Reconnaissance automatique de la parole en langue somalienne

Abdillahi Nimaan [1,2], *Pascal Nocera*[1] *et Jean-François Bonastre*[1]

[1] Laboratoire Informatique d'Avignon (Université d'Avignon et des pays du Vaucluse)
BP 1228 84911 Avignon Cedex 9, France
[2] Institut des Sciences et des Nouvelles Technologies (Centre d'Études et des Recherches de Djibouti)
BP. 486 Djibouti, Djibouti
{nimaan.abdillahi, pascal.nocera, jean-françois.bonastre}@univ-avignon.fr
http ://www.lia.univ-avignon.fr

ABSTRACT

Most African countries follow an oral tradition system to transmit their cultural, scientific and historic heritage through generations. This ancestral knowledge accumulated during centuries is today threatened of disappearing. Automatic transciption and indexing tools seem potential solution to preserve it. This paper presents the first results of automatic speech recognition (ASR) of Djibouti languages in order to index the Djibouti cultural heritage. This work is dedicated to process Somali language, which represents half of the targeted Djiboutian audio archives. We describe the principal characteristics of audio (10 hours) and textual (3M words) training corpora collected and the first ASR results of this language. Using the the specificities of the Somali language, (words are composed of a concatenation of sub-words called "roots" in this paper), we improve the obtained results. We will also discuss future ways of research like roots indexing of audio archives.

1. INTRODUCTION

Le patrimoine scientifique, culturel, historique, *etc.* des pays africains se transmet oralement de générations en générations. Ce savoir ancestral, accumulé durant des siècles est menacé de disparition, du fait du processus de mondialisation, de la transformation sociale ainsi que du manque de moyen de sauvegarde. De nombreuses organisations nationales et internationales [15] oeuvrent pour endiguer ce phénomène. Aujourd'hui, la plupart des pays concernés disposent d'importantes bases de données audio, archivées, le plus souvent, par les stations radio locales depuis plusieurs décennies. Ces pays sont confrontés à deux questions : sauvegarder ce patrimoine par un programme de numérisation et le rendre plus accessible. Concernant le premier point, les techniques sont bien connues, et la numérisation, en cours dans de nombreux pays, n'est qu'un problème d'ordre logistique. Le second point est plus délicat, car l'exploitation de bases de données audio de grandes tailles [1] nécessite des traitements informatiques de haut niveau pour toutes les langues des pays concernés, tels que des outils de transcription et d'indexation automatiques. Ce papier présente les prémices du traitement automatique du patrimoine culturel audio de la république de Djibouti. Dans un premier temps, les langues djiboutiennes et les différents corpus constitués pour cette étude sont présentés. Nous décrivons ensuite les expériences de reconnaissance de la parole somalienne effectuées sur les mots et sur les racines. Finalement, nous tirons les conclusions de ces travaux, et énonçons les futurs axes de recherche.

2. LANGUES DJIBOUTIENNES

Quatre langues sont parlées à Djibouti, le français et l'arabe sont officiels, l'afar et le somalien sont autochtones et largement utilisés. Nos travaux actuels portent uniquement sur la langue somalienne qui concerne la moitié des archives audio ciblées. Cette langue est parlée dans plusieurs autres pays de l'Afrique de l'est par une population estimée entre 12 et 15 millions [2]. Elle est répertoriée dans la sous famille couchitique des langues afro-asiatique dans la classification internationale SIL [3]. La variante somali-somali, communément appelée langue somalienne, et parlée à Djibouti, est plus précisement visée dans nos recherches. Son système phonétique est composé de 22 consonnes et de 10 voyelles (5 longues et 5 courtes) [14]. La table 1 présente la structure phonétique des consonnes. C'est également une langue tonale avec deux ou trois tons différents [6], [7], [13]. Sa forme graphique est relativement jeune, puisqu'elle n'est écrite que depuis 1972 en caractères latins. Il n'existe donc aucun document écrit antérieur à cette date. La transcription d'un mot est directement issue de sa réalisation phonétique (chaque phonème est représenté par une lettre).

3. CONSTITUTION DES CORPUS

3.1. *Corpus textuel*

La reconnaissance automatique de la parole (RAP) basée sur des méthodes stochastiques atteint d'excellents niveaux de performances pour de nombreuses langues si des corpus d'entraînements (textuels et audio) de tailles suffisantes sont disponibles [8]. Le principal obstacle au développement de systèmes de RAP pour les langues africaines est le manque ou l'insuffisance de corpus textuels, du fait précisément de la tradition orale de ces pays et de leur récent système graphique. Depuis l'émergence de l'Internet et du Web, et surtout grâce aux journaux électroniques, des bases de données se constituent progressivement. Des travaux ont récemment été effectués par différentes équipes de chercheurs pour la construction automatique de corpus textuels à partir d'Internet pour les langues peu dotées [5], [16]. Inspiré par ces travaux - nous avons obtenu pour la langue somalienne - un texte brut de 3 millions de mots issus d'articles de journaux. La table 2 montre la composition du corpus textuel. Il est composé de 2 820k mots,

[1] Certains pays comme Djibouti, disposent d'archives culturelles audio enregistrées depuis 40 ans.

[2] http ://www.ethnologue.com

[3] http ://www.sil.org

	Labiales	Labiodentales	Dentales	Alvéolaires	Retroflèxes	Palatales	Vélaires	Uvulaires	Pharyngales	Glottales
Occlusives voisées	b		d		dh		g	q		’
Occlusives non voisées		t				k				
Nasales	m			n						
Fricatives non voisées		f		s		sh		kh	x	h
Fricatives voisées						j			c	
Roulées				r						
Latérales				l						
Approximantes	w					y				

TAB. 1: Structure phonétique des consonnes de la langue somalienne.

avec 121k mots différents.

TAB. 2: Composition du corpus textuel issu de l’Internet.

Phrases	84,7k
Mots	2 820k
Mots distincts	121k
Racines	6 042k
Racines distinctes	4,4k
Phonèmes	14 104k
Phonèmes distincts	36

3.2. Corpus audio : Asaas

Un sous-ensemble du corpus textuel a été isolé pour servir de base aux enregistrements sonores. Ce texte a été lu par 10 locuteurs âgés de 20 à 60 ans. Les enregistrements se sont déroulés dans un environnement non bruité. Nous avons ainsi constitué un corpus de 10 heures de parole à une fréquence d’échantillonnage de 16Khz et un codage sur 16 bits ainsi que sa transcription exacte au format Transcriber [1]. Ce premier corpus audio somalien, nommé Asaas [4], contient 59k mots et 10k mots différents. Les répartitions phonétiques de Asaas et du corpus textuel sont similaires (figure 1). Ce corpus est divisé en deux parties : 9,5 heures pour l’apprentissage et 0,5 heure pour les tests.

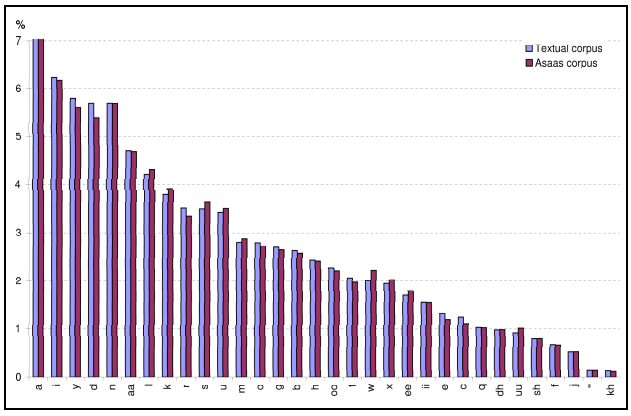

FIG. 1: Répartitions phonétiques pour le corpus textuel et le corpus audio Asaas.

[4] Asaas signifie fondation en langue somalienne

4. BOÎTES À OUTILS SOMALI

Afin de rendre les textes somaliens exploitables, nous avons été amenés à développer une série d’outils informatiques [10]. La langue somalienne est une langue “jeune” dans sa version écrite et présente une orthographe non standardisée. Le même mot peut se trouver écrit de différentes manières. Ces transcriptions multiples ne peuvent pas être considérées comme fausses, puisse qu’aucune standardisation ne s’est imposée à ce jour. Cependant, elles perturbent la qualité des modèles stochastiques et la robustesse des systèmes automatiques. Afin de corriger ce problème, nous avons écrit un programme qui “standardise” les transcriptions en utilisant l’orthographe la plus fréquemment rencontrée comme orthographe de référence.
La plupart des mots somaliens sont formés par la concaténation d’un nombre limité de "sous-mots", nommés “racines” dans ce papier. Leur forme est en général [3] : CVC, CVVC, CVV, VC [5], *etc.* Par exemple :
. *birlab* (un aimant) – *bir* (CVC) and *lab* (CVC) ;
. *galab* (après-midi) – *gal* (CVC) and *ab* (VC).
Cette particularité de la langue somalienne nous semble très intéressante, car les racines représentent un niveau intermédiaire entre les mots complets et les phonèmes/lettres. Nous verrons dans le chapitre “reconnaissance de la parole” qu’elles peuvent être utilisées pour la transcription. Afin d’étudier ces racines, nous avons mis au point un programme qui extrait les racines des mots somaliens. 4 400 racines ont ainsi pu être extraites.
Différents transducteurs ont également été développés pour traiter les abréviations, les dates, les nombres, etc.. qui apparaissent dans les corpus. Un phonétiseur (SOMPHON) de textes somaliens, inspiré du phonétiseur LIA_PHON [2] a également été créé.

5. EXPÉRIENCES

5.1. Modélisation acoustique

Un modèle acoustique somalien de départ utilisant un modèle français a été construit à l’aide d’une table de concordance entre les phonèmes des deux langues. Ce premier modèle a permis d’initialiser un processus itératif constitué d’une phase d’alignement puis d’apprentissage, afin d’obtenir un modèle acoustique somalien. Nous avons utilisé deux types de tables de concordance pour construire le modèle initial. La première table a été définie en utilisant les connaissances expertes des 2 systèmes phonétiques. Nous avons utilisé une méthode basée sur la matrice de confusion entre les phonèmes des

[5] C=Consonne, V=Voyelle

2 langues pour construire la deuxième table de concordance (sans connaissance *a priori* du système phonétique de la langue cible). Les résultats obtenus avec les deux méthodes [9] sont comparables et confirment les précédents travaux [4]. Nous avons adopté une représentation en 36 modèles pour la langue somalienne. Les voyelles longues et courtes sont très différentes dans leur durée d'exécution comme le montre la figure 2. Les voyelles longues sont en moyenne 1,86 fois plus longues. Ce constat nous a amené à modéliser les voyelles longues et courtes avec des modèles différents.

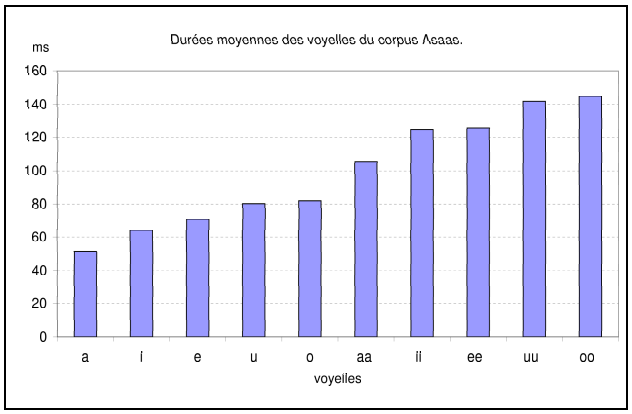

FIG. 2: Durées moyennes des voyelles du corpus Asaas. (Le rapport moyen des durées longue/courte est de 1,86)

Le signal est paramétrisé par 39 coefficients : 12 coefficients MFCC et l'énergie, plus leurs dérivées premières et secondes. Les paramètres sont centrés et réduits. Les modèles acoustiques sont composés de 3 états par phonème, excepté le " ' "[6] qui lui est codé avec 1 état, compte tenu de sa vitesse d'exécution. Nous avons utilisé pour les expériences décrits dans ce papier des modèles non contextuels avec 128 gaussiennes par état.

5.2. Modélisation linguistique

Un modèle de langage trigramme a été appris sur le corpus textuel somalien avec les outils du LIA et du CMU [12]. Ce modèle est composé de 726k bigrammes et de 1,75M trigrammes. La perplexité, calculée sur le corpus de test, est de 63,97 avec un taux de mot hors vocabulaire (OOV)[7] de 6,77%. Un lexique composé des 20k mots les plus fréquents en a été extrait et a été phonétisé à l'aide de SOMPHON. De la même façon, nous avons appris un modèle de langage basé sur les racines. Pour cela, le corpus de texte a été entièrement décomposé en racines. Le modèle de langage obtenu compte 4,4k racines, 189k bigrammes de racines et 996k trigrammes. L'ensemble des 4 400 racines a été retenu et phonétisé pour constituer le lexique. La perplexité calculée sur le corpus de test, lui même transformé en racines, est de 19,05 et le taux de racines hors vocabulaire est de 0,03%.

5.3. Reconnaissance de la parole

Les premiers résultats de reconnaissance automatique de la parole en langue somalienne avec le moteur Speeral du LIA [11] ont donné un taux d'erreur mot de 20,9%. C'est un résultat encourageant, compte tenu de la petite taille de nos corpus de départ. Signalons toutefois que les mêmes locuteurs se retrouvent dans le test et l'apprentissage. La normalisation des formes orthographiques a amené un gain relatif de 34% (WER=32% avec les données non normalisées). La table 3 donne les détails des résultats obtenus.

TAB. 3: Taux d'erreur mots pour la reconnaissance de la parole en langue somalienne, avec et sans normalisation.

	Corrects	Sub	Dél	Ins	WER
Non normalisé	75,2	19,2	5,6	7,1	32,0
Normalisé	84 ;2	13,2	1,9	5,2	20,9

Le but final de notre travail n'est pas de retranscrire le plus fidèlement possible le somalien, mais de trouver un mode de représentation des données audio permettant leur indexation. C'est pourquoi, nous avons voulu évaluer les performances du système de transcription au niveau des racines. Pour cela, les hypothèses fournies par le système Speeral ainsi que les fichiers de références ont été décomposés en racines. Nous avons obtenu un taux d'erreur Mot-racines[8] (WRER) de 14,2%. Ce résultat est intéressant, compte tenu de nos objectifs. Une indexation basée sur les racines et non sur les mots pourrait s'avérer plus appropriée pour les données que nous projetons de traiter. Afin d'étayer notre hypothèse, nous avons également effectué une reconnaissance basée uniquement sur les racines, en utilisant le lexique des 4 400 racines et le modèle de langage appris sur ces racines. Le taux d'erreur racines[9] (RER) est de 18,3%. La table 4 donne les détails des résultats du WRER et du RER.

TAB. 4: Taux d'erreur mots-décomposés-en-racines (WRER) et taux d'erreur racines (RER) pour la reconnaissance de la parole en langue somalienne.

	Corrects	Sub	Dél	Ins	Taux erreur
WRER	87,8	8,0	4,2	1,9	14,2
RER	83,3	10,8	5,9	1,7	18,3

Comme déjà mentionné, il est difficile pour ne pas dire impossible de trouver des corpus textuels correspondants aux périodes des données que nous souhaitons traiter. Le présent corpus, issu de l'Internet, n'est pas forcément adapté à cette tâche. Un décalage temporel et thématique est à prévoir entre les données d'apprentissage et les archives culturelles, qui se traduira, entre autres, par une augmentation des mots hors vocabulaire et une baisse du taux de reconnaissance. Les racines présentent de nombreux avantages par rapport aux mots. Etant à la base de la constitution de la langue, elles ont la capacité de la représenter avec peu d'individus (4 400 pour la totalité de notre corpus). Ce nombre de racines augmente la représentativité d'un corpus textuel de taille limité, diminue les mots hors vocabulaires et permettra également d'accroître la portée

[6] Occlusive glottale.
[7] Out Of Vocabulary
[8] Word-root error rate
[9] Root error rate

des modèles de langage (4 ou 5 grammes). Afin de comparer la robustesse des différentes représentations, nous avons calculé les taux d'erreur mots et d'erreur racines obtenus sur deux corpus de tests différents [10]. Les résultats obtenus sont présentés dans la figure 3. Le WER augmente de 24,8%, le RER de 7,1% et le WRER de 12,6%. Le RER semble donc moins sensible par rapport au WER et également par rapport au WRER. D'autres expériences sur des corpus de période différentes devront être menées pour confirmer nos hypothèses.

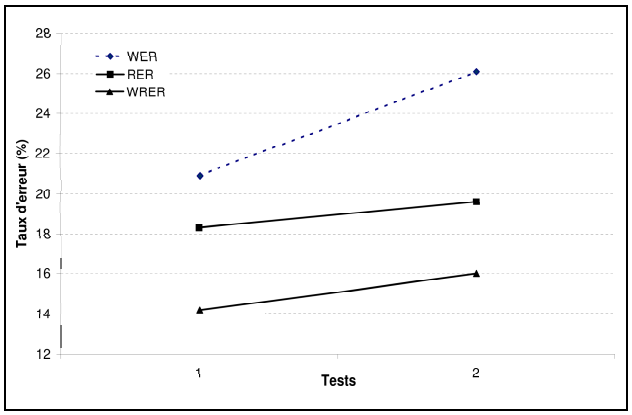

FIG. 3: Comparaison de l'évolution du WER, du WRER et du RER pour deux corpus différents.

6. CONCLUSIONS ET PERSPECTIVES

Dans ce travail, nous avons constitué un premier corpus audio de la langue somalienne. Les premiers résultats obtenus sur la reconnaissance de la parole en langue somalienne sont encourageants, compte tenu de la taille réduite de nos corpus. Nous avons aussi montré qu'un travail préalable de normalisation est nécessaire (gain relatif de 34% du WER) pour cette langue et probablement aussi pour l'ensemble des langues récemment transcrites. Nous avons également confirmé les travaux précédents concernant [16], [5] l'utilisation des documents provenant de l'Internet pour la constitution d'un corpus textuel et [4] pour la méthode rapide de modélisation acoustique. La reconnaissance automatique de la langue somalienne, basée sur les racines, semble une voie intéressante, du fait de sa moindre sensibilité aux décalages entre les données de test et d'apprentissages.

Les travaux futurs concerneront prioritairement la confirmation des résultats montrés par ces travaux par des expériences sur une plus grande échelle. Si l'utilisation des racines se confirme au niveau du traitement des données audio comme étant plus robuste, il faudra étudier l'impact d'une telle représentation plutôt que celle en mots en recherche documentaire. Enfin, nous tenterons de transposer les résultats obtenus à la langue afare, parlée à Djibouti, et qui concerne également une grande partie des données ciblées.

[10] Il s'agit uniquement d'une différence thématique dans cette expérience. Les deux corpus sont de la même période.

7. REMERCIEMENTS

Ce travail a été financé par le centre d'études et de recherches de Djibouti (CERD), le service de coopération et d'action culturelle (SCAC) du ministère des affaires étrangères français et le laboratoire informatique d'Avignon (LIA).

RÉFÉRENCES

[1] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, 1-2(33) :5–22, 2001.

[2] F. Bechet. Lia_phon : Un système complet de phonétisation de textes. *Traitement Automatique des Langues*, 2(1) :47–67, 2001.

[3] Sabrina Bendjaballah. La palatisation en somali. *Linguistique Africaine*, (21 - 98), 1998.

[4] Huerta J.M. Khudanpur S. Marthi B. Morgan J. Peterek N. Picone J. Wang W. Beyerlein P., Byrne W. Towards language independant acoustic modeling. *IEEE workshop on automatic speech recognition and understanding*, 1999.

[5] Rayid Ghani, Rosie Jones, and Dunja Mladenic. In *Mining the web to Create Minority Language Corpora*, Berlin, 2000.

[6] Larry Hyman. Tonal accent in somali. *Studies in African linguistics*, (12) :169–203, 1981.

[7] David Le-Gac. Structure prosodique de la focalisation : cas du somali et du français, 2001.

[8] R. De Mori. *Spoken dialogues with computers*. Academic Press, 1998.

[9] A. Nimaan, P. Nocera, and J.F. Bonastre. Towards automatic transcription for indexing the djibouti oral cultural patrimony. In *LREC 2006*, Genevo, ITALIA., 2006.

[10] A. Nimaan, P. Nocera, and J.M Torres-Moreno. Boîte à outils tal pour des langues peu informatisées : le cas du somali. In *JADT 2006 Journées d'Analyses des Données Textuelles*, Besançon, FRANCE., 2006.

[11] P. Nocera, G. Linares, D. Massonie, and L. Lefort. Brno. In *Phoneme lattice based A* search algorithm for speech recognition*, TSD2002, 2002.

[12] R. Rosenfeld. The cmu statistical language modeling toolkit, and its use. In *ARPA Spoken Language Technology Workshop*, Austin, TEXAS, USA., 1995.

[13] John Saeed. *Somali reference grammar*. Dunwoody Press, MD, 1993.

[14] John Saeed. *Somali (London Oriental and African Language 10)*. Johns Benjamins Publishing Company, Amsterdam/Philadelphia, 1999.

[15] Unesco. Convention pour la sauvegarde du patrimoine culturel immateriel. *http ://www.unesco.org/*, 2003.

[16] D. Vaufreydaz, M. Akbar, and J. Roullard. Asru'99. In *Internet documents : a rich source for spoken language modelling*, pages pp. 177 – 280, Keystone Colorado (USA), 1999. Workshop.