Towards automatic transcription of Somali language

Nimaan Abdillahi [*†], Nocera Pascal [†], Bonastre Jean-François [†]

[†] Laboratoire Informatique d'Avignon - CNRS / Université d'Avignon et des pays du Vaucluse
BP 1228 84911 Avignon, Cedex 9, France
[*] 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

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. This paper presents the first steps in the building of an automatic speech to text transcription for African oral patrimony, particularly the Djibouti cultural heritage. This work is dedicated to process Somali language, which represents half of the targeted Djiboutian audio archives. The main problem is the lack of annotated audio and textual resources for this language. We describe the principal characteristics of audio (10 hours) and textual (3M words) training corpora collected. Using the large vocabulary speech recognizer engine, Speeral, developed at the Laboratoire Informatique d'Avignon (LIA) (computer science laboratory of Avignon), we obtain about 20.9% word error rate (WER). This is an encouraging result, considering the small size of our corpora. This first recognizer of Somali language will serve as reference and will be used to transcribe some Djibouti cultural archives. We will also discuss future ways of research like sub-words indexing of audio archives, related to the specificities of the Somali language.

1. Introduction

In most African countries, the cultural and historic patrimonies are inherited orally through generations. This ancestral knowledge gathered during centuries is today threatened of disappearing due to the lack of interest of the young generations for the traditional way of life. Several national and international organizations (Unesco, 2003) are elaborated policies to save this human richness.

Today, most of the African countries have databases of cultural audio archives, coming mostly from radio broadcast sources, and recorded during the last forty years. They are now concerned by two main issues: saving this patrimony by digitalizing the recordings and exploiting the data. Concerning the first problem, the techniques are well known and digitalization is mostly a logistic problem. The second problem is less straightforward as facing a huge amount of data requires automatic tools for each of the different African languages involved (Berment, 2004). Particularly, automatic transcription and indexing tools are necessary for accessing the richness of the databases.

This paper presents the first step of the automatic transcription and indexing of Djibouti multicultural heritage. First, we present the Djibouti languages, the different corpora collected and theirs characteristics. Secondly, we describe the normalization tools as well as a first Somali large vocabulary speech recognizer. We also describe the different experiments and their results. Finally, we will discuss future works and perspectives.

2. Djibouti languages

Four languages are spoken in Djibouti. French and Arabic are official languages, Somali and Afar are native and widely spoken. This work is dedicated to process Somali language, which represents half of the targeted audio archives. This language is spoken in several countries of the East of Africa (Djibouti, Ethiopia, Somalia and Kenya) by a population estimated between 12 to 15 millions of inhabitants [1]. It is a Cushitic language within the Afro-asiatic family. The different variants are Somali-somali, Somali-maay, Somali-dabarre, Somali-garre, Somali-jiiddu and Somali-tunni. Somali-somali and Somali-maay are the most widely spread dialects (80% and 17%). We only process the Somali-somali variant, frequently known as Somali language and spoken in Djibouti.

The phonetic structure of this language (Saeed, 1999) has 22 consonants and 10 vowels, 5 long and 5 short. Table 1 resumes the Somali consonant phonetic structure. Somali is also a tone accent language with 2 to 3 lexical tons (Hyman, 1981), (Saeed, 1993), (Le-Gac, 2001). The written system was adopted in 1972 (SIL, 2004), and there are no textual archives before this date. It uses Roman letters and doesn't consider the tonal accent. Somali words are composed by the concatenation of a small number of sub words, named "roots" in this paper. Their forms are mostly (Bendjaballah, 1998) CVC, CVVC, CVV, VC [2] , etc. For example:

- *birlab* (a magnet) – *bir* (CVC) and *lab* (CVC);
- *galab* (afternoon) – *gal* (CVC) and *ab* (VC).

3. Corpora constitution

3.1. Somali textual corpus

The main difficulty for ASR development in African languages is the lack of textual corpus. This is mainly due

[1] http://www.ethnologue.com
[2] C=Consonant, V=Vowel

	Labial	Labiodental	Dental	Alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Voiced plosives	b		d		dh		g	q		’
Voiceless plosives		t				k				
Nasal	m			n						
Voiceless fricatives		f		s		sh		kh	x	h
Voiced fricatives						j			c	
Trill				r						
Lateral				l						
Approximants	w					y				

Table 1: Somali-consonant phonetic structure.

to the oral tradition and the industrial development of these countries.

With the development of the information technologies, many works have been undertaken to solve this problem by using Internet documents for the resource-scarce languages (Ghani et al., 2000), (Vaufreydaz et al., 1999). We applied this kind of strategy and downloaded from Internet various Somali documents (total of 3M words). As shown on table 2, we split it in two subsets (one for the speech corpus recordings and the other for the language model training). The textual corpus contains 2 820k words and 121K different words. Table 3 shows the distributional properties of this textual corpus.

	Words	Sentences
Speech corpus (Asaas)	59k	1.6k
Textual corpus	2 820k	84.7k
Total	2 879k	86.3k

Table 2: Distribution of the Somali speech and textual corpora.

Unit	Total
Sentences	84.7k
Words	2 820k
Distinct words	121k
Roots	6 042k
Distinct roots	4.4k
Phones	14 104k
Distinct phones	36

Table 3: Distributional properties of the Somali textual corpus.

3.2. Somali audio corpus : Asaas

The text selected for the speech corpus was read by 10 speakers, chosen in Djibouti area. All speakers are Somali natives from 20 to 60 years old. The recordings were done in a quiet environment. Durations vary between 30 to 120 minutes, depending of the speaker availability. We obtained a Somali audio corpus named “Asaas” composed of 10 hours of speech and the corresponding transcriptions in Transcriber format (Barras et al., 2001). It contains 59k words (10k different words) and it is digitalized with a sampling rate of 16 KHz and a precision of 16 bits. The figure 1 shows the phonetic repartition of the audio and the textual corpus. This corpus was divided into two subsets: 9,5 hours for the training subset and 0,5 hours for the evaluation subset. The figure 2 shows the phoneme duration in Asaas corpus.

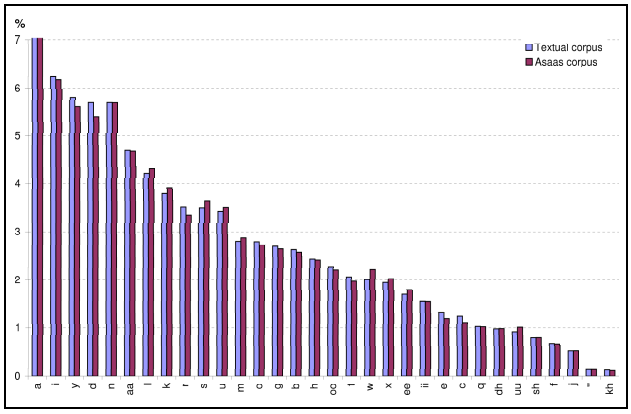

Figure 1: Phonetic distribution of the two corpora audio and textual. (The phone “a” itself represents 21% of all the phones)

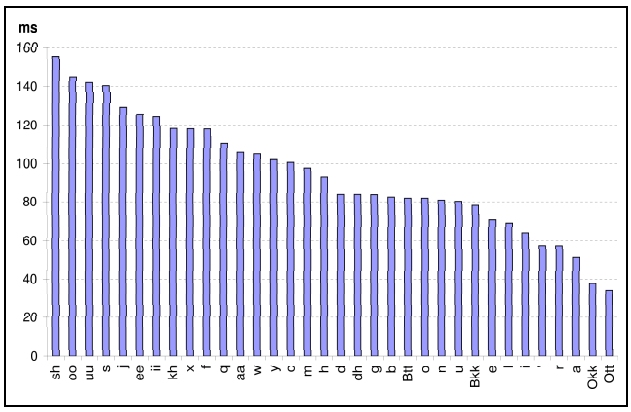

Figure 2: Phoneme duration in Asaas audio corpus. The *Okk Ott* and *Bkk Btt* correspond to the occlusive and the burst parts of the phones “k” and “t”.

4. Somali text toolbox

Several tools (Nimaan et al., 2006) have been developed to process Somali texts for audio and language processing.

4.1. Normalization

Somali language is a recent written language. As explained in the previous chapter, the spelling is not normalised. The same word can be written with a wide range of different forms. For example :

- *jibuuti, jabuuti, jibbuuti, jabbuuti, jabuudti* (Djibouti);
- *wargeys* and *wargays* (newspaper);
- *dhow* and *dhaw* (near);
- *weftiga* and *waftiga* (delegation) etc.

Another difficulty is due to the morphology of Somali words (concatenation of roots). Some words appear sometimes splited in two components. For example:

- *ka dib* and *kadib* (after);
- *mahad celin* and *mahadcelin* (thanks) etc.

These multi-spelling forms must be taken into account for the development of human language technologies for languages with recent written form. To solve this problem, we have developed a set of tools of Somali text normalization. To each word in a text, is associated its most frequent written form. If the word *dhaw* appears 11 times in the corpus and *dhow* 7 times, *dhaw* will be considered as the exact transcription.

4.2. Transducers

As in other languages, a series of transducers have been developed to transform into textual-form the different abbreviations and numbers which appear in the corpus, like dates, telephone numbers, money, etc. Examples:

- 00-253-343098 : *eber eber laba shan sadex sadex afar sadex eber sagaal sideed* (zero, zero, two, five, three, three, four, three, zero, nine eight)
- 14/10/2005 : *afar iyo tobankii bishii tobnaad laba kun iyo shan* (fourteen of the tenth month two thousand five)
- 452548 : *afar boqol laba iyo konton kun shan boqol sideed iyo afartan* (Four hundred and fifty two thousand five hundred and forty and eight)

4.3. Other tools

For future works, a morphological analyzer has also been developped for extracting roots from Somali words. We chose 4 types of roots : CVC, CV, VC and V. We first extract the CVC roots from words, after the CV roots, and finally the VC and V. This algorithm produces 4400 different roots for the whole corpus. We also developed a Somali phonetizer named SOMPHON to transform text into phonemes, inspired by LIA_PHON (Bechet, 2001), for the audio modelling.

5. Experiments

In this section, we describe our first Somali large vocabulary recognition system.

5.1. Acoustic models

The first Somali acoustic model was obtained from a French one, and was used, as a baseline, to produce the first audio segmentation of the Asaas corpus. To build this model, we established a concordance table between Somali and French phonemes. The first audio segmentation was used to produce a new Somali acoustic model with the LIA acoustic modelling toolkit. We iterated the segmentation and learning processes many times. We also tried a different initialisation by using the confusion matrix between French and Somali phonemes, to obtain an automatic baseline model. Figure 3 resumes the results obtained by the two initialisations (knowledge-based and automatic). After 3 iterations, the results are similar. This confirm the previous studies done for a fast language independent acoustic modelling methods (Beyerlein P., 1999).

We adopted 36 models for the Somali. The speech signal is parameterized using 39 coefficients: 12-mfcc coefficients plus energy and their first- and second-order derivative parameters. The cepstral mean removal and the normalization of the variance have been performed sentence by sentence.

Acoustic models are composed of 3 states by phoneme, except for the glottal plosive phoneme coded on one state (taking into account its duration). For the moment, we used non contextual models with 128 Gaussian components by state.

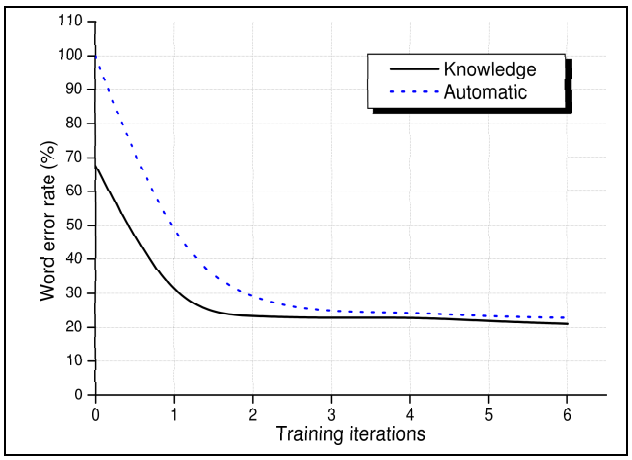

Figure 3: Learning process for the Somali acoustic model with knowledge-based and automatic methods. The decoding was done with a trigram language model.

5.2. Language model

A trigram language model trained on the Somali textual corpus with the CMU toolkit (Rosenfeld, 1995) has been obtained. We extract a 20K word lexicon from the most frequent words and a canonical phonetic form was produced

for each entry using SOMPHON tool. The language model is composed of 726K bigram and 1.75M trigram. The perplexity of the language model on the test corpus is 63.88 with 6.69% of Out-Of-Vocabulary words.

5.3. Results

This paragraph presents the first results of the ASR system for the Somali language. Speech decoding is made with the LIA large vocabulary speech recognition system Speeral (Nocera et al., 2002). The same speakers are in the test and the training sets. We obtain a word error rate of 20.9% on the 30 minutes test corpus as shown in table 4. This is an encouraging result according to the size of the training corpora (9,5 hours for the audio and 3M words for LM).

Without the spelling normalization presented int section 4.1, the error rate is 32%. This shows that the normalization process is necessary for recent written languages. When the evaluation is done at the root instead of the word level, we obtain a word-root error rate of 14.2% as shown in table 5.

	Correct	Sub	Del	Ins	WER
Not normalized	75.2	19.2	5.6	7.1	32.0
Normalized	84.2	13.9	1.9	5.2	20.9

Table 4: Results of the Somali automatic speech recognition in %, with a normalized and a raw text.

	Correct	Sub	Del	Ins	Error rate
Root	87.8	8.0	4.2	1.9	14.2

Table 5: The Word-root error rate (WRER) of 14.2% is obtained with the word hypothesis files. It is an encouraging result for indexing the audio archives with roots.

6. Conclusions and perspectives

Results of this first Somali large vocabulary recognizer are encouraging. We demonstrate that a normalizing process is necessary for Somali language and probably for all recent written languages. We reduce the WER of about 34%, after the normalization process. This work is the first step for the automatic transcription for indexing Djibouti cultural audio heritage. Our final objective is not to transcribe exactly audio archives, but rather to obtain an index table (based on an approximate transcription) in order to build a speech mining system.

One perspective of this work is to work in a root-based decoder in order to be more robust to thematic and temporal mismatch between training and testing corpora. We also project to transpose our results to the Afar language spoken in Djibouti. We believe that the work done within this project will be useful not only to the Somali language but to several oral tradition countries.

7. acknowledgment

This research is supported by the Centre d'Études et des Recherches de Djibouti [3] (CERD), the Service de Coopération et d'Action Culturelle [4] (SCAC) and the Laboratoire Informatique d'Avignon [5] (LIA).

8. References

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

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

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

Vincent Berment. 2004. Méthodes pour informatiser des langues et des groupes de langues "peu dotées".

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

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

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

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

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

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

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

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

International SIL. 2004. *Ethnologue : Language of the World. 14th edition*. USA.

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

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

[3] http://www.cerd.dj

[4] http://www.ambafrance-dj.org/

[5] http://www.lia.univ-avignon.fr