



Information Retrieval Strategies for Accessing African Audio Corpora

Abdillahi Nimaan^{1,2}, Pascal Nocera¹, Frédéric Bechet¹, Jean-François Bonastre¹

¹Laboratoire Informatique d'Avignon - UAPV, Avignon, France

²Institut des Sciences et des Nouvelles Technologies - CERD, Djibouti

nimaan.abdillahi, pascal.nocera, frederic.bechet, jean-francois.bonastre@univ-avignon.fr

Abstract

In this paper we present a first approach to access African oral corpora, combining automatic speech recognition and information retrieval. Firstly, we present the principal characteristics of our Somali speech recognizer [8] and the results obtained on real audio archives gathered from Djibouti Radio. Secondly, we present a Hybrid Language Model (HLM) including words and sub-words to improve the robustness against OOV words. We proceed to Information Retrieval experiments with various strategies. We search on the different outputs of the ASR system (words, sub-words and hybrid). We finally present a new strategy combining sub-words and words to enhance the information retrieval results.

Index Terms: speech recognition, information retrieval, hybrid language model, Somali language.

1. Introduction

Most of African countries have audio archives coming from radio broadcast sources. A large part of these corpora represents their cultural, historical and scientific heritage. 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 African language targeted. This work is focused on the Somali language.

Significant research has been directed toward automatic audio indexing and retrieval. The main part of these efforts has focused on spoken documents like broadcasts news [15, 12, 5]. Some other works have focused on data collections containing spontaneous speech like academic and/or scientific lecture material [11] or the recorded interview of the oral testimonies collected within the MALACH project. These works show the effectiveness of IR tools, when the transcriptions present high WER (Word Error Rate) values [7]. When for the above works related text corpora were available, it is not the case for our application context. No large text corpora are available in Somali language. Furthermore, there is no textual data matching the speech corpora we want to process. This makes the task very challenging.

In this paper, we first describe our test corpora gathered from Djibouti Radio¹ and our Somali speech recognizer, already pre-

sented in [8]. We introduce a decoding system based on a Hybrid Language Model (HLM) to improve the IR performance. We also propose an information retrieval system and study different strategies to enhance the precision and the recall rates. Finally, we highlight some perspectives.

2. Somali language

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 Afroasiatic family. Somali language is spoken in several countries of the East of Africa (Djibouti, Ethiopia, Somalia and Kenya) by a population estimated between 11 to 13 millions of inhabitants². 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 vowels occur in long and short pairs, giving 20 in total [14]. There are 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 [6, 13]. The written system was adopted in 1972, 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 [14]. In this work we choose only four structures : V, CV, VC and CVC³ named "roots" in this paper.

3. Somali Speech Recognizer

We presented in [8] a first Somali speech recognizer. We used a trigram language model trained on a Somali textual corpus gathered from the Web and composed of 2 820k words and of 121k different words. This corpus is collected from several Internet newspapers. We extracted a 20k words lexicon from the most frequent words and a canonical phonetic form was produced for each entry using a Somali phonetizer. The obtained language model is composed of 726k bigrams and 1.75M trigrams.

The acoustic model is composed by 36 Hidden Markov Models (HMM). Each acoustic model corresponds to one phoneme and is represented by 3 states, except for the glottal plosive phoneme coded on one state (taking into account its

radio broadcast archives. <http://www.rtd.dj>

²Ethnologue : Language of the World. 15th edition. USA 2005.

³C=Consonant, V=Vowel

¹The republic of Djibouti launched a wide digitalization program of

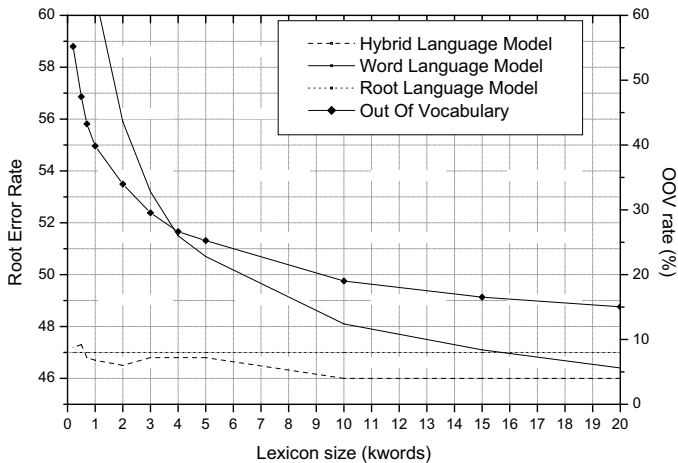


Figure 1: Comparison of the root error rate of decoding using word, root and hybrid language models and their behaviour in regards to the OOV rate.

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. We developed several tools [9] to process Somali texts for audio and language processing. Somali language is a recent written language. The spelling is not rather normalized. The same word can be written with a wide range of different forms. To solve this problem, we developed a Somali normalization tool.

The first evaluation of our Somali speech recognition was made with a 0.5 hour read test corpus, HAATUF1.0⁴. It gave a 21% Word Error Rate (WER). The system is based on the large vocabulary speech recognition engine Speeral [10]. The same speakers were in the test and train corpora. We have also considered a decoding system based on roots in order to deal with the mismatch between the text corpora obtained from the World Wide Web and the audio archives that we want to transcribe and to retrieve. This mismatch increases significantly the OOV rate and of course the WER. The roots present many advantages because they are the fundamental basic elements of the Somali words and because this set doesn't vary along the decades. The limited size⁵ of the root set helps also to decrease the Out-Of-Vocabulary (OOV) rate. Root Error Rate (RER) seems less sensitive to OOV rate than the WER.

4. RTD01.1 speech Corpus

We collected five cultural broadcasts⁶, approximately 20 minutes each from Djibouti Radio. These broadcasts are composed of interviews of seniors. Eight subjects are approached: traditional songs, the relation between horn of Africa and ancient

⁴Newspaper in Somali language: www.haatuf.net.

⁵The Somali textual corpus contains 5k roots. We used all in our experiments.

⁶They are from *War iyo waayo arag* and *Sooyaal* issues.

Egypt, the benefits of seafood, silk, two historical leaders (*Aale Boore* and *Cabdiraxmaan Saylici*⁷), the prophet of Islam and a comparison of divorce in the modern and ancient society. This corpus, denoted RTD01.1⁸, contains 7 803 words (2 378 distinct words) and 15 619 roots (931 distinct roots). It is digitalized with a sampling rate of 16 KHz and a precision of 16 bits. It is manually transcribed with Transcriber [1]. The perplexity of the language model on this test corpus is 804 with 12.60% of OOV words. The perplexity of HAATUF1.0 was 68.97 and 6.77% OOV words in our previous experiments. This shows the distance between the RTD01.1 corpus and the textual corpus used to train the language model. RTD01.1 describes situations and events of the 19th, 16th and 7th centuries and the second contains news and events between 2002 and 2004. This mismatch is intrinsically due to the oral tradition way of life of these countries⁹. The non-normalized perplexities calculated on the roots based language model are 19.05 for HAATUF1.0 and 56.66 for RTD01.1. The OOV roots rate is 0.03% both.

5. Information retrieval

5.1. Automatic transcription

With a word based language model (lexicon of 20k), the WER obtained on the RTD01.1 corpus is 62.1%. When we transform the output hypothesis files into roots we obtain 46.4% of Root Error Rate. These high rates were predictable due to the important value of OOV words. An OOV word is never recognized and the neighbouring words are often misrecognized. In order to deal with the high OOV word rate, we experiment a strategy based on roots. The text corpus is transformed in roots corpus by representing word in their roots structure. We build a new lexicon composed of 5k roots with a OOV roots rate of 0.03%. With a trigram language model trained on roots, we obtain a Root Error Rate of 47%. The output sentences composed of roots are not easily understandable but they can be used for Information Retrieval as we will demonstrate in section 5.2.

5.2. Hybrid Language Model

We also use a Hybrid Language Model (HLM) combining roots and words. This idea of implementing a sub-word based language model was proposed by [2, 16, 3]. Our method is similar to [16]. We use words and roots uniformly. We partition our lexicon in two sets. The first set is composed of the most N frequent words and called In-Vocabulary (IV) lexicon and the second is made of roots from remaining words. On this basis, we build several HLM with different sizes (N=0.2k to N=20k) of the IV lexicon. We calculate the recognition results in root error rate like shown in figure 1. The HLM is not disturbed by the size of the lexicon and the rate of the OOV words. Table 1 shows some OOV words misrecognized by the WLM and the behaviour of the HLM and the RLM. All the words not belonging to the lexicon are represented by roots. We also measured the accuracy of the words recognized with the Hybrid Model. The accuracy is calculated by dividing the correct words by the hypothesis ones. A 700 words based HLM is 26% more accurate than a 20k words based HLM in our experiments.

⁷The first is a Somali hero of the 16th century and the second is a well known religious man of the 19th century.

⁸Radio Television of Djibouti.

⁹The Somali language is officially written since 1972, so it's not possible to find a training corpus corresponding to the period of the targeted archives.

Table 1: OOV words recognized with HLM and the RLM based decoding.

Reference	WLM 20k	HLM 20k	HLM 1k	RLM
asnaamtaasi	wasaaradaasi	as naam ta si	as naam ta si	as naam ta si
tafaraaruqa	taf abaabulka	taf ar aar uq a	taf ar aar uq a	taf ar aar uq a
faaqidi	nafaqada	faaq id i	faaq id i	aq ad i
(shiinaha) qudhooda	bishii lagu looga	(shiinaha) qudh ood a	(bishii ina) qudh ood a	(bish iib a) qudh ood a
(laba) dakhare	labadaba	laba dakh ar e	labada kale	lab ad a sar e

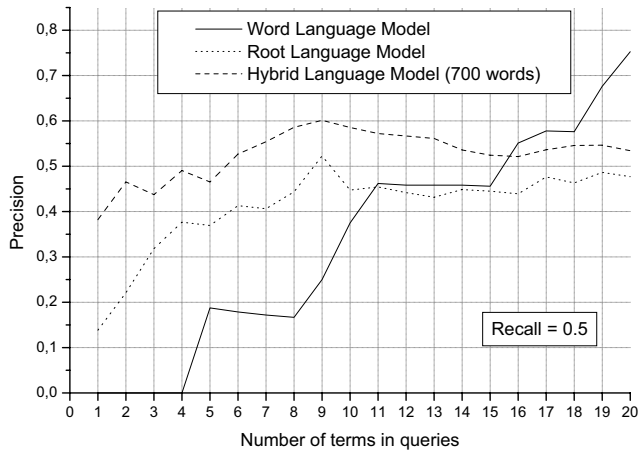


Figure 2: Impact of the number of terms in queries on the precision for the three output sentences: WLM-, RLM- and HLM-using decoding.

5.3. Information Retrieval Experiment Setup

The aim of our study is to find an efficient representation and a strategy to improve the information retrieval on the noisy ASR transcriptions. Thus, we avoid the queries definition and expansion problem. We also use a standard search engine based on vector-space model.

Within the several approaches on vocabulary creation and topic detection [4], we use the *tfidf* method. We bring together all the documents of RTD01.1 corpora related to each topic described on section 4. A list of stopwords is removed and a vector is extracted for each topic. These vectors are used like queries. We transform the queries in roots when we search within the roots outputs of the ASR system. We made the same operation with the hybrid outputs. These queries are optimal towards the documents to process. Therefore, it is only the retrieval strategy that is going to be evaluated.

5.4. Results

Table 2 indicates the number of terms in queries and the number of QOV¹⁰. The high QOV rate shows that the OOV words are often content-words. The impact of the number of terms in queries¹¹ on the information retrieval results is shown in figure 2. When the number of terms in queries is small (<10) the retrieval results from root and hybrid transcriptions are better

¹⁰Query Out-of-Vocabulary items calculated on the WLM.

¹¹We only considered the 20 most weighted terms in queries.

Table 2: Number of terms in queries and number of QOV.

	Query	Terms	#QOV
1	Traditional songs	51	21
2	Horn of Africa history	50	15
3	Benefits of seafood	51	10
4	Silk	51	16
5	<i>Aale Borre</i>	50	18
6	The prophet of Islam	50	15
7	<i>Cabdiraxmaan saylici</i>	50	21
8	Divorce	51	9

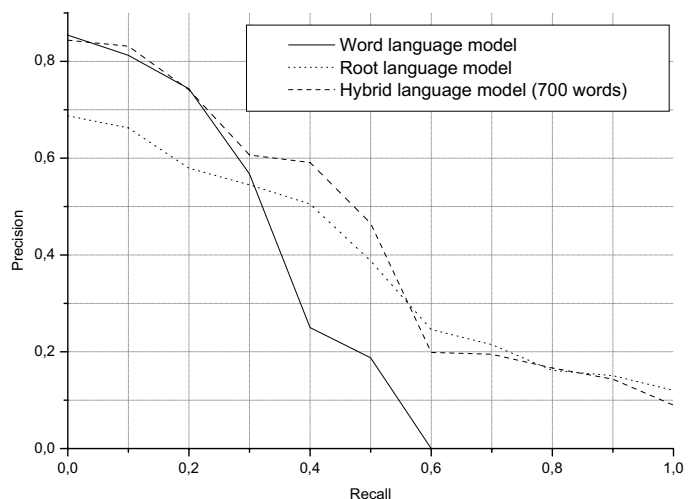


Figure 3: Precision and recall for the WLM, RLM and HLM based decoding.

than the word transcription. For a recall of 50%, the precision for 10 terms queries is: 37.5% for words, 44.73% for roots and 58.55% for hybrid. If the number of terms in queries increases, the word-based retrieval is better. For a recall of 50% and a number of terms equal to 20 the precision is: 75.26% for words, 47.69% for roots et 53.42% for hybrid. We note that the sequence of roots inside the words is lost in vectorial model. For instance, to the hybrid or root system, the word *madax*¹² will be transformed in *mad ax*. The IR system will retrieve all the documents who contain *madax*, *axmad*¹³, *ax*, *mad*, *ax**mad*, *mad**ax*, etc.

The number of terms in queries is usually less than 5 in document retrieval applications. On this basis, we compare

¹²head.

¹³firstname.

the results with queries of 5 terms. Table 3 shows the number of QOV in the queries. For the word based transcription the first retrieved documents are accurate like shown in figure 3. But the precision decreases quickly when the recall increases. A part of the relevant documents will never be retrieved. The reason is that the QOV are not presents on the transcriptions. The root-based system is less accurate for the first retrieved documents, but all the relevant documents are retrieved. The precision is 8% for a recall of 100%. For a recall of 50%, we have a precision of 38.78% for the roots approach and 18.75% for the words approach.

Table 3: Number of QOV in queries of 5 terms.

	Query	Terms	#QOV
1	Traditional songs	5	2
2	Horn of Africa history	5	0
3	Benefits of seafood	5	0
4	Silk	5	0
5	<i>Aale Borre</i>	5	1
6	The prophet of Islam	5	0
7	<i>Cabdiraxmaan saylici</i>	5	4
8	Divorce	5	0

The hybrid approach recovers the two advantages of the word approach and the root approach. For the first retrieved documents the precision is similar to using the words. When the recall increases the hybrid approach is similar to root approach. For a recall of 50%, the precision of the hybrid approach (46.53%) is better than the two others (38.78% for the roots and 18.75% for the words).

6. Discussions and future work

Due to the oral tradition of most of African countries, there is not enough corpora to model their audio archives. The non-standardization of the orthography is also another difficulty. Through this work, we demonstrate that it is possible to retrieve information in this crucial situation. In order to deal with the high OOV rate on this corpus, we demonstrate that using a Hybrid Language Model can improve the recognition results. We applied the same approach to the information retrieval part by decomposing the QOV in roots. On this basis, we improved the retrieval results.

In future work, we will study the information retrieval results with real queries taken from the web. We will investigate a better weighting of words by taking into account the degree of confidence in the output of the ASR system. We will also take into account of the root sequence in the search process.

7. Acknowledgements

This research is supported by the Centre d'Études et des Recherches de Djibouti¹⁴ (CERD), the Service de Coopération et d'Action Culturelle¹⁵ (SCAC) and the Laboratoire Informatique d'Avignon¹⁶ (LIA).

¹⁴<http://www.cerd.dj>

¹⁵<http://www.ambafrance-dj.org/>

¹⁶<http://www.lia.univ-avignon.fr>

8. References

- [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] I. Bazzi and J. Glass. Modeling out-of-vocabulary words for robust speech recognition. *In Proc. ICSLP, Beijing CHINA*, 2000.
- [3] M. Bisani and H. Ney. Open vocabulary speech recognition with flat hybrid models. *In Eurospeech 2005*, Lisbon, Portugal, 2005.
- [4] A. Brun, K. Smaili, and J.P. Haton. Nouvelle approche de la sélection de vocabulaire pour le détection de thème. *In TALN, Batz-sur-Mer, France*, 2003.
- [5] Benoît Favre. Résumé automatique de parole pour un accès efficace aux bases de données audio, 2007.
- [6] Larry Hyman. Tonal accent in somali. *Studies in African linguistics*, (12):169–203, 1981.
- [7] D.Z. Inkpen, M. Alzghool, G. Jones, and D.W. Oard. Investigating cross-language speech retrieval for a spontaneous conversational speech collection. *In HLT-NAACL, Beijing CHINA*, 2006.
- [8] A. Nimaan, P. Nocera, and J.F. Bonastre. Automatic transcription of somali language. *In ICSLP 2006*, Pittsburg, USA., 2006.
- [9] 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.
- [10] P. Nocera, G. Linares, D. Massonnie, and L. Lefort. Phoneme lattice based A* search algorithm for speech recognition. *In TSD2002*, 2002.
- [11] A. Park, T.J Hazen, and J.R Glass. Automatic processing of audio lectures for information retrieval. *In Proc. of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pages 497–500, 2005.
- [12] S. Renals, D. Abberley, D. Kirby, and T. Robinson. Indexing and retrieval of broadcast news. *In Proc. ICSLP, Beijing CHINA*, 2000.
- [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] J.-M. Van Thong, P.J. Moreno, B. Logan, B. Fidler, K. Maffey, and M. Moores. Speechbot: an experimental speech-based search engine for multimedia content on the web. *In Multimedia, IEEE Transactions on, Vol.4, Iss.1*, pages 88–96, 2002.
- [16] Ali Yazgan and Murat Saraclar., "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition.", Montreal, Canada, 2004.