knowledge-based method

Traditionally, knowledge-based methods had been applied to find the phoneme of the source language that best matches a phoneme in the target language [1, 6]. However, no knowledge-based method is known that allows to calculate the similarity between two phonemes. Thus, in this section, we propose a new knowledge-based method to calculate the phoneme similarity. As we know, similarities of sounds are documented in international phonetic inventories like the International Phonetic Alphabet (IPA)[1] which classifies sounds based on phonetic knowledge.

Based on the IPA phoneme classification we propose a *bottom-up algorithm* to determine a distance-based similarity between two phonemes. This algorithm consists of two steps: *top-down classification using a hierarchical graph* and *bottom-up phoneme distance estimation.*

a) Step 1: Top-down classification

Figure 1 shows a hierarchical graph where each node is classified into different layers. To each node we manually assigned a group of phonemes following the IPA phoneme classification scheme. Each group of phoneme has a user-defined similarity value assigned that represents the similarity of the elements within this group. All nodes corresponding to the same layer obtain the same similarity value. Let k be the number of layers and G_i be the user-defined similarity value for layer i (i = 0...k-1). In our work, we investigated several settings of k and G_i and set G = {0.9; 0.45; 0.25; 0.1; 0.0} with k = 5 based on a cross-evaluation in crosslingual acoustic modeling experiments.

To grow this graph, we start with the group PHONEME, which contains all the phonemes, at layer 0 and divide it into a CONSONANT group and VOWEL group at layer 1. This top-down classification is applied with increasingly specified grouping criteria until each group contains only one phoneme.

b) Step 2: Bottom-up estimation

To estimate the distance between two phoneme s and t, we locate them in the leaves of the graph and then trace back from their respective leaves until the nearest common parent node is reached. The similarity between s and t is thus given by the similarity value of layer i, which contains this parent node, we have:

$$d(s, t) = G_i \quad (2)$$

For example, the parent node of vowel [i] and [u] is CLOSE, we have:

$$d([i], [u]) = G_2 \quad (= 0.25 \text{ in our experiment }).$$

[1] http://www2.arts.gla.ac.uk/IPA/ipa.html

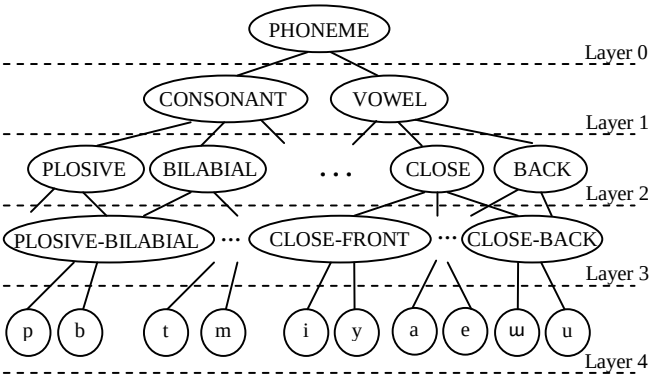


Figure 1 : Hierarchical graph for phoneme similarity

2.2. Polyphone Similarity

Let L be the left and the right context length of a polyphone. We assume that the context length of polyphones in source and target language are the same. If not, a context normalization procedure is needed. Let S be the phoneme set in source language, T be the phoneme set in target language.

Let $P_S = (s_{-L}, s_{-L+1},\ldots, s_{-1}, \mathbf{s_0}, s_1,\ldots, s_L)$ and $P_T = (t_{-L}, t_{-L+1},\ldots, t_{-1}, \mathbf{t_0}, t_1,\ldots, t_L)$ be polyphones in source and target language, where $s_{-L},\ldots, s_{-1}, s_0, s_1,\ldots, s_L \in S$ and $t_{-L},\ldots, t_{-1}, t_0, t_1,\ldots, t_L \in T$ denote the central phoneme, left phonemes or right phonemes of P_S and P_T.

The distance-based similarity of P_S and P_T is calculated as a weighted sum of distance between corresponding source/target phonemes along their context:

$$d(P_S, P_T) = \propto_0 .d(s_0, t_0) + \propto_1 .[d(s_{-1}, t_{-1}) + d(s_1, t_1)] + \ldots + \propto_L .[d(s_{-L}, t_{-L}) + d(s_L, t_L)] \quad (3)$$

where $\propto_0, \propto_1,\ldots \propto_L$ are contextual weight coefficients which represent the influence of contextual phoneme to the central phoneme; $d(s_k, t_k)$ is the phoneme distance (k = -L,…L). In the same way, the triphone similarities are calculated in [5].

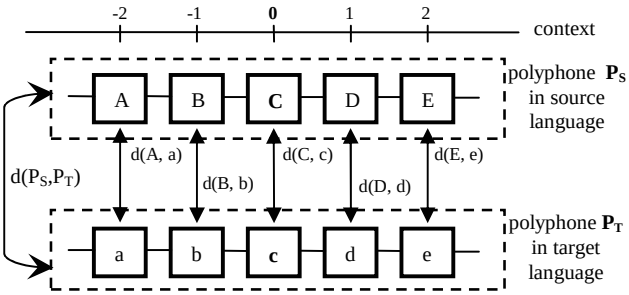


Figure 2 : Distance-based polyphone similarity

Figure 2 shows an example of the similarity between polyphone P_S = (A B **C** D E) and P_T = (a b **c** d e) in the source and target language.

For each polyphone of the target language, the nearest polyphone P_S* in source language is obtained that satisfies the following relation:

$$\forall P_S \in S,\ d(P_S^*, P_T) = \min[\ d(P_S, P_T)\ ] \quad (4)$$

2.3. Clustered Polyphonic Model Similarity

Since the number of polyphones in a language is very large (e.g.,

over 100,000 triphones for English), a limited training corpus usually does not cover enough occurrences of every polyphones. As a consequence many polyphones in the test set have never been seen in training. Thus, we need to find models that are accurate and trainable in acoustic modeling. A decision tree-based clustering (figure 3) or an agglomerative clustering [5] procedure is needed to cluster and model similar polyphones in a clustered polyphonic model.

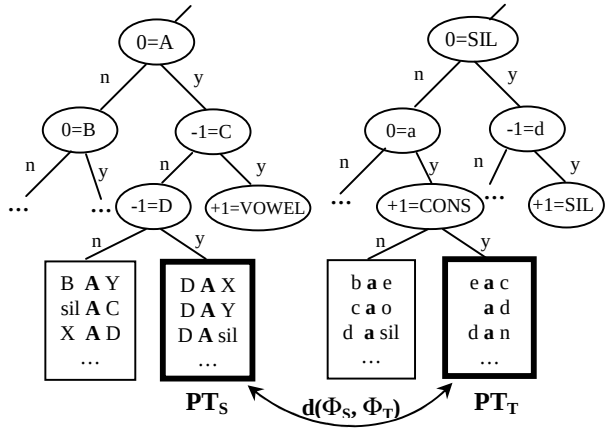


Figure 3 : Clustered polyphone similarity across languages

Therefore, for crosslingual context dependent modeling, a clustered polyphonic model similarity evaluation method must be proposed to find two nearest clustered polyphonic models across languages (figure 3).

Let $\Phi_S = (P_{S1}, P_{S2}\ldots P_{Sm})$ be a clustered polyphonic model of m polyphones in the source language and $\Phi_T = (P_{T1}, P_{T2},\ldots, P_{Tn})$ be a clustered polyphonic model of n polyphones in the target language, the similarity between Φ_S and Φ_T is the average of all distances between any two polyphones in Φ_S and Φ_T. We have:

$$d(\Phi_S, \Phi_T) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d(P_{Si}, P_{Tj})}{m.n} \quad (5)$$

For each clustered polyphone set in the target language, the nearest clustered polyphone set P_S^* in source language is obtained if it satisfies the following relation:

$$\forall\Phi_S,\ d(\Phi_S^*, \Phi_T) = \min\ [d(\Phi_S, \Phi_T)] \quad (6)$$

3. CROSSLINGUAL ACOUSTIC MODELING

3.1. Context Independent Acoustic Model Portability

For context independent acoustic modeling, the phonetic unit is the monophone and a distance between monophone models is calculated. Φ_S and Φ_T are calculated using the distance between two phonemes. Equation (5) leads to:

$$d(\Phi_S, \Phi_T)\ ] = d(P_S, P_T) = d(s, t) \quad (7)$$

where d(s, t) is calculated by equation (1) or (2).

Equation (6) leads to:

$$\forall\Phi_S,\ d(\Phi_S^*, \Phi_T) = \min\ [d(\Phi_S, \Phi_T)\ ] = \min[\ d(s, t)\ ] \quad (8)$$

By applying equation (8), a *phoneme mapping table* between source and language can be obtained. Based on this mapping table, the acoustic models in the target language can be borrowed from the source language and adapted by a small amount of target language speech data (see [3] for more details).

3.2. Context Dependent Acoustic Model Portability

In this section, a context dependent acoustic model portability method is proposed based on the phonetic similarities described in the previous section.

Firstly, by using a small amount of speech data in the target language, a decision tree for polyphone clustering (PT_T) can be built. We suppose that such a decision tree (PS_S) is also available in the source language (figure 3).

Secondly, by applying the equation (5), we can evaluate the distance between any two source/target clustered polyphonic models. That allows us, by applying the equation (6), to determine for each model in target language, the most similar model in the source language. This model is then copied into the acoustic model in the target language.

Finally, while acoustic models borrowed directly from the source language did not perform very well, an adaptation procedure (Viterbi training, MLLR, MAP) can successfully be applied with a small amount of speech data in the target language (see also [6]).

4. EXPERIMENTS AND RESULTS

This section presents our experiments in portability of context dependent acoustic models to new language using acoustic-phonetic unit similarities. Experiments and results in crosslingual context independent modeling were already presented in [3].

4.1. Experimental framework

4.1.1. ASR system

All recognition experiments use the JANUS toolkit [8] developed at the ISL Laboratories. The model topology is a 3- state left-to-right HMM with 48 Gaussian mixtures per state. The pre-processing of the system consists of extracting a 43 dimensional feature vector every 16 ms. The features consist of 13 MFCCs, energy, the first and second derivatives, and zero-crossing rate. An LDA transformation is used to reduce the feature vector dimensionality to 32.

Since Vietnamese language is a monosyllabic and tonal language with 6 tones (figure 4), we used syllables rather than words as recognition units (*syllable-based ASR system*). Furthermore, in the described experiments, the Vietnamese phones are modeled without tone indication. Since tone is a discriminative feature in Vietnamese, decisions between two different words with the same phone sequence but two different tones, are made by the language model.

4.1.2. Vietnamese Text and Speech Resources

<table>
<tr><td colspan="4">Tonal syllables (6,492)</td><td rowspan="3">Tone (6)</td></tr>
<tr><td colspan="3">Base syllables (2,376)</td></tr>
<tr><td rowspan="2">INITIAL(22)</td><td colspan="3">FINAL (155)</td></tr>
<tr><td>Medial(1)</td><td>Nucleus(16)</td><td>Ending(8)</td></tr>
</table>

Figure 4 : The phonological hierarchy of Vietnamese syllables

Firstly, since there are 6,492 syllables in the Vietnamese language (figure 4), a vocabulary of 6,492 syllables was extracted from a 40,000 word vocabulary. Then a pronunciation dictionary

for Vietnamese was built by applying our *VNPhoneAnalyzer* [9] on this syllable vocabulary.

Secondly, documents were gathered from Internet and filtered for building a text corpus. After data preparation, the text corpus has a size of 868 MB. A syllable-based statistical trigram language model was trained from this text corpus by using the SRILM toolkit [10] with a Good-Turing discounting and Katz backoff for smoothing. It is very important to note that with this toolkit, the unknown words are removed in our case, since we are in the framework of closed-vocabulary models. The perplexity value evaluated on our test corpus is *108.5*.

Finally, speech data was extracted from the *VNSpeechCorpus* [9], which was built at CLIPS-IMAG and MICA laboratories. In order to build a polyphonic decision tree and to adapt the crosslingual acoustic models, 2.25 hours of data spoken by 8 speakers were used. The test set contains 400 utterances spoken by 3 speakers different from the training speakers.

4.2.Experimental Results

4.2.1. Baseline System

By using 2.25 hours of Vietnamese speech data, decision trees for 500, 1000 and 2000 sub-triphone models were built respectively by a clustering procedure. These models are trained using LDA calculation, codebooks initialization (kmeans) and 6 iterations of Viterbi training.

4.2.2. Comparative Experiments

For crosslingual experiments, we use multilingual context dependent models (MM6-Mix with 12,000 sub-quinphone models) developed by ISL Laboratories [6]. Speech data from six languages were used to build these models: Arabic, Chinese, English, German, Japanese and Spanish. After the crosslingual transfer procedure, initial sub-models were adapted with 2.25 hours of Vietnamese speech data by 6 iterations of Viterbi training.

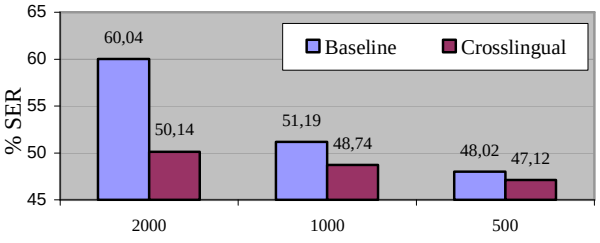


Figure 5 : Performance (syllable error rate) of baseline system and crosslingual method with different numbers of sub-triphone models

Figure 5 shows the Syllable Error Rate (SER) of the baseline system and the proposed crosslingual system. The crosslingual system improves 1.87%, 4.79% and 16.49% of absolute SER for 500, 1000, and 2000 sub-triphone models respectively. As the number of clustered sub-models increases, SER of the baseline system increases proportionally since the amount of data per model decreases due to the limited training data. However, the crosslingual system is able to overcome this problem by indirectly using data in other languages.

Figure 6 presents the influence of adaptation data size and number of speakers on the baseline system and two methods of phoneme similarity estimation: *proposed knowledge-based* and *data-driven using confusion matrix*. We find that the knowledge-based method outperforms the data-driven method.

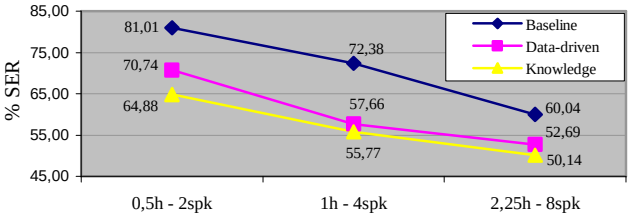


Figure 6 : Performance of phoneme similarity estimation methods with different amount of adaptation data (2000 sub-triphone models used): a) baseline system b) data-driven c) proposed knowledge-based

5. CONCLUSIONS AND PERSPECTIVES

This paper presents different methods of estimating the similarities between two acoustic-phonetic units. Based on these similarities, some crosslingual context independent and dependent acoustic modeling methods are proposed in our work. By using 2.25 hours of Vietnamese adaptation data, results from the obtained baseline system are outperformed by the proposed system (up to 16.49% of absolute SER). We note that, by using the vocabulary of 6,492 syllables, our syllable-based system almost covers all of the possible words in Vietnamese language (LVCSR). The potential of our method is demonstrated even though the use of trigrams the in syllable-based language modeling might be insufficient to obtain acceptable error rates (best SER is 47.12% obtained with 2.25h Vietnamese data only).

In the future, we will investigate word-based ASR systems to obtain the most likely recognition unit in Vietnamese language. We also plan to try different size of polyphone context and different contextual weight coefficients in order to obtain the suitable crosslingual acoustic models.

6. REFERENCES

[1] P. Beyerlein et al., “Towards language independent acoustic modeling”, *ASRU’99*, Keystone, CO, USA, December 1999.

[2] R. Bayeh et al., “Towards multilingual speech recognition using data driven source/target acoustical units association”, *ICASSP’04*, vol. I, pp. 521-524, Montreal, Canada, May 2004.

[3] V. B. Le, L. Besacier, “First steps in fast acoustic modeling for a new target language: application to Vietnamese”, *ICASSP’05*, vol. 1, pp. 821-824, Philadelphia, PA, USA, March 2005.

[4] J. Köhler, “Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds”, *ICSLP’96*, pp. 2195-2198, Philadelphia, PA, USA, October 1996.

[5] B. Imperl et al., “Agglomerative vs. Tree-based clustering for the definition of multilingual set of triphones”, *ICASSP’00*, vol. 3, pp. 1273-1276, Istanbul, Turkey, June 2000.

[6] T. Schultz, A. Waibel, “Language independent and language adaptive acoustic modeling for speech recognition”, *Speech Communication*, vol. 35, no. 1-2, pp. 31-51, August 2001.

[7] J. J. Sooful, E. C. Botha, “An acoustic distance measure for automatic cross-language phoneme mapping”, *PRASA’01*, pp. 99-102, South Africa, November 2001.

[8] M. Finke et al., “The Karlsruhe-Verbmobil Speech Recognition Engine”, *ICASSP’97*, vol. 1, pp. 83-86, Munich, Germany, 1997.

[9] V. B. Le et al, “Spoken and written language resources for Vietnamese”, *LREC’04*, pp. 509-602, Lisbon, Portugal, May 2004.

[10] A. Stolcke, “SRILM - An Extensible Language Modeling Toolkit”, *ICSLP’02*, vol. 2, pp. 901-904, Denver, CO, USA, September 2002.

Bibliographie

[Abdou 2004] S. Abdou et al., *The 2004 BBN Levantine Arabic and Mandarin CTS Transcription Systems*, RT-04 Workshop, Palisades, NY, USA, 2004.

[Abramson 1963] N. Abramson, *Information Theory and Coding*, McGraw-Hill, New York, 1963.

[Adda 1997] G. Adda, M. Adda-Decker, J-L. Gauvain, L. Lamel, *Text normalization and speech recognition in French*, EuroSpeech'97, vol. 5, pp. 2711-2714, Rhodes, Greece, September 1997.

[Allauzen 2004] A. Allauzen, J-L. Gauvain, *Construction automatique du vocabulaire d'un système de transcription*, JEP'04, Fès, Maroc, Avril 2004.

[Andersen 1994] O. Anderson, P. Dalsgaard, W. Barry, *On the use of data-driven clustering technique for identification of poly- and mono-phonemes for four European languages*, ICASSP'94, vol 1, pp. 121-124, Adelaide, Australia, April 2004.

[Antoine 2004] F. Antoine, D. Zhu, P. Boula de Mareüil, M. Adda-Decker, *Approches segmentales multilingues pour l'identification automatique de la langue : phones et syllabes*, JEP'04, Fez, Maroc, April 2004.

[Bahl 1977] L-R. Bahl, J-K. Baker, F. Jelinek, R-L. Mercer, *Perplexity - A Measure of the Difficulty of Speech Recognition Tasks*, Journal of the Acoustical Society of America, vol. 62, pp. S63, 1977.

[Baker 1975] J. K. Baker, *Stochastic modeling for automatic speech understanding*, Speech Recognition, Academic Press, pp. 521–542, 1975.

[Balakrishnama 1999] S. Balakrishnama, A. Ganapathiraju, J. Picone, *Linear discriminant analysis for signal processing problems*, Southeastcon'99, pp. 78-81, Lexington, KY, USA, Mars 1999.

[Berment 2002] V. Berment, *Several directions for minority languages computerization*, COLING'02, Taipei, Taiwan, 2002.

[Berment 2004] V. Berment, *Méthodes pour informatiser des langues et des groupes de langues peu dotées*, Thèse de doctorat de l'Université J. Fourier - Grenoble I, France, Mai 2004.

[Besacier 2006] L. Besacier, V-B. Le, C. Boitet, V. Berment, *ASR and translation for under-resourced languages*, 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06), Toulouse, France, 15-19 May 2006.

[Beyerlein 1998] P. Beyerlein, *Discriminative Model Combination*, ICASSP'98, vol. 1, pp. 481-484, Seattle, Washington, May 1998.

[Beyerlein 1999] P. Beyerlein, W. Byrne, J. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, W. Wang, *Towards language independent acoustic modeling*, ASRU'99, Keystone, CO, USA, December 1999.

[Boite 2000] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, H. Leich, *Traitement de la parole*, Presses Polytechniques et Universitaires Romandes, Collection Electricité, Lausanne, Switzerland, 2000.

[Boitet 2001] C. Boitet, *Méthodes d'acquisition lexicale en TAO : des dictionnaires spécialisés propriétaires aux bases lexicales généralistes et ouvertes*, TALN'01, vol. 2, pp. 249-265, Tours, France, juillet 2001.

[Boula-de-Mareüil 1997] P. Boula de Mareüil, *Étude linguistique appliquée à la synthèse de la parole à partir du texte*, Thèse de doctorat de l'Université Paris XI, Orsay, 1997.

[Calliope 1989] Calliope, *La parole et son traitement automatique*, Masson, Paris, France, 1989.

[Chang 2000] E. Chang et al., *Large Vocabulary Mandarin Speech Recognition With Different Approaches in Modeling Tones*, ICSLP'00, vol.2, pp. 983-986, Beijing, China, October 2000.

[Chen 2003] F. Chen, *Syllable Clustering and Spectral Discontinuity in Syllable-based TTS Systems*, ICASSP'03, vol. 1, pp. 688-691, Hong Kong, April 2003.

[Church 2003] K-W. Church, *Speech and Language Processing: Where Have Been and Where Are We Going?*, Eurospeech'03, Geneva, Switerland, September 2003.

[Cole 1997] R-A. Cole , J. Mariani, H. Uszkoreit, G-B. Varile A. Zaenen, A. Zampolli, *Survey of the State of the Art in Human Language Technology*, Cambridge University Press, 533 pages, ISBN 0-521-59277, 1997 (http://cslu.cse.ogi.edu/HLTsurvey/).

[Cover 1991] T. Cover, J. Thomas, *Elements of Information Theory*, Wiley and Sons, New York, 1991.

[Davis 1980] S-B. Davis, P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Transactions on ASSP, Vol. 28, No. 4, pp. 357-366, August 1980.

[Dieguez-Tirado 2005] J. Dieguez-Tirado, C. Garcia-Mateo, L. Docio-Fernandez, A. Cardenal-Lopez, *Adaptation strategies for the acoustic and language models in bilingual speech transcription*, ICASSP'05, Philadelphie, PA, USA, March 2005.

[Dinh 2001] D. Dinh, K. Hoang, V-T. Nguyen, *Vietnamese Word Segmentation*, NLPRS'01, pp. 749-756, Tokyo, Japan, November 2001.

[Dinh 2003] D. Dinh, P-H. Pham, Q-H. Ngo, *Some Lexical Issues in Building Electronic*

Vietnamese Dictionary, Atelier Papillon'03, Sapporo, Japan, Juillet 2003.

[Doan 1999] T-T. Doan, *Ngữ âm Tiếng Việt - phonétique vietnamienne*, Editions de l'Université Nationale du Vietnam, 363 pages, Hanoï, Vietnam, 1999.

[Doan-Nguyen 1998] H. Doan-Nguyen, *Techniques génériques d'accumulation d'ensembles lexicaux structurés à partir de ressources dictionnairiques informatisées multilingues hétérogènes*, Thèse de doctorat de l'Institut National Polytechnique de Grenoble, 168 pages, Grenoble, France, Décembre 1998.

[Dutoit 2002] T. Dutoit, L. Couvreur, F. Malfrère, V. Pagel, C. Ris, *Synthèse Vocale et Reconnaissance de la Parole : Droites Gauches et Mondes Parallèles*, CFA'02, Lille, France, Avril 2002.

[Finke 1997] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, M. Westphal, *The Karlsruhe-Verbmobil Speech Recognition Engine*, ICASSP'97, pp. 83-86, Munich, Germany 1997.

[Fügen, 2003] C. Fügen, S. Stüker, H. Soltau, F. Metze, T. Schultz, *Efficient Handling of Multilingual Language Models*, ASRU'03, Virgin Islands, USA, December 2003.

[Gao 2005] Y. Gao, L. Giu, H-K. Kuo, *Portability challenges in developing interactive dialogue systems*, ICASSP 2005, Philadelphia, USA, Mars 2005.

[Gauvain 1994] J-L. Gauvain, C-H. Lee, *Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains*, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 2, pp. 291-298, April 1994.

[Ghani 2005] R. Ghani, R. Jones, D. Mladenic, *Building Minority Language Corpora by Learning to Generate Web Search Queries*, Knowledge and Information Systems, vol. 7, issue 1, pp. 56-83, January 2005.

[Godfrey 1992] J. Godfrey, E. Holliman, J. McDaniel, *SWITCHBOARD: Telephone speech corpus for research and development*, ICASSP'92, vol. 1, pp. 517-520, San Francisco, CA, USA, March, 1992.

[Godfrey 1994] J. Godfrey, *Multilingual speech databases at LDC*, ARPA HLT'94 Workshop, pp 23-26, Plainsboro, NJ, USA, 1994.

[Gu 2002] L. Gu, S-A. Zahorian, *A New Robust Algorithm for Isolated Word Endpoint Detection*, ICASSP'02, vol. 4, pp. 4161, Orlando , FL, USA, May 2002.

[Haeb-Umbach 1992] R. Haeb-Umbach, H. Ney, *Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition*, ICASSP'92, vol I, pp 13-16, San Francisco, CA, USA, Mars 1992.

[Hamming 1986] R-W. Hamming, *Coding and Information Theory*, 2nd Edition, Englewood Cliffs NJ, Prentice-Hall, 1986.

[Haton 1991] J-P. Haton, J-M. Pierrel, G. Perennou, J.Calen, J-L.Gauvain, *Reconnaissance*

automatique de la parole, ISBN 2-04-018827-9, Bordas, Paris, 1991.

[Hieronymus 1993] J-L. Hieronymus, *ASCII Phonetic Symbols for the World's Languages: Worldbet*, Journal of the IPA, 1993.

[Hoang 2004] P. Hoang, *Từ điển vần – Vietnamese Rhyming Dictionary*, 211 pages, Vietnamese Lexicography Centre - Maison d'édition Danang, Vietnam, 2004.

[Hoge 1999] H. Höge, C. Draxler, H. van den Heuvel, F-T. Johansen, E. Sanders, H-S. Tropf, *SpeechDat multilingual speech databases for teleservices: Across the finish line*, EuroSpeech'99, vol. 6, pp. 2699-2702, Budapest, Hungary, September, 1999.

[Hovy 1999] E. Hovy, N. Ide, R. Frederking, J. Mariani, A. Zampolli, *Multilingual Information Management: Current Levels and Future Abilities*, Linguistica Computazionale, Volume XIV-XV, ISSN 0392-6907, Insituti Editoriali e Poligrafici Internazionli, Pisa, Italy, 2001 (version électronique : http://www.cs.cmu.edu/~ref/mlim/).

[Huang 2001] X. Huang, A. Acero, H-W. Hon, *Spoken Language Processing – A Guide to Theory, Algorithm, and System Development*, Practice Hall, 2001.

[Huffman 1970] F-E. Huffman, *Cambodian System of Writing and Beginning Reader*, Yale University Press, March 1970.

[Imperl 2000] B. Imperl, Z. Kacic, B. Horvat, A. Zgank, *Agglomerative vs. Tree-based clustering for the definition of multilingual set of triphones*, ICASSP'00, pp. 1273-1276, Istanbul, Turkey, 2000.

[Imperl 2003] B. Imperl, Z. Kacic, B. Horvat, A. Zgank, *Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones*, Speech Communication, vol. 39, issue 3-4, pp. 353-366, February 2003.

[IPA 1999] IPA, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, ISBN 0-521-63751-1, Cambridge University Press, 1999.

[Jelinek 1976] F. Jelinek, *Continuous Speech Recognition by Statistical Methods*, IEEE Trans. on ASSP, vol 64(4), pp. 532-556, Avril 1976.

[Jones 1997] R-J. Jones, S. Downey, J-S. Mason, *Continuous speech recognition using syllables*, EuroSpeech'97, vol. 3, pp. 1171-1174, Rhodes, Greece, September 1997.

[Juang 1985] B. Juang, L. Rabiner, *A probabilistic distance measure for hidden Markov models*, AT&T Technical Journal, vol. 64, no 2, pp. 391-408, February 1985.

[Kahn 1976] D. Kahn, *Syllable-based Generalizations in English Phonology*, PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1976.

[Kanthak 2002] S. Kanthak, H. Ney, *Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition*, ICASSP'02, pp. 845-848, Orlando FL, USA, 2002.

[Kanthak 2003] S. Kanthak, H. Ney, *Multilingual Acoustic Modeling Using Graphemes*, Eurospeech'03, pp. 1145-1148, Geneva, Switzerland, September 2003.

[Katz 1987] S-M. Katz, *Estimation of probabilities from sparse data for the language model component of a speech recognizer*, IEEE Trans. on ASSP, vol. 35, pp. 400-401, Mars 1987.

[Killer 2003a] M. Killer, *Grapheme Based Speech Recognition*, Master thesis, Interactive Systems Labs, Carnegie Mellon University, Pittsburgh, PA, USA, March, 2003.

[Killer 2003b] M. Killer, S. Stüker, T. Schultz, *Grapheme Based Speech Recognition*, Eurospeech'03, pp. 3141-3144, Geneva, Switzerland, September 2003.

[Kohler 1996] J. Köhler, *Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds*, ICSLP'96, pp. 2195-2198, Philadelphia, PA, USA, 1996.

[Kosawat 2003] K. Kosawat, *Méthodes de segmentation et d'analyse automatique de textes thaï*, Thèse de doctorat de l'Université de Marne la Vallée, Marne la Vallée, France, 2003.

[Kuhn 1990] R. Kuhn, R. De Mori, *A Cache-Based Natural Langage Model for Speech Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-12(6), pp. 570-583, 1990

[Kunzmann 2001] J. Kunzmann, K. Choukri, E. Jahnke, A. Kiessling, K. Knill, L. Lamel, T. Schultz, S. Yamamoto, *Portability of Automatic Speech Recognition Technology to New Languages: Multilinguality Issues and Speech/Text Resources***,** Panel Session, ASRU'01, Madonna di Campiglio, Italy, Decembre 2001.

[Le 2003a] V-B. Le, *Génération automatique des énoncés pour l'enregistrement d'un grand corpus vocal du vietnamien*, Rapport de stage, 14 pages, Centre MICA, Hanoi, Vietnam, Août 2003.

[Le 2003b] V-B. Le, B. Bigi, L. Besacier, E. Castelli, *Using the Web for fast language model construction in minority languages*, Eurospeech'03, pp. 3117-3120, Geneva, Switzerland, September 2003.

[Le 2004] V-B. Le, D-D. Tran, E. Castelli, L. Besacier, J-F. Serignat, *Spoken and written language resources for Vietnamese*, LREC'04, pp. 599-602, Lisbon, Portugal, May 2004.

[Le 2005] V-B. Le, L. Besacier, *First steps in fast acoustic modeling for a new target language: application to Vietnamese*, ICASSP'05, Vol. 1, pp. 821-824, Philadelphia, USA, March 2005.

[Le 2006] V-B. Le, L. Besacier, T. Schultz, *Acoustic-Phonetic Unit Similarities for Context Dependent Acoustic Model Portability*, ICASSP'06, Toulouse, France, May 2006.

[Lee 2002] T. Lee et al., *Using tone information in Cantonese continuous speech recognition*, ACM TALIP'02, vol. 1, issue 1, pp. 83-102, ACM Press, March 2002.

[Leggetter 1995] C-J. Leggetter, P-C. Woodland, *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models*, Computer Speech and

Language, vol. 9, no. 2, pp. 171-185, April 1995.

[Lejeune 2005] R. Lejeune, J. Baude, C. Tchong, H. Crepy, C. Waast-Richard, *Flavoured acoustic model and combined spelling to sound for asymmetrical bilingual environment*, Interspeech'05. Lisbonne, Portugal, September 2005.

[Levinson 1983] S. Levinson, L. Rabiner, M. Sondhi, *An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition*, AT&T Technical Journal, vol. 62, no. 2, pp. 1035-1074, April, 1983.

[Martin 2003] T. Martin, T. Svendsen, S. Sridharan, *Cross-Lingual Pronunciation Modelling for Indonesian Speech Recognition*, Eurospeech'03, pp. 3125-3128, Geneva, September 2003.

[Martin 2005] T. Martin, S. Sridharan, *Cross-language acoustic model refinement for the Indonesian language*, ICASSP'05, vol. I, pp. 865-868, Philadelphia, USA, Mars 2005.

[Meknavin 1997] S. Meknavin, P. Charoenpornsawat, B. Kijsirikul, *Feature-based Thai Word Segmentation*, NLPRS'97, Phuket, Thailand, 1997.

[Monroe 2002] G-A. Monroe, J-C. French, A-L. Powell, *Obtaining language models of web collections using query-based sampling techniques*, HICSS'02, vol. 03, no. 3, p. 67b, Big Island, Hawaï, USA, January 2002.

[Muthusamy 1992]Y-K. Muthusamy, R-A. Cole, B-T. Oshika, *The OGI Multi-language Telephone Speech Corpus*, ICSLP'92, pp. 895-898, Banff, Alberta, Canada, October 1992.

[Nguyen 2002] Q-C. Nguyen, *Reconnaissance de la parole en langue vietnamienne*, Thèse de doctorat de l'Institut National Polytechnique de Grenoble, 167 pages, Grenoble, Juin 2002.

[Nguyen 2004] T-B Nguyen, T-M-H Nguyen, L. Romary, X-L Vu, *Developing Tools and Building Linguistic Resources for Vietnamese Morpho-Syntactic Processing*, LREC'04, pp. 1231-1234, Lisbon, Portugal, May 2004.

[Nguyen-Duc 2006] H-H. Nguyen-Duc, H-Q. Vu, *Selection of phonetic units for Vietnamese large vocabulary continuous speech recognition*, RIVF'06, Ho Chi Minh Ville, Vietnam, Février, 2006.

[Nguyen-Thi 2000] B-M. Nguyen-Thi, *Regards sur l'enseignement de la phonétique dans la formation des étudiants en F.L.E à l'Université Pédagogique de Ho Chi Minh ville*, Thèse de doctorat de l'Université de Rouen, 706 pages, Rouen, France, Avril 2000.

[Nguyen-Thi 2001] N-H. Nguyen-Thi, *Interférences phonologiques de la langue maternelle dans l'apprentissage du français*, Thèse de doctorat de l'Université de Rouen, Rouen, France, October 2001.

[Nie 2000] J-Y. Nie, J. Gao, J. Zheng, M. Zhou, *On the Use of Words and N-grams for Chinese Information Retrieval*, IRAL'00, Hong Kong, September 2000.

[Nishimura 2001] R. Nishimura, K. Komatsu, Y. Kuroda, K. Nagatomo, A. Lee, H. Saruwatari,

K. Shikano, *Automatic N-gram Language Model Creation from Web Resources*, Eurospeech'01, pp.2127-2130, Aalborg, Danemark, Septembre 2001.

[Pham 1969] H-L. Pham, *Structure économique de la phonologie vietnamienne*, Thèse de doctorat de l'Université de Paris, 333 pages, Paris, France, 1969.

[Promchan 1998] P. Promchan, T. Yunyong, *Performance Comparison of Thai Word Separation Algorithms*, NCSEC'98, Thailand, 1998

[Rabiner 1978] L-R. Rabiner, R-W. Schafer , *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[Rabiner 1993] L-R. Rabiner, B-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[Rao 1992] G-V. Ramana Rao, *Detection of word boundaries in continuous speech using pitch and duration*, SST'92, Brisbane, Australia 1992.

[Rao 1996] G-V. Ramana Rao, J. Srichand, *Word Boundary Detection Using Pitch Variations*, ICSLP'96, pp. 813-816, Philadelphia, USA, October 1996.

[Ren 2005] X. Ren, X. He, Y. Zhang, *Mandarin / English Mixed-lingual Name Recognition for Mobile Phone*, Interspeech'05. Lisbon, Portugal, September 2005.

[Rosenfeld 1995] R. Rosenfeld, *Optimizing lexical and N-gram coverage via judicious use of linguistic data*, Eurospeech'95, vol. 3, pp. 1763-1766, Madrid, Spain, September, 1995.

[Sam 2004] S. Sam, *Traitement Automatique De La Langue Khmère*, Mémoire de fin d'études d'ingénieur, Institut de Technologie du Cambodge, Phnom Penh, Cambodge, Juin 2004.

[Scannell 2003] K-P. Scannell, *Automatic thesaurus generation for minority languages: an Irish example*, Atelier TALN'03, vol. 2, pp 203-212, Batz-sur-Mer, France, 11-14 Juin 2003.

[Schultz 1997] T. Schultz, A. Waibel,, *Fast bootstrapping of LVCSR systems with multilingual phoneme sets*, Eurospeech'97, pp. 371–374, Rhodes, Greece, 1997.

[Schultz 1999] T. Schultz, A. Waibel, *Experiments towards a multi-language LVCSR interface*, ICMI'99, Hong Kong, China, January 1999.

[Schultz 2001] T. Schultz, A. Waibel, *Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition*, Speech Communication, vol. 35, issue 1-2, pp 31-51, August 2001.

[Schultz 2002] T. Schultz, *GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University*, ICSLP'02 , Denver, CO, USA, Septembre 2002.

[Shannon 1951] C-E. Shannon, *Prediction and entropy of printed English*, Bell System Technical Journal, vol. 30, pp. 50-64, January 1951.

[SIL 2005] SIL International, *Ethnologue: Languages of the World, 15th Edition*, ISBN 1-

55671-159-X, 1272 pages**,** SIL International, Dallas, 2005.

[Silverman 1990] H-F. Silverman, D-P. Morgan, *The application of dynamic programming to connected speech recognition*, IEEE ASSP magazine, vol.7, pp.6-25, 1990.

[Singh 2002] R. Singh, B. Raj, and R-M. Stern, *Automatic Generation of Subword Units for Speech Recognition Systems*, IEEE Transactions on Speech and Audio Processing, vol. 10, pp. 98-99, 2002.

[Sloboda 1995] T. Sloboda, *Dictionary learning: Performance through consistency*, ICASSP'95, vol. 1, pp. 453-456, Detroit, MI, USA, 1995.

[Sloboda 1996] T. Sloboda, A. Waibel, *Dictionary learning for spontaneous speech recognition*, ICSLP'96, Philadelphia, PA, USA, 1996.

[Sooful 2001] J-J. Sooful, E-C Botha, *An acoustic distance measure for automatic cross-language phoneme mapping*, PRASA'01, pp. 99-102, Afrique du Sud, 2001.

[Special 2004] Special Session, *Multilinguality in Speech Processing*, ICASSP'04, Montréal, Canada, Mai 2004.

[Stolcke 2002]A. Stolcke, *SRILM - An Extensible Language Modeling Toolkit*, ICSLP'02, vol. 2, pp. 901-904, Denver, Colorado, September 2002.

[Stuker 2002] S. Stüker, *Automatic Generation of Pronunciation Dictionaries For New, Unseen Languages by Voting Among Phoneme Recognizers in Nine Different Languages*, Master thesis, Carnegie Mellon University, Pittsburgh, PA, USA, April, 2002.

[Stuker 2004] S. Stüker, T. Schultz, *A Grapheme based Speech Recognition System for Russian***,** SPECOM'04, St. Petersburgh, Russia, September 2004.

[Suebvisai 2005] S. Suebvisai, P. Charoenpornsawat, A. Black, M. Woszczyna, T. Schultz, *Thai Automatic Speech Recognition*, ICASSP'05, vol. I, pp. 857-860, Philadelphia, USA, Mars 2005.

[Talk 2005] Projet TALK, *Rapport Scientifique 1ère Tranche*, ITC-MICA-CLIPS, Mai 2005.

[Tomlison 1991] M-J. Tomlison, *Guide to Database Generation - Recording Protocol, Final Version*, SAM-RSRE-015, Marlvern, England, 1991.

[Tran 2003] D-D. Tran, *Building a large Vietnamese speech database*, Rapport de Master TIC, 87 pages, Institut Polytichnique de Hanoi, Vietnam, 2003.

[Tran 2005] D-D. Tran, E. Castelli, J-F. Serignat, V-L. Trinh, X-H. Le, *Influence of F0 on Vietnamese Syllable Perception*, Eurospeech'05, pp. 1697-1700, Lisbon, Portugal, September 2005.

[Truong 1970] V-C. Truong, *Structure* de *la langue vietnamienne*, Imprimerie nationale, 478 pages, Libraire orientaliste Paul Geuthner, Paris, 1970.

[UIT 2001] UIT, *Mise à jour des indicateurs des Télécommunications de l'UIT*,

Telecommunication Data and Statistics, Union Internationale des Télécommunications, Genève, Suisse, 2001.

[UIT-ICS 2001] UIT-ICS, *Internet Case Studies*, Workshop on the Internet in South East Asia, Thailand, November 2001.

[Vaufreydaz 1998] D. Vaufreydaz, M. Akbar, J. Caelen, J-F. Serignat, *EMACOP Environnement Multimédia pour l'Acquisition et la gestion de COrpus Parole*, JEP'98, pp. 175-178, Martigny, Switzerland, June 1998.

[Vaufreydaz 2000] D. Vaufreydaz, C. Bergamini, J-F. Serignat, L. Besacier, M. Akbar, *A new methodology for speech corpora definition from internet documents*, LREC'00, vol. I, pp. 423-426, Athens, Greece, 2000.

[Vaufreydaz 2001] D. Vaufreydaz, M. Géry, *Internet Evolution and Progress in Full Automatic French Language Modelling*, ASRU'01, Madonna di Campiglio, Italie, December 2001.

[Vaufreydaz 2002] D. Vaufreydaz, *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*, Thèse de doctorat de l'Université J. Fourier - Grenoble I, France, 226 pages, Janvier 2002.

[Waheed 2002] K. Waheed, K. Weaver, F-M. Salam, *A Robust Algorithm for Detecting Speech Segments Using an Entropic Contrast*, MWSCAS'02, vol 3, pp. 328-331, East Lansing, MI, USA, August 2002.

[Waibel 2004] A. Waibel, T. Schultz, S. Vogel, C. Fügen, M. Honal, M. Kolss, J. Reichert, S. Stüker, *Towards Language Portability in Statistical Machine Translation*, Special Session on Multilinguality in Speech Processing, ICASSP'04, Montreal, Canada, May 2004.

[Wang 2003] Z. Wang, T. Schultz, *Non-Native Spontaneous Speech Recognition through Polyphone Decision Tress Specialization*, Eurospeech'03, pp. 1449-1452, Geneva, Switzerland, September 2003.

[Wheatley 1994] B. Wheatley, K. Kondo, W. Anderson, Y. Muthusamy, *An evaluation of cross-language adaptation for rapid HMM development in a new language*, ICASSP'94, pp. 237-240, Adelaide, Australia, 1994.

[Wilpon 1987] J-G. Wilpon, L-R. Rabiner, *Application of hidden Markov models to automatic speech endpoint detection*, Computer Speech & Language, Academic Press Limited, no. 3/4, vol. 2, pp. 321-341, September-December 1987.

[Wu 2003] A. Wu, *Chinese word segmentation in MSR-NLP*, SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, July 2003.

[Yang 1998]Y-J. Yang, L-S. Lee, *A Syllable-Based Chinese Spoken Dialogue System for Telephone Directory Services Primarily Trained with A Corpus*, ICSLP'98, vol. 4, pp. 1247-1250, Sydney, Australia, November-December 1998.

[Young 1994] S-J. Young, J-J. Odell, P-C. Woodland, *Tree-based state tying for high accuracy*

acoustic modelling, ARPA Workshop on Human Language Technology, pp. 307-312, 1994.

[Zhu 2001] X. Zhu, R. Rosenfeld, *Improving Trigram Language Modelling with the World Wide Web*, ICASSP'01, pp. 533-536, Salt Lake City, USA, Mai 2001.