# Contribution
# to the
# ITU and UNESCO Global Symposium on Promoting the Multilingual Internet

**Geneva, 9-11 May 2006**

| | |
|---|---|
| Submitted by: | Université d'Avignon<br>Nimaan Abdilahi |
| Title: | Speech mining to make African oral patrimony accessible |
| For Discussion in Session: | 3: Scripting operational experiences |

# Speech mining to make African oral patrimony accessible

**Nimaan Abdillahi** *†, **Nocera Pascal** †, **Bonastre Jean-François** †, **Bechet Fréderic** †

† 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, frederic.bechet}@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 for automatic transcription and indexing of African oral tradition heritage, particularly the Djibouti cultural heritage.

## 1. Context of this study

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 (Unesco, 2003) are elaborated policies to save this human richness. Today, 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. 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 the processing of the Djibouti oral patrimony [1]. The republic of Djibouti launched a wide digitalization program of radio broadcast archives [2].

## 2. Methodology

Accessing the richness of the African oral patrimony requires automatic search engine. Most of the known search engines concern text. So, a transcription phase is necessary for the automatic indexing of audio archives. Nowadays, statistical Automatic Speech Recognition (ASR) systems can reach a good level of performance for a wide range of languages if training data (both for the acoustic and linguistic models) are available. Unfortunately, it is difficult to get enough textual corpora for African languages. This is mainly due to the oral tradition system of these countries. With the development of the information technologies, Internet documents appear like a probable solution. Several works have been undertaken by different teams (Ghani et al., 2000), (Vaufreydaz et al., 1999) for the automatic construction of textual corpus for the resource-scarce languages [3] by using web documents. However, this method is limited by the number and quality of the websites available on Internet for each language. For some of them, there are not enough data. In this case, others solutions must be considered. Anyway, this kind of collected corpus is not adapted to the audio archives we project to transcribe. There are temporal and thematic mismatch between them. This mismatch will produce high out of vocabulary (OOV) rate and high word error rate (WER). However, the NIST Topic Detection and Tracking (Fiscus and Doddington, 2002) and TREC document retrieval evaluation programs has studied the impact of recognition errors in the overall performance of Information Extraction systems for tasks like story segmentation or topic detection and retrieval. The results obtained by (Fiscus and Doddington, 2002) have shown that this impact was very limited, compared to those obtained on *clean* text corpora. Similar results were obtained during the TREC program for a document retrieval task (Barnett et al., 1997). Then, to automatically index African audio patrimony we present the speech mining system shown in figure 1. It is composed of

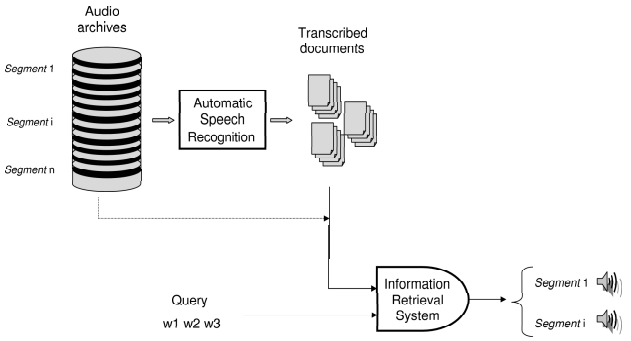


Figure 1: Speech mining system

[1] A part of the Djiboutian oral heritage is in Somali language. Other part is in Afar language

[2] http://www.rtd.dj

[3] In the literature, the terms "minority language", "less-equipped languages" and "resource-scarce languages" appears to design "less computerized" languages

| | 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-consonants phonetic structure.

two modules: automatic speech recognition (ASR) one and an information retrieval (IR) one. The information retrieval system uses the hypothesis files provided by the ASR module. In this paper, we present, the Djibouti languages and more precisely Somali one. Secondly, we describe our first Somali speech recognizer and the results obtained. Finally, we discuss about using Somali “roots” to deal with the inevitable mismatch between the audio archives and the training corpus available.

## 3. 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 [4]. 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 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 (Saeed, 1999) has 22 consonants and 20 vowels (5 basic distinctions which all occur in front and back versions. These 10 all occur in long and short pairs). Table 1. resumes the consonants structure. Somali is also a tone accent language with 2 to 3 lexical tones (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, CV, VC and V [5], etc.

## 4. Automatic Speech Recognition

### 4.1. Corpora constitution

In order to build an ASR system for the Somali language, we collected an audio and a textual corpus. With the development of the information technologies, many works have been undertaken 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. The textual corpus gathered contains 2 820k words and 121K different words. Table 2 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 |
| Distinct phones | 36 |

Table 2: Distributional properties of the Somali textual 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 obtain a Somali audio corpus named “Asaas” [6] 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 audio corpus is phonetically well balanced. 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.

### 4.2. Normalisation tools

Several tools (Nimaan et al., 2006) have been developed to process Somali texts for audio and language processing. Somali language is a recent written language and the spelling is not 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

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

[5] C=Consonant, V=Vowel

[6] Asaas means beginnings in Somali language

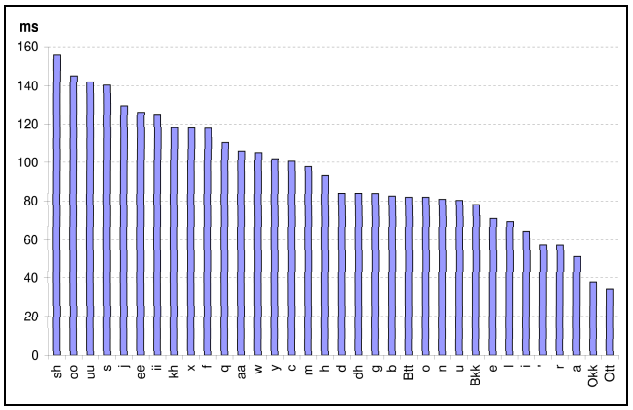


Figure 2: Phoneme duration in Asaas audio corpus.

(*ka dib* and *kadib*). 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. 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 (Bechet, 2001), for the audio modelling.

### 4.3. Experiments

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

#### 4.3.1. 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 (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[7]. 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. 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.

[7]We considered only 10 vowels (5 longs and 5 shorts)

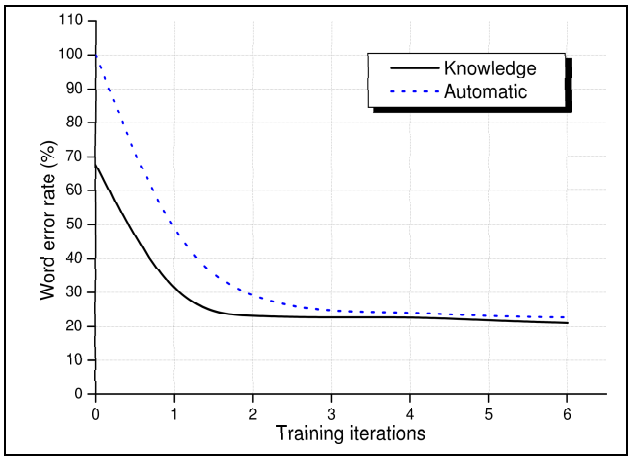


Figure 3: Learning process for the Somali acoustic model with knowledge-based and automatic methods. The decoding was done with a trigram language model.

#### 4.3.2. Language model

A trigram language model trained on the Somali textual corpus with the LIA toolkit and 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 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, 189k 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.

### 4.4. 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 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 normalization presented above, 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 4. It is an encouraging result for indexing the audio archives with roots.

| | 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 |
|---|---|---|---|---|---|
| Root | 87.8 | 8.0 | 4.2 | 1.9 | 14.2 |

Table 4: The Word-root error rate (WRER) of 14.2% is obtained with the word hypothesis files.

## 5. Information retrieval

Our aim is to make Djibouti audio archives more accessible. Obtaining exact transcriptions of this audio data is an extremely difficult task, the main difficulty being collecting text corpora matching all the different kinds of speech recorded. In most of the cases, no written corpus at all is available. For this reason our goal in this work is not to produce transcriptions of audio archives that would replace the original recording but rather word and sub-word transcriptions that can be use for performing Information Retrieval processes for accessing the audio data. Indeed, even in perfect transcriptions all the non textual information included in the audio data (prosody, emotions, ...) is lost.

## 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. One perspective 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 [8] (CERD), the Service de Coopération et d'Action Culturelle [9] (SCAC) and the Laboratoire Informatique d'Avignon [10] (LIA).

## 8. References

J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. Kuo. 1997. Experiments in spoken queries for document retrieval. In *In Eurospeech 97*, pages 1323–1326.

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).

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*.

Jonathan G. Fiscus and George R. Doddington. 2002. Topic detection and tracking evaluation overview. *Topic detection and tracking: event-based information organization*, pages 17–31.

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.

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

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.

[8] http://www.cerd.dj
[9] http://www.ambafrance-dj.org/
[10] http://www.lia.univ-avignon.fr