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'Etudes 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. Automatic transcription and indexing tools seem potential solution to preserve it. This paper presents the first steps 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 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 also discuss future ways of research like roots indexing of audio archives.

Index Terms : resource-poor languages, speech recognition, African languages, oral patrimony indexing

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 globalization process, the economic situation and the lack of interest of the young generations for the traditional way of life. Several national, regional and international organizations [17] are elaborated policies to save this human richness. Today, African countries have databases of cultural audio archives coming mostly from radio broadcast sources, [1] 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. Particularly, automatic transcription and indexing tools are necessary for accessing the richness of the databases. These tools are language-dependent and need to be adapted to each of the African languages targeted. This work is focused on Somali language. First, we present the Djiboutian languages and more precisely the Somali one. We describe also the different corpora collected and the normalisation tools. Secondly, we present the results obtained on a word based system and a root based system. Finally, we discuss about future works.

[1] The republic of Djibouti launched a wide digitalization program of radio broadcast archives. http://www.rtd.dj

2. Djibouti languages

Four languages are spoken in Djibouti. French and Arabic are official languages, Somali and Afar are native and widely spoken. Somali and Afar are Cushitic languages within the Afro-asiatic family. Somali 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 [2]. 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 variants (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 has 22 consonants and 5 basic vowels which all occur in front and back versions (+ATR or -ATR). These 10 all occur in long and short pairs, giving 20 in total [14]. There is also 5 diphthongs which occur in front and back, long and short versions. Somali is also a tone accent language with 2 to 3 lexical tones [9], [13], [10]. The written system was adopted in 1972 [16], and there are no textual archives before this date. It uses Roman letters and doesn't consider the tonal accent in the current form. Somali words are composed by the concatenation of syllable structures [4], [14]. In this work we choose only four structures : V, CV, VC and CVC [3] named "roots" in this paper.

3. Corpora constitution

3.1. Textual corpus

Automatic speech recognition can reach a good level of performance if enough data (both textual and audio corpora) are available. The difficulty for ASR development in African languages is the lack of corpora. This is mainly due to the oral tradition system and the industrial development of these countries. With the development of the information technologies, many works have been undertaken by using Internet documents for the resource-scarce languages [8], [18]. We applied this kind of strategy and downloaded from Internet various Somali documents. The textual corpus gathered contains 2 820k words and 121K different words.

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

Table 1 shows the distributional properties of this textual corpus.

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

Table 1: Distributional properties of the Somali textual corpus.

3.2. Audio corpus

We also downloaded a subset of text from Internet for the audio recordings. This text was read by 10 speakers. The recordings were done in a quiet environment. We obtained a Somali audio corpus named "Asaas" [4] composed of 10 hours of speech and the corresponding transcriptions in Transcriber format [2]. It contains 59k words (10k different words) and it is digitalized with a sampling rate of 16 KHz and a precision of 16 bits. This corpus was divided into two subsets: 9,5 hours for the training subset and 0,5 hours for the evaluation subset. Figure 1 shows the phoneme duration in Asaas corpus. The figure 2 shows the phonetic distribution of the textual and audio corpora. The two distributions are similar. The audio corpus is phonetically well balanced.

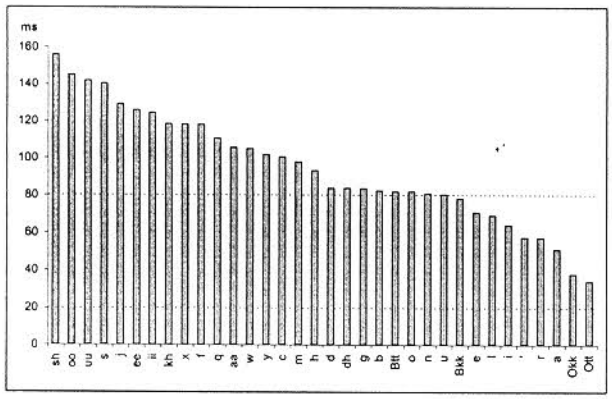

Figure 1: Phoneme duration in Asaas audio corpus.

4. Normalisation tools

Several tools [?] have been developed to process Somali texts for audio and language processing. As explained before, Somali language is a recent written language. The spelling is not rather normalised. The same word can be written with a wide range of different forms (*jibuuti, jabuuti, jibbuuti, jabbuuti, jabuudti*). Another difficulty is due to the morphology of Somali words (concatenation of roots). Some words appear sometimes splited in two components (*ka dib* and *kadib*). These multi-spelling forms must be taken into account for the development of human language technologies for this language. To solve this problem, we have developed a Somali normalization tool. 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. A serie 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. A morphological analyzer has also been developed for extracting roots from Somali words. We choose 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 the French one LIA_PHON [3], for the audio modelling.

[4] Asaas means beginnings in Somali language

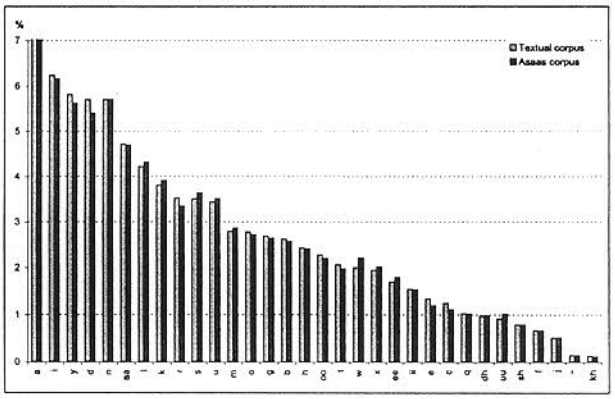

Figure 2: Phonetic distribution of textual and audio corpus. The *Okk Ott* and *Bkk Btt* correspond to the occlusive and the burst parts of the phonemes "k" and "t".

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 shows the results obtained by the two initialisations methods (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 [15], [5].

In this work, we take into account only 10 vowels (5 longs and 5 shorts). We don't consider the front and back features

and the diphthongs. So, our acoustic model is composed by 36 models. Each acoustic model corresponds to one phoneme and is composed of 3 states, except for the glottal plosive phoneme coded on one state (taking into account its duration). We use non contextual models with 128 Gaussian components by state. 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.

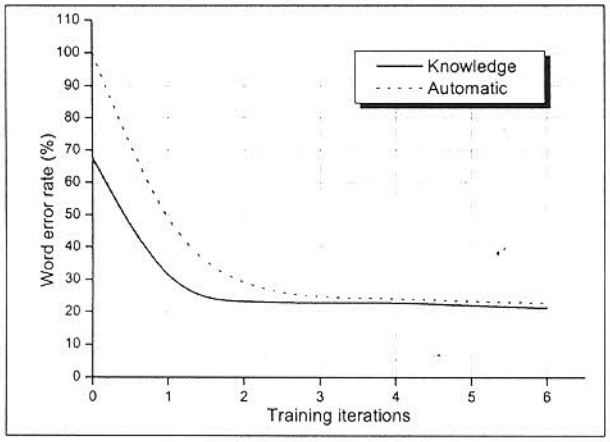

Figure 3: Learning process for the Somali acoustic model with knowledge-based and automatic methods. The decoding was done with a trigram language model.

5.2. Language model

A trigram language model is trained on the Somali textual corpus with the LIA toolkit and CMU toolkit [12]. We extracted a 20K word lexicon from the most frequent words and a canonical phonetic form was produced for each entry using Somali phonetizer. The language model is composed of 726K bigram and 1.75M trigram. The perplexity of the language model on the test corpus is 63.97 with 6.77% of Out-Of-Vocabulary words. Likewise, we trained a trigram language model based on roots. The entire textual corpus was transformed in roots form. We obtained 4.4k unigram, 189k bigram and 996k trigram of roots. The perplexity of this model is 19.05. With the test corpus, we obtained 0.03% of Out-Of-Vocabulary roots. Table 2 resumes the principal characteristics of the two models.

	unigram	bigram	trigram	perplexity	OOV (%)
Word	20k	726k	1750k	63.97	6.77
Root	4.4k	189k	996k	19.05	0.03

Table 2: Principal characteristics of the word- and root-based language model.

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 [11]. The same speakers are presents 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 3. 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 normalisation presented above, the error rate is 32%. This shows that the normalisation 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 3. We decomposed the hypothesis and the reference files in roots.

We also do a root based decoding by using the root language model. We obtain a root error rate (RER) of 18.3% as shown in table 4. These results are encouraging for indexing the audio archives.

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

	Correct	Sub	Del	Ins	Error rate
WRER	87,8	8,0	4,2	1,9	14,2
RER	83,3	10,8	5,9	1,7	18,3

Table 4: Word-root error rate (WRER) and root error rate (RER) in %.

There is a big mismatch between the text corpora obtained from the World Wide Web in order to build the language models and the audio archives we want to transcribe. This mismatch will increase significantly the OOV rate and of course the WER. The roots present many advantages because they are the fundaments basic elements of the Somali words and because this set doesn't vary along the decades. The limited size of the root set helps also to decrease the OOV rate. The figure 4 shows the results obtained with another test corpus showing a higher OOV rate. A 24.8% relative WER increase could be noticed when the RER increases of only 7.1%. RER seems less sensitive than the WER and the WRER. Different experiments might be done to confirm this result.

We planned also to combine the two languages models (words/roots) in order to increase the recognition rate as explained in [6].

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. We also confirm the fast acoustic modeling for a new language and the use of Internet documents for resource-scarce language modeling. We obtain a first result which shows that