Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof

Elodie Gauthier[1], **Laurent Besacier**[1], **Sylvie Voisin**[2], **Michael Melese**[3], **Uriel Pascal Elingui**[4]
[1]Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France
[2]Laboratoire Dynamique Du Langage (DDL), CNRS - Université de Lyon, France
[3]Addis Ababa University, Ethiopia
[4]Voxygen SAS, Pleumeur-Bodou, France & Dakar, Sénégal
elodie.gauthier@imag.fr, laurent.besacier@imag.fr, sylvie.voisin@ish-lyon.cnrs.fr,
michael.melese@gmail.com, elinguiuriel@voxygen.fr

Abstract

This article presents the data collected and ASR systems developped for 4 sub-saharan african languages (Swahili, Hausa, Amharic and Wolof). To illustrate our methodology, the focus is made on Wolof (a very under-resourced language) for which we designed the first ASR system ever built in this language. All data and scripts are available online on our *github* repository.

Keywords: african languages, under-resourced languages, wolof, data collection, automatic speech recognition

1. Introduction

Today is very favorable to the development of a market for speech in sub-saharan african languages. People's access to information and communications technologies (ICT) is done mainly through mobile (and keyboard) and the need for voice services can be found in all sectors, from higher priority (health, food) to more fun (games, social media). For this, overcoming the language barrier is needed. This paper is done in the context of ALFFA project[1] where two main aspects are involved: fundamentals of speech analysis (language phonetic and linguistic description, dialectology) and speech technologies (automatic speech recognition (ASR) and text-to-speech (TTS)) for African languages. In the project, developed ASR and TTS technologies will be used to build micro speech services for mobile phones in Africa. For this, speech fundamental knowledge for targeted languages has to be upgraded while African language technologies are still at their very beginning. For these reasons, the ALFFA project is really interdisciplinary since it does not only gather technology experts but also includes fieldwork linguists/phoneticians.

Paper contribution. This article presents the data collected and ASR systems developped for 4 sub-saharan african languages (Swahili, Hausa, Amharic and Wolof). To illustrate our methodology, the focus is made on Wolof (a very under-resourced language) for which we designed the first ASR system ever built in this language.

Paper outline. The outline of this paper is the following: first, section 2 will summarize ASR sytems already available on our github[2] repository for Swahili, Hausa and Amharic (together with scripts and resources needed to reproduce ASR experiments). Then, sections 3, 4 and 5 will focus respectively on text data collection, speech data collection and ASR for Wolof language. Finally, section 6 will conclude this work and give some perspectives.

2. ASR systems made available in the ALFFA project

2.1. Target languages of the project

Language choice for the project is mainly governed by population coverage and industrial perspectives. We focus on Hausa spoken by around 60 million people, as first or second language. Hausa is part of the family of Afroasiatic languages. Specifically, Hausa is the most spoken language among the Chadic languages (Vycichl, 1990). It is the official language of northern Nigeria (around 30 million speakers) and a national language of Niger (around 9 million speakers) but is also spoken in Ghana, Benin, Cameroon, Togo, Chad and Burkina Faso (Koslow, 1995). Hausa is considered to be a common language of West Africa and Central Africa: it is spoken in many large commercial cities such as Dakar, Abidjan, Lome, Ouagadougou or Bamako. Hausa is a tonal language. About a quarter of the Hausa words comes from the Arabic language but Hausa was also influenced by French. Hausa can be written with the Arabic spelling since the beginning of the 17th century. This writing system is called *'Ajami*. However, the official spelling is based on the Latin alphabet called *Boko*. There are different varieties of Hausa, depending on whether it is spoken in eastern (i.e: Kano, Zaria, Bauchi, Daura), western (i.e: Sokoto, Gobir, Tahoua), northern (i.e: Katsina, Maradi, Zinder) or southern (i.e: Zaria, Bauci) areas.

Enlarging the subregion coverage, we also consider Bambara, Wolof, and Fula languages, to cover major West African languages. All the languages targeted cover more than half of the 300 million people of West Africa.

Bambara (or Bamanankan) is largely spoken in West Africa by around 40 million people. It is mainly spoken in Mali with 4 million speakers (census from the SIL[3] in 2012) and is used as a lingua franca. The Bambara is an agglutinative language. It uses the Roman script written system and is a tonal language.

Wolof belongs to the Atlantic languages which is part of the

[1] see `http://alffa.imag.fr`

[2] see `https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR`

[3] see `https://www.ethnologue.com/`

Niger-Congo phylum. This language is spoken in Senegal, Gambia and Mauritania where it is considered as a national language. In addition, Wolof is considered as one of the common languages and is spoken by 10 million people in total. More details on the language are given in the section 3.1..

Fula is a set of dialects spoken in all West Africa countries by 70 million people. They include both tonal and non-tonal languages. In the ALFFA project, we have decided to focus on Pulaar, the western dialect of the fula languages. Like the Wolof, the Pulaar belongs to the Atlantic languages which is part of the Niger-Congo phylum. The dialect is mainly spoken in Senegal by almost 3.5 million speakers (census in 2015 from the SIL) but also in dozens of other African countries. The Pulaar dialect use the Latin-based orthography.

As far as East Africa is concerned, we designed ASR system for Swahili which is the most widespread language in the East of the continent: spoken by more than 100 million people[4]. Swahili (or Kiswahili) belongs to the Niger-Congo phylum. It is the most spoken language among the Bantu language: it is used as a first language but also as a common language by a large population. In Tanzania, the language has an official status but Swahili is a national language in numerous East and Central Africa. For centuries the Arabic script has been used to written the Swahili but currently the standard written system is based on Latin script.

We also work on Amharic, a Semitic language which is part of the Afro-asiatic phylum. It is the second most spoken language among the Semitic languages. It is spoken mostly in Ethiopia by 22 million speakers (census from the SIL in 2010) where the language has an official status. Amharic uses an alphasyllabary written system named fidel : the consonant-vowel sequence represents a discrete unit (Comrie, 2009).

2.2. Automatic Speech Recognition for 3 languages

ASR systems for Swahili, Hausa and Amharic have been built so far. All the data and scripts to build a complete ASR system for these 3 languages are already available to the public on the github repository already mentioned in section 1. We used Kaldi speech recognition toolkit (Povey et al., 2011) for building our ASR systems. For the Swahili and Amharic ASR systems, the transcribed speech corpora, pronunciation lexicons and language models (LMs) are also made available while for Hausa ASR, users need to buy the corpus and the lexicon at ELDA first.

The ASR for Swahili was trained on about 10 hours of speech and the evaluation was done on about 1.8 hours. The language model was trained on text data grabbed from online newspaper (about 28M words) and cleaned as much as it could. More details on the Swahili corpus and how it was collected can be found on (Gelas et al., 2010).

For Hausa, the GlobalPhone Speech Corpus (Schlippe et al., 2012) was used. About 7 hours of data was used to train the system and 1 hour to evaluate it. The language model is composed of transcribed speech data from Globalphone corpus (41k words) and was converted into lower case. Finally, the lexicon contains 42,662 entries.

The Amharic system was retrained from the corpus described in (Tachbelie et al., 2014) and represents about 20 hours of speech data, while the testing set represents about 2 hours. Concerning the language model, it was created using SRILM using 3-grams and the text is segmented in morphemes using Morfessor 2.0 (Creutz and Lagus, 2002). A summary of the ASR performance (measured with Word Error rate (WER)) obtained for the three languages is given on table 1 but more experimental details can be found in the README files of the Github repository.

Table 1: *ASR performance for Swahili, Hausa and Amharic - HMM/SGMM acoustic modeling - all scripts available on github to reproduce experiments.*

Task	**WER (%)**
Swahili broadcast news	20.7
Hausa read speech	10.0
Amharic read speech	8.7

3. Collecting text in Wolof

3.1. Our focus: Wolof language

As we said in 2.1., Wolof is mainly spoken in Senegal but also in Gambia and Mauritania. Even if people who speaks Wolof understand each other, the Senegalese Wolof and the Gambian Wolof are two distincts languages: both own their ISO 639-3 language code (respectively ”WOL” and ”WOF”). For our studies, we decided to focus on Senegalese Wolof, and more precisely on the urban Wolof spoken in Dakar.

About 90% of the Senegalese people speak Wolof, while almost 40% of the population uses it as their mother tongue[5]. Wolof is originally used as an oral language. The writing system was developed at a later time, in Arabic orthography named *Wolofal* first (back to the pre-colonial period, through Islam spreading) and then in Latin orthography (through pre-colonial period). Nowadays, the latter is officially in use. 29 Roman-based characters are used from the Latin script and most of them are involved in digraphs standing for geminate and prenasalized stops. Wolof is a non-tonal language, with no diphthong and a moderate syllabic complexity.

The official language in Senegal is French. By definition, it is the language used in all the Institutions of the government like administrations and all administrative papers, courts, schools, etc. Consequently, Wolof is not learned at school. The orthography has no real standard rules and can be written in many ways. We will see later in this paper that it causes some problems in ASR. Nonetheless, the Center of Applied Linguistics of Dakar (CLAD)[6], coordinates the orthographic standardization of the Wolof language.

Finally, very few electronic documents are available for this language. We started from a first set of initial documents gathered as part of (Nouguier Voisin, 2002) and then tried to collect Wolof text from the Web.

[4] see `http://swahililanguage.stanford.edu`

[5] see `http://www.axl.cefan.ulaval.ca/afrique/senegal.htm`

[6] `http://clad.ucad.sn`

3.2. Initial documents available

First of all, in order to build a textual corpus in Wolof, we used some texts collected for the purpose of (Nouguier Voisin, 2002). It gathers proverbs (Becker et al., 2000), stories from (Kesteloot and Dieng, 1989), transcripts of debates about healers, a song entitled "Baay de Ouza" and two dictionaries: "Dictionnaire wolof-français" written by Aram Fal, Rosine Santos, Jean-Léonce Doneux (Fal et al., 1990) and "Dictionnaire wolof-français et français-wolof" by Jean Léopold Diouf (Diouf, 2003). These files were in different formats such as PDF, MS Word/Excel, HTML, etc.

We extracted all these documents to TXT format. We post-processed them by converting text into lower case and by cleaning them from non-wolof data (like section numbering, numbered list, french notes, etc.) and punctuation. Overall, it represents an overall usable text corpus of 20,162 utterances (147,801 words).

3.3. Retrieving data from the Web

148k words is small for statistical language modeling, so we decided to collect more text data in Wolof using the Web. Very few documents written in Wolof are actually available. In addition, we looked for well-structured data (in accordance with syntactic rules). For this purpose, we found some PDF files from educational, religious and news websites. We have extracted contents from the Universal Declaration of Human Rights, the Bible and a book written by an humanist in TXT format. In term of post-processing, we removed symbols, punctuation characters and non meaningful text (like section numbering, numbered list, etc.) from the collected texts and converted characters to lower case. In total, we got 197,430 additional words. Also, given the limited data manually found, we decided to crawl the Wikipedia database to collect a larger amount of data in Wolof. We retrieved all the articles indexed in the Wolof language using Wikipedia Extractor (Attardi and Fuschetto, 2013). As this kind of open database is only lightly supervised, some articles can be multilingual. To remove non Wolof text, we applied the Google Compact Language Detector (CLD2)[7]. As CLD2 cannot recognize Wolof but can detect the most widely used languages, we used the tool to filter out the non Wolof languages detected (and hypothesized that the remaining documents were in Wolof). To improve the precision on Wolof text retrieval, we also applied a data selection tool called Xenc (Rousseau, 2013). After these two filtering passes and quick manual cleaning, we obtained an additional collection of about 311k words.

Table 2 summarizes the data finally retrieved from the Web.

Table 2: *Additional Wolof text retrieved from the Web.*

Text	#utterances	#tokens
Universal Declaration of Human Rights	112	1,923
Silo's Message	602	10,443
The Bible	14,474	185,064
Wikipedia	10,738	311,995
Total	**25,926**	**509,425**

4. Audio corpus in Wolof

From our initial textual corpus (developed in section 3.2.), we randomly extracted a corpus of 6,000 utterances with length between 6 and 12 words. Then, from this small homogeneous corpus we extracted 18 sub-corpora of 1,000 utterances that will be used as recording sessions. We took advantage of the TTS recording campaign in Dakar (Senegal) of our project partner Voxygen to collect our read speech corpus. They recorded for us 18 natives Wolof speakers (10 male, 8 female) from different socio-professional categories (journalist, student, manager, teacher, teleoperator) and from 24 to 48 years old, using a Samson G-track microphone in a clean environment. A sociolinguistic questionnaire was also collected for each speaker.

The 18,000 recorded utterances represent 21h22mn of signal. We chose 14 speakers for the traning set, 2 for the development set and 2 for the testing set. We checked that each set was composed of an equivalent quantity of literary genre.

Training / development / testing partition is given in table 3.

Table 3: *Wolof speech corpus overview.*

Set	Male	Female	#utterances	#tokens	Duration
Training	8	6	13,998	132,963	16 h 49 mins
Development	1	1	2,000	18,790	2 h 12 mins
Testing	1	1	2,000	18,843	2 h 20 mins
Total	**10**	**8**	**17,998**	**170,596**	**21 h 21 mins**

5. First ASR system for Wolof

We built two language models. The first 3-gram model we trained with SRILM toolkit (Stolcke and others, 2002) was trained on 106,206 words (11,065 unigrams). These training data have been generated from the initial text corpus (20,162 utterances represented by 147,801 words) from which we removed utterances used for the speech recordings (41,595 words removed). The perplexity of the language model is 294 on the dev set (7.1% of OOVs) and 301 on the test set (7.2% of OOVs). This model will be called LM1 in the next sections. The second 3-gram language model trained with SRILM toolkit is an interpolation between the first one and another built from the data collected on the Web (509k) cleaned, mentioned in 3.3.. It finally corresponds to 601,609 words (29,148 unique words). Its perplexity is 314 (5.4% of OOVs) on the dev set and 323 (5.1% of OOVs) on the test set. This model will be called LM2 in the next sections. The table 4 summarizes the language models built so far.

Table 4: *Summary table of the 2 language models built so far.*

Language model	#Words of the textual corpus	Ngram perplexity		Out of vocabulary words (%)	
		dev	test	dev	test
LM1	∼106k	294	301	7.1	7.2
LM2	∼600k	314	323	5.4	5.1

Regarding the pronunciation dictionary, we used a seed of 8,724 entries corresponding to the concatenation of the

[7] see https://github.com/CLD2Owners/cld2

entries phonetically transcribed in (Fal et al., 1990) and (Diouf, 2003). From this, we trained a 7-gram pronunciation model for Wolof using Phonetisaurus (Novak, 2011), a Grapheme-to-Phoneme (G2P) conversion system, which allows us to automatically transcribe into phonetic symbols the remaining vocabulary of the LMs not phonetized yet. The vocabulary that covers LM1 language model is made of 15,575 entries (including 302 variants) while the one for LM2 has 32,039 entries. Each pronunciation dictionary was used both for training and decoding stages.

We used Kaldi speech recognition toolkit for building our ASR systems (and consequently, acoustic models). Three systems based on different acoustic modeling techniques were built: one based on the classical hidden Markov model and Gaussian mixture model (HMM/GMM) approach, one based on the subspace Gaussian mixture model (SGMM) approach and another one using deep neural networks (DNNs).

For the HMM/GMM system, the acoustic models were built using 13 Mel-frequency cepstrum coefficients (MFCCs) and Gaussian mixture models on 16.8h training data. We trained triphone models by employing 3,401 context-dependent states and 40k Gaussians. Besides that, we implemented delta delta coefficients on the MFCCs, linear discriminant analysis (LDA) transformation and maximum likelihood transform (MLLT) (Gopinath, 1998), as well as speaker adaptation based on feature-space maximum likelihood linear regression (fMLLR) (Gales, 1998).

To build DNNs, we trained the network using state-level minimum Bayes risk (Kingsbury, 2009) (sMBR) and the network had seven layers, each of the six hidden layers had 1024 hidden units. The network was trained from 11 consecutive frames (5 preceding and 5 following frames) of the same MFCCs as in the GMM systems. Furthermore, same HMM states were used as targets of the DNN. The initial weights for the network were obtained using Restricted Boltzmann Machines (RBMs) (Hinton, 2010) then fine tuning was done using Stochastic Gradient Descent.

Table 5: *Wolof ASR systems performance for different AMs and LMs - with speaker adaptation.*

	WER (%)			
Acoustic model	**LM1 (∼106k)**		**LM2 (∼600k)**	
	dev	test	dev	test
HMM/GMM	33.51	37.95	31.70	35.97
no diacritic	*31.60*	*36.20*	*29.70*	*34.10*
SGMM+MMI	30.37	35.24	28.56	33.56
no diacritic	*28.40*	*33.50*	*26.60*	*31.70*
DNN+sMBR	29.10	35.45	**27.21**	33.63
no diacritic	*27.20*	*33.60*	*25.10*	*31.70*

We can see in the table 5 the performance for the first Wolof ASR system trained using HMM/GMM, SGMM and DNN approaches. These baseline performances show that our first Wolof ASR system can reach WER around 30%. Since we have only two speakers per evaluation set, the standard deviation is large. Indeed, we observed until 6% of WER difference between speakers, on both dev and test sets. Also, results are close between language models. Our hypothesis is that LM2 does not provide much better performances because it comes for the most part from the Web, thus the vocabulary and syntax are much different from the ones of the speech corpus (similar to LM1). We observe, however, a small improvement of the performance with LM2 compared to LM1. We also analysed the outputs of the ASR systems and found that many errors are due to normalization issues in the text. Wolof is a morphologically complex language and some errors can appear at this stage. In addition, as we said in 3.1., a same word can have several surface forms, especially when it contains diacritics. For example, the word ”jél” can also be written ”jël” and mean ”took” or ”steal”. Also, the word ”randal” can be written ”ràndal” and mean the same: ”keep away”, but can also be spelled ”dandal”. About this problem of orthography, the word ”céetal” for example can be spelled ”sétal” and means ”organise the wedding”. Concerning diacritics, we also evaluate WER by removing all of them on both hypothesis and references. The results are also provided in table 5. WERs are slightly lower in that case and the difference between both numbers (WER with/w-o diacritics) corresponds to diacritic errors. Moreover, we can observe that DNNs bring some improvements on the dev set but not on the test set in comparison with the SGMMs.

6. Conclusion

This paper presented the data collected and ASR systems developped for 4 sub-saharan african languages (Swahili, Hausa, Amharic and Wolof). All data and scripts are available online on our github[8] repository. More precisely, we focus on Wolof language by explaining our text and speech collection methodology. We trained two language models: one from some data we already owned and another one with the addition of data crawled from the Web. Finally, we present the first ASR system ever built in this language. The system which obtains the best score is the one using the LM2 and the DNNs, for which we got 27.21% of WER.

Perspectives. In the short run, we intend to improve the quality of the LM2 by using neural networks. We also currently work on a duration model for the Wolof and the Hausa ASR systems.

In the medium term, we want to deal with text normalization. As we have seen, words can have several surface forms and can be written in many ways. By selecting one among a number of several possible variants, we want to normalize the orthography of our corpus. We expect that this desambigation will allow to reduce the number of OOVs and improve the quality of our language model.

Last but not least, we continually develop and improve Lig-Aikuma (Blachon et al., 2016), a fork of the Aikuma application (Bird et al., 2014). Lig-Aikuma is destinated to field linguists and is designed to make the data collection work easier and faster. Through the use of this application, we will check, by an expert of the Wolof, our data collected (right correspondence between the audio file and its transcription) to assess our system on more reliable data.

[8] see `https://github.com/besacier/ALFFA_PUBLIC/tree/master/ASR`

7. Bibliography

Attardi, G. and Fuschetto, A. (2013). Wikipedia extractor. *Medialab, University of Pisa*.

Becker, C., Martin, V., and Mbodj, M. (2000). Proverbes et énigmes wolof cités dans le dictionnaire volof-français de mgr kobès et du rp abiven.

Bird, S., Hanke, F. R., Adams, O., and Lee, H. (2014). Aikuma: A mobile app for collaborative language documentation. *ACL 2014*, page 1.

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., and Rialland, A. (2016). Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. *SLTU*. Submitted.

Comrie, B. (2009). *The world's major languages*. Routledge.

Creutz, M. and Lagus, K. (2002). Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 workshop on Morphological and phonological learning-Volume 6*, pages 21–30. Association for Computational Linguistics.

Diouf, J. L. (2003). *Dictionnaire wolof-français et français-wolof*. KARTHALA Editions.

Fal, A., Santos, R., and Doneux, J. L. (1990). *Dictionnaire wolof-français: suivi d'un index français-wolof*. Karthala.

Gales, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. In *Computer Science and Language*, volume 12, pages 75–98.

Gelas, H., Besacier, L., Rossato, S., and Pellegrino, F. (2010). Using automatic speech recognition for phonological purposes: study of vowel length in punu (bantu b40).

Gopinath, R. A. (1998). Maximum likelihood modeling with gaussian distributions for classification. In *Proc. of ICASSP*, pages 661–664.

Hinton, G. E. (2010). A practical guide to training restricted boltzmann machines. Utml tr 2010-003, Dept. Computer Science, University of Toronto.

Kesteloot, L. and Dieng, B. (1989). *Du tieddo au talibé*, volume 2. Editions Présence Africaine.

Kingsbury, B. (2009). Lattice-based optimization of sequence classification criteria for neural network acoustic modeling. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3761–3764, April.

Koslow, P. (1995). *Hausaland: the fortress kingdoms*. Chelsea House Pub.

Nouguier Voisin, S. (2002). *Relations entre fonctions syntaxiques et fonctions sémantiques en wolof*. Ph.D. thesis, Lyon 2.

Novak, J. R. (2011). Phonetisaurus: A wfst-driven phoneticizer. *The University of Tokyo, Tokyo Institute of Technology*, pages 221–222.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit.

Rousseau, A. (2013). Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100:73–82.

Schlippe, T., Djomgang, E. G. K., Vu, N. T., Ochs, S., and Schultz, T. (2012). Hausa large vocabulary continuous speech recognition. In *SLTU*, pages 11–14.

Stolcke, A. et al. (2002). Srilm-an extensible language modeling toolkit. In *INTERSPEECH*.

Tachbelie, M. Y., Abate, S. T., and Besacier, L. (2014). Using different acoustic, lexical and language modeling units for asr of an under-resourced language–amharic. *Speech Communication*, 56:181–194.

Vycichl, W. (1990). Les langues tchadiques et l'origine chamitique de leur vocabulaire. In *Relations interethniques et culture matérielle dans le bassin du lac Tchad: actes du IIIème Colloque MEGA-TCHAD, Paris, ORSTOM, 11-12 septembre 1986*, page 33. IRD Editions.