# DEVELOPMENTS OF SWAHILI RESOURCES FOR AN AUTOMATIC SPEECH RECOGNITION SYSTEM

*Hadrien Gelas[1,2], Laurent Besacier[2], François Pellegrino[1]*

[1]Laboratoire Dynamique Du Langage, CNRS - Université de Lyon, France
[2]Laboratoire Informatique de Grenoble, CNRS - Université Joseph Fourier Grenoble 1, France
{hadrien.gelas, francois.pellegrino}@univ-lyon2.fr, laurent.besacier@imag.fr

## ABSTRACT

This article describes our efforts to provide ASR resources for Swahili, a Bantu language spoken in a wide area of East Africa. We start with an introduction on the language situation, both at linguistic and digital level. Then, we report the selected strategies to develop a text corpus, a pronunciation dictionary and a speech corpus for this under-resourced language. We explore methodologies as crowdsourcing or collaborative transcription process. Besides, we take advantage of some linguistic characteristics of the language such as rich morphology or shared vocabulary with English to improve performance of our baseline Swahili ASR system in a broadcast speech transcription task.

***Index Terms***— Swahili, under-resourced languages, automatic speech recognition, speech resources

## 1. INTRODUCTION

Due to world's globalization and answering the necessity of bridging the numerical gap with the developing world, speech technology for under-resourced languages is a challenging issue. Applications and usability of such tools in developing countries are proved to be numerous and are highlighted for information access in Sub-Saharan Africa [1, 2], agricultural information in rural India [3], or health information access by community health workers in Pakistan [4]. However, Human Language Technologies still face the lack of numerical resources and thus barely reflect the world language diversity (over 6000 languages [5]). As many other world languages, African languages highly suffer from this. Answering this issue, there is a growing research interest towards speech and language processing for under-resourced and more specifically African languages. Specific workshops in this domain recently appeared, such as SLTU conferences (Spoken Languages Technologies for Under-resourced languages), AfLaT (African Language Technology[1]) and the recent special session "Speech Technology for Under-Resourced Languages" at Interspeech 2011.

[1]http://aflat.org/

In this contribution, we report recent development work on a broadband automatic speech recognition (ASR) system for Swahili. In the next section we present some insight into the language linguistic background and situation. Section 3 addresses data collection, while section 4 focuses on the resulting ASR system and experimental results. Finally, section 5 concludes and discusses future work.

## 2. BACKGROUND: SWAHILI

### 2.1. Swahili language

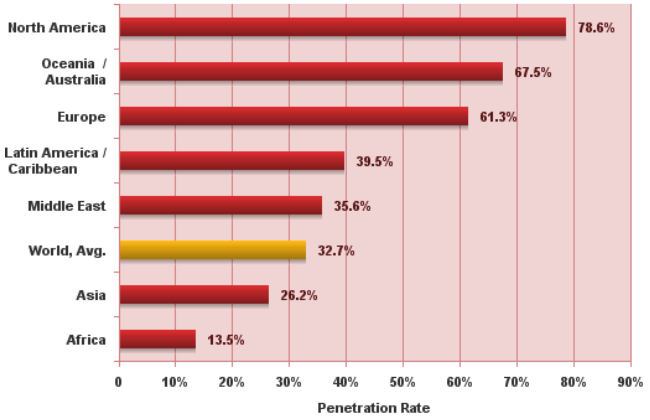
Swahili is a widely spoken language often used as a vehicular language in a large area of East Africa [6]. It is the national language of Kenya and Tanzania and is broadly used in public administration, education and the media. It is also spoken in different parts of Democratic Republic of Congo, Mozambique, Somalia, Uganda, Rwanda and Burundi. Most estimations give between 40 and 100 million speakers (with only less than 5 million native speakers). Several Swahili dialects are spoken nowadays and this study only focuses on the major variety, the so-called standard Swahili (Kiswahili sanifu).

Swahili is a member of the large Bantu family that spreads over an important part of Africa and is more specifically part of group G of Guthrie's referential classification [7]. Structurally, Swahili is often considered as an agglutinative language [8] and has typical Bantu features, such as noun class, agreement systems and complex verbal morphology. However, it distinguishes itself from many other Bantu languages by the absence of tone and also an important share of vocabulary with an arabic origin. It was first written with an arabic-based orthography before it adopted the Roman script (standardized since 1930 [9]).

### 2.2. Digital review of Swahili

If Africa is the continent with the lowest internet penetration rate of 13.5% (while the world average is 32.7%, see figure 1), it still represents 6.2% of all the Internet users in the world (over 139 million people) and the continent had a significant growth of users of 2988.4% between 2000 and

**Fig. 1.** *World Internet penetration rates by geographic regions - 2011*




Source: Internet World Stats - www.internetworldststs.com/stats.htm
Penetration Rates are based on a world population of 6,930,055,154 and 2,267,233,742 estimated Internet users on December 31, 2011.
Copyright © 2012, Miniwatts Marketing Group

2011[10]. Moreover, Swahili being an impactful vehicular language of East Africa explains why many of mainstream IT services are already proposing a localization in this language. Among others:

- Microsoft launched Swahili version of Microsoft Office and Windows in 2005[2].
- Wikipedia reached 23k articles in December 2011 (80th on 283 languages) after a launch in 2003[3]. It is the first Bantu language and is second after Yoruba (30k articles) in the Niger-Congo family.
- Facebook Swahili version was launched in 2009[4] and was made by a group of scholars with the firm permission.
- Google also offers many of its services in Swahili[5]: Google search interface in 2004, Google Translate since 2009, Text to speech, Gmail, Google Chrome and Google Maps in 2010, but not yet Voice Search ASR.

Other initiatives for Swahili promotion over the web does exist. This includes the following websites: the `Kamusi project` ("the internet living Swahili dictionary") or the one-stop Swahili portal `goswahili.org` regrouping many resources on the language. It is also to be mentioned the Kiswahili Linux Localization Project (`klnX`) who made great efforts to localize free and open source software to the Swahili language.

Regarding Natural Language Processing researches on Swahili, previous works have been made on different language analyzers (morphology analyzers, parsers, POS-taggers, lemmatizers...). Some are rule-based approaches as in [11] when others follow data-driven approaches [12, 13, 14, 15]. Concerning Human Language Technologies, should be mentioned studies on Text-To-Speech [16], Machine Translation in [17] and [18] but also a first dictation system in [19].

## 3. RESOURCES

### 3.1. Text collection / Text corpus

A text corpus is known to be essential for language modeling and thus ASR. Previous attempts have been made concerning text collection in Swahili. The Helsinki Corpus of Swahili [20] is available for research and contains more than 12M words from news texts and extracts of books. In [17], a 2M word parallel text corpus english-swahili was built for machine translation. Finally in [21], a 5M words corpus was collected using bigrams search, and where only Swahili web pages were filtered among other languages using an identification score based on unigrams from preselected text in the respective languages.

As it has been said in section 2, Swahili benefits from a good visibility on the web. Thus, as many news websites can be found, we decided to build our own corpus based on 16 of them preselected to be strictly monolingual (avoiding a multilingual filtering step). As in [22], we downloaded all news pages under different format and then applied a classical text extraction, cleaning and filtering process as follows :

- Text extraction
- Sentences identification and segmentation using punctuations and some specific html tags
- Removal of all html tags and irrelevant texts
- Cleaning of bad/different encoding characters to UTF8 encoding
- Removal of different brackets and text in between
- Digit conversion and replacement of common abbreviations

We collected through this process over 28M words (tokens). As already seen above, Swahili is an agglutinative language with a rich morphology. In a morphological template for Swahili verbs, ten positions can be identified [8]. If not all can be filled at the same time, it is common to found six or seven positions filled as in example 1[6].

Verbs can also be modified by verbal extensions common in Bantu languages, as among others, the affixes for the passive voice (-w-) or the causative (-ish-, -esh-). Such

[2]http://news.bbc.co.uk/2/hi/africa/4527876.stm
[3]http://stats.wikimedia.org/EN/TablesWikipediaSW.htm
[4]http://news.bbc.co.uk/2/hi/8100295.stm
[5]http://googleblog.blogspot.com/

[6]NEG= Negation, SM2= Subject marker of noun class 2 (one of the 16 different noun classes, it is common in Bantu linguistics to name noun classes according to a numerical system), FUT= Future tense, OM2= Object marker of class 2, *tell*= verbal root, FIN= Final vowel, PL= Post final plural

**Example 1.** *Morphological segmentation of a Swahili verb*

| | |
|---|---|
| Word | hawatakuambieni |
| Segmentation | ha-wa-ta-ku-ambi-e-ni |
| Glosis | NEG-SM2-FUT-OM2-*tell*-FIN-PL |
| Translation | they will not tell you (plural) |

characteristic involves an important lexical variety and for ASR, it results in data sparsity and in a much worse lexical coverage than state of the art speech recognition set-up (as one for English). Resulting high Out-Of-Vocabulary (OOV) rates have obvious consequences on Word-Error rate (WER) as each OOV word will not be recognized but can also affect their surrounding words and strongly increase WER. Many research tried to tackle morphologically rich languages in NLP [23]. Concerning ASR, the frequent answer to reach a larger lexical coverage is to segment words in sub-word units as in [24] for Amharic. In [25] is presented a recent overview of different studies on morph-based language modeling for speech recognition. After investigating different sub-units for Swahili (experiments not reported here), we decided to use the morpheme obtained with an unsupervised approach. For this, we used a publicly available tool called Morfessor [26]. Its data-driven approach learns a subword lexicon from a training corpus of words using a Minimum Description Length algorithm.

Table 1 gives type OOV rates calculated on reference transcriptions and depending on segmentation level and different vocabulary size. The segmentation in morphemes allows to reach a better lexical coverage while keeping the same size of lexicon. Due to decoder limitations, we restrained this study to 65k lexicon, but for a 200k word vocabulary we get a type OOV rate of 12.46% and with 400k words (full vocab) it is still 10.28%. In the same time, growing Morfessor lexicon to 200k would be more advantageous as it reduces the type OOV rate to 1.61%.

**Table 1.** *Type OOV rates depending on different type of text segmentation and vocabulary size*

| LM | Type OOV (%) |
|---|---|
| Word-65k | 19.17 |
| Word-200k | 12.46 |
| Word-400k | 10.28 |
| Morf-65k | **11.36** |
| Morf-200k | 1.61 |

### 3.2. Pronunciation dictionary

Pronunciation dictionary is of primary importance in acoustic modeling. To generate one, we extracted from the news text corpus the 65k most frequent words. The following step is to provide a pronunciation for each lexical entry using a limited set of phones, the basic unit of acoustic models. Swahili spelling is really close to its pronunciation and for each distinct phoneme, the same distinct written form is used. Only two written forms exist for a phoneme in Swahili : either a single letter or a digraph. Therefore, a grapheme-to-phoneme script taking benefits of this regularity automatically generates most of the words' pronunciation. All the phonemes (basic units of the language) of Swahili are here considered as phones (basic units of the acoustic model), further investigations are needed to decide if rarest sounds could be avoided and thus improve or not the acoustic model. Indeed, for example, sounds like [θ], [ð], [x] and [ɣ] (respectively transcribed th, dh, kh and gh) are quite rare since they only appear in some of the words with Arabic origin and may be pronounced by Swahili speakers, respectively [s], [z], [h], [r]. The table 2 present all graphemes, phonemes and their corresponding phones in the ASR system. Among the different published grammars and linguistic studies, prenasal consonants status as a phoneme is subject of controversy among researchers [27]. As others, we decide to keep them as distinct phonemes and also distinct phones in the acoustic model since they may be distinct phonetic realization with their standalone counterparts [28] ([$^{m}$b] is distinct of a [m] followed by a [b]). Again, additional explorations on phone selection are required.

A remaining issue is the generation of pronunciation for English words, proper names and acronyms, which appear frequently in the corpus. Most English words and proper names are pronounced in news as they are in English. If those words are too rare to add a specific English phone set for them in the acoustic model, they are also too frequent to leave them with their erroneous grapheme-to-phoneme Swahili pronunciation (as in [29] with Mandarin or [30] with Cantonese). In the 65k Swahili lexicon, 8.77% of the words are found in the publicly available CMU English dictionary [31]. They are highly supposed to be English words or proper names. Words in the 500 Swahili most frequent word list were removed as they are assumed to be short Swahili functional or grammatical words. Therefore, to handle this, when one word was common to CMU dictionary and our 65k Swahili vocabulary, the CMU pronunciation was added as a variant in the Swahili dictionary using a theoretical mapping of CMU English phones to Swahili phones. We preferred to keep also the original grapheme-to-phoneme pronunciation and let the system decide, since it still may be in some cases closer than the English pronunciation. The pronunciation dictionary with variants included was only used while decoding the test set and not during the training phase. Example 2 shows extracts of English words in our pronunciation dictionary.

Considering acronym pronunciations, it is often pronounced as if it was spelled. Therefore, to generate closer transcriptions, a script was aiming to detect short entry containing clusters of letters that are not allowed by Swahili phonotactic. For those entries, a variant with the spelling pronunciation was added (ex. TFF became in our dictionary

**Table 2.** *Phonemes, graphemes and ASR Phones for Swahili*

| Phoneme | Grapheme | Phone | Phoneme | Grapheme | Phone | Phoneme | Grapheme | Phone |
|---|---|---|---|---|---|---|---|---|
| /a/ | a | a | /b/ | b | b | /ɣ/ | gh | RR |
| /e/ | e | e | /d/ | d | d | /f/ | f | f |
| /i/ | i | i | /ɟ/ | j | j | /θ/ | th | TT |
| /o/ | o | o | /g/ | g | g | /s/ | s | s |
| /u/ | u | u | /p/ | p | p | /ʃ/ | sh | SS |
| /m/ | m | m | /t/ | t | t | /x/ | kh | XX |
| /n/ | n | n | /c/ | ch | CC | /h/ | h | h |
| /ɲ/ | ny | YY | /k/ | k | k | /r/ | r | r |
| /ŋ/ | ng' | NN | /$^{m}$v/ | mv | VV | /l/ | l | l |
| /$^{m}$b/ | mb | BB | /$^{n}$z/ | nz | ZZ | /j/ | y | y |
| /$^{n}$d/ | nd | DD | /v/ | v | v | /w/ | w | w |
| /$^{ɲ}$ɟ/ | nj | JJ | /ð/ | dh | LL | | | |
| /$^{ŋ}$g/ | ng | GG | /z/ | z | z | | | |

**Example 2.** *Extracts of English words and their variant imported from CMU English dictionary in our pronunciation dictionary*

```
...                 ...
corporation         k o r p o r a t i o n
corporation(2)      k o r p e r e y SS e n
...                 ...
games               g a m e s
games(2)            g e y m z
...                 ...
ukraine             u k r a i n e
ukraine(2)          y u k r e y n
...                 ...
```

"t i e f e f").

### 3.3. Audio collection / Speech data

Any research in speech recognition requires audio data and matching transcriptions in order to build the necessary acoustic models. However, in an under-resourced language situation, it is expected that one will not have access to speech transcriptions and thus being a major constraint while it represents both a time consuming and an expensive task [32]. Different studies conduct methodologies to accelerate such corpora development as in [33] and [34].

In our case, we first started to collect a read speech audio corpus where the sentences read by speakers were directly providing our corresponding transcriptions. In our protocol, texts were first extracted from news websites and then segmented into sentences. Recordings were made by native speakers reading sentence by sentence with the possibility to re-record any time they considered having mispronounced. We retained from these first steps of development, a set of 3 hours and a half read by 5 speakers (3 males and 2 females).

To have a more substantial corpus, we also collected web broadcast news which have the clear advantage to be massively and directly available. The main issue one can found by mining web broadcast news speech is concerning the audio quality which is often low. However, we managed to collect more than 200 hours with a 64kbps bitrate which was considered as good enough for ASR acoustic models in [35] and [36]. Each radio show was containing both music and speech segments which may also differ by its audio quality: studio (high quality), telephone (low quality but without noise) or noisy (really bad quality, mostly outdoors or with ambient noise, speech or music in background).

In order to quickly provide transcriptions to this audio corpus, we investigated the use of web crowdsourcing tools as Amazon's Mechanical Turk (MTurk). MTurk is an on-line market place for work where one can submit simple tasks to human willing workers around the web. It has one major benefit: repetitive, time consuming and costly tasks can be completed quickly for low payment. It has been explored through many recent studies as a powerful tool for NLP tasks [37]. It has also a great potential to reduce their cost with good enough quality as in [38]. But Mturk is still a subject of controversy among researchers for some legal and ethical issues (mentionned in [39] and [40]), explaining why we evaluated first the possibility to use it on the small read speech corpus. The sentences read by speakers during recordings were used as our gold standards to compare with the transcriptions obtained by MTurk. The resulting acoustic model trained using MTurk transcriptions were quite similar to the one trained using our reference transcriptions. Respectively 38.5% and 38.0% WER on a small 82 sentences test set (more details in [39] where the task is also applied to Amharic). The task to transcribe three hours and a half of read speech completed in 12 days for Swahili by three MTurk workers. It is clearly a lower completion rate than for a language as English. For that reason and added to some potential legal/ethical issue, we

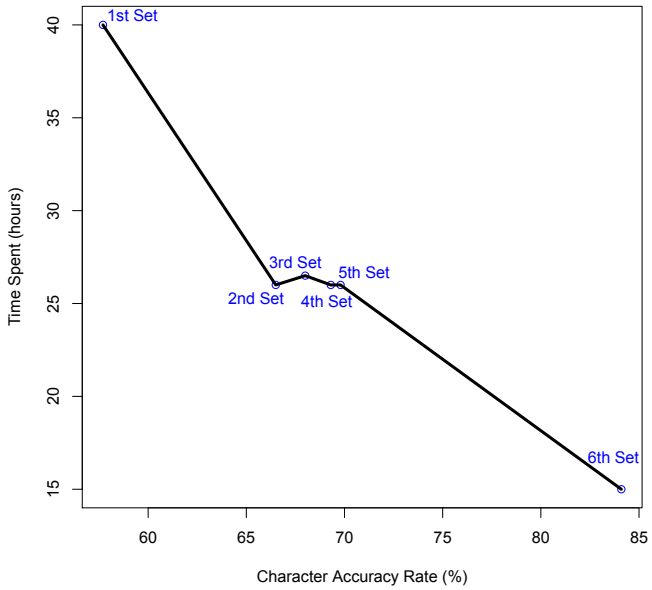
consequently decided to work directly with a Kenyan institute[7] to collaboratively transcribe 12 hours of our web broadcast news corpus.

Once more, in order to reduce the repetitive and time consuming transcription task, we considered a collaborative transcription process based on the iterative application of the following protocol:

- A first acoustic model is trained using read speech corpus.
- Each radio show were first segmented using standard automatic silence detection. Only files duration between 2 and 6 seconds were kept in order to pre-filter some music and too noisy segments. It also allows to considerably simplify the transcription task [41]. For a one hour show, we approximately keep 25mn of speech (where around 8% are rejected aposteriori by transcribers) and reject 25min of long segments (mainly bad quality speech, music, jingles...). Around 10mn of silence are also removed this way.
- A two-hour audio set (with the automatically segmented and pre-filtered speech) is then transcribed using our first ASR system.
- The two-hour pre-transcribed audio set of speech is sent to the Taji Institute for correction (post-edition).
- After being corrected by transcribers, annotated data were added to the training corpus and a new acoustic model was trained to transcribe the next two-hour audio set.

We repeated this procedure until 12 hours of transcribed audio were reached, keeping 10 hours for training and 2 hours as a test set. Figure 2 presents the different results at each iteration. As it appears, the time spent to correct transcriptions is correlated with the quality of transcriptions provided. In this figure the character accuracy rate is evaluated on the following audio set which is different each time (explaining the look of the plot between 2nd and 5th set). But results in Table 3 (when evaluated on a same test set) shows that each correctly transcribed set added to the training pool improves the acoustic model which provides better transcriptions for the next set. Thus, taking less time to correct. Using this protocol, the time for transcription task has been reduced from 40 hours to 15 hours.

**Fig. 2.** *Time spent (hours) to correct a 2 hours data set transcriptions depending on the character accuracy rate of the proposed corresponding transcriptions*




## 4. AUTOMATIC SPEECH RECOGNITION SYSTEM

### 4.1. System configuration

Once all resources described above collected, we used Sphinx-Train[8] toolkit from open source Sphinx project for developing Hidden Markov Models (HMMs) based acoustic models (AMs) for Swahili using our 37 predefined modeling units. The initial step was to extract features from overlapping frames of acoustic data. Each frame is encoded as 13 dimensional Mel Frequency Cepstral Coefficients (MFCCs) and has a window size of 25ms while the beginning of the frame is incremented of 10ms. Then, we used the acoustic features derived to train a standard 3-state context dependent (CD) model with 3000 tied states and 8 Gaussian mixtures. During the collaborative transcription process, only context independent (CI) models were trained until reaching 10 hours of training audio.

Concerning language models (LMs), both word and morpheme-based trigram LMs were built using the SRI[9] language model toolkit. Each language model is smoothed with modified Kneser-Ney smoothing technique and included a special unknown unit token.

### 4.2. System results

Different recognition experiments were conducted on a two-hour test set (1991 sentences) and results are presented in table 3. As expected during the collaborative transcription process, each two-hour audio set added to the training pool improved significantly (excepted between 4th and 5th set) the

[7] http://www.taji-institute.com/
[8] cmusphinx.sourceforge.net/
[9] www.speech.sri.com/projects/srilm/

WER. The visible shift from CI to CD acoustic models was our last step since our baseline model reach 35.8% WER.

It is also shown how adding variants in our pronunciation dictionary for English words and acronyms improved the performance in a clear acoustic environment (26.9% without and 26.5% with). Worst audio quality brings too much acoustic confusion to be strongly beneficial (35.8% to 35.7%).

Finally, in our sub-word units experiment, the ASR decoder output is a sequence of sub-word units, inducing the need to rebuild this output back to word level. Thus, a morpheme boundary tag is added on each side of segmentation. To rebuild up to word, we reconnect every time two morpheme boundaries are appearing consecutively (example, `kiMB MBtabu` becomes `kitabu`). The use of sub-word units for language modeling improves significantly WER in both poor and good acoustic environment (34.8% for all quality types and 25.9% with only studio audio quality). It can be explained by the improved lexical coverage. The morfessor-based 65k lexicon coverage represents 30.83% of the full vocabulary while word 65k lexicon represents only 13.95%. As shown in 3.1, it has a direct impact on OOV words. As a matter of fact, one other value of sub-word units for ASR is the OOV word recovery. Among the OOV words that can possibly be recognized, 36,04% were rebuilt.

**Table 3.** *Word Error Rates (WER) depending on different acoustic models (CI or CD), language models (word-based or Morfessor-based), dictionary (with or without variants) and quality types (all, lowband, noisy or studio)*

| ASR system | Quality type | Number sentences | WER (%) |
|---|---|---|---|
| 1st Set CI Word(65k) | All quality | 1991 | 72.8 |
| 2nd Set | All quality | 1991 | 59.0 |
| 3rd Set | All quality | 1991 | 57.4 |
| 4th Set | All quality | 1991 | 56.2 |
| 5th Set | All quality | 1991 | 56.1 |
| Baseline CD Word(65k) | All quality | 1991 | 35.8 |
| | Lowband | 424 | 60.0 |
| | Noisy | 402 | 36.4 |
| | Studio | 1165 | 26.9 |
| Baseline + Dict variants | All quality | 1991 | 35.7 |
| | Studio | 1165 | 26.5 |
| CD Morf(65k + variants) | All quality | 1165 | **34.8** |
| | Studio | 1165 | **25.9** |

## 5. SUMMARY AND PROSPECTS

In the present contributon, a set of newly-developed resources for Swahili ASR has been described. Different approaches to quicken the creation of a transcribed speech corpus have been explored. The powerful crowdsourcing tool which is Mturk has been tested to provide transcriptions on a small controlled corpus. Even if successful for Swahili, a collaborative transcription process with a Kenyan institute was found to be more rewarding. To help transcribers in their tasks a pre-transcription of a two-hour audio set was submitted to correction. Each audio set finally corrected was added to the training pool in order to re-train acoustic model and improve new proposed transcriptions. This protocol has been successful and the transcription task for a two-hour audio set went from 40h to 15h.

Special care has been provided to some linguistic singularities of Swahili but which can be extended to other languages with similar features. Concerning language modeling, the advantage of sub-unit improved performance for this morphologically rich language and can be performed without any specific knowledge through the application of unsupervised methods, for instance the publicly available tool Morfessor. From 35.7% WER for word-based model, we improved performance to 34.8% WER with the morfessor-based segmentation. A similar experiment has been performed on an Amharic read speech recognition task with also significant improvment [42].

Regarding the development of the pronunciation dictionary, the same attention was made to some language characteristics such as the strong presence of English words in the vocabulary. Variants of pronunciation were automatically added to the dictionnary taking advantage of already available materials as the CMU pronunciation dictionary for English. This process improves performance in a clear audio environment (studio quality) from 26.9% to 26.5%.

While we have promising results, a number of new lines of research will be followed. As it has been raised in Section 3.2, further investigations will be made on unit selection using different approaches. Also, taking benefit of the important number of mined web news (more than 200 hours of pre-segmented audio) is planned and since they are non-transcribed, different unsupervised or cross-sharing approaches could be taken into consideration. Finally, although it is important to develop language independent methodologies, some specific linguistic features seem difficult to avoid and others shared among many languages may greatly improve ASR performance. For example, rich morphology or tones are often considered in recent studies. But such characteristic like the strongly regular syllabic structure of Swahili could be interesting to exploit. Exploring syllabic acoustic models for Swahili and similar languages is planned.

## 6. ACKNOWLEDGEMENTS

This project was made possible through the support of the french ANR PI. The authors would like to thank Lutz Marten and Swahili speakers for their help concerning the read speech corpus. As well as Peter Waiganjo Wagacha, Anne Waiganjo

and people from the Taji Institute for their great work and help with transcriptions in this project.

## 7. REFERENCES

[1] Etienne Barnard, Marelie Davel, and Gerhard van Huyssteen, "Speech technology for information access: a South African case study," in *AAAI Symposium on Artificial Intelligence*, 2010, pp. 22–24.

[2] Etienne Barnard, Johan Schalkwyk, Charl van Heerden, and Pedro Moreno, "Voice search for development," in *Interspeech*, 2010.

[3] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T.S. Parikh, "Avaaj otalo: a field study of an interactive voice forum for small farmers in rural India," in *CHI*. ACM, 2010, pp. 733–742.

[4] A. Kumar, A. Tewari, S. Horrigan, M. Kam, F. Metze, and J. Canny, "Rethinking speech recognition on mobile devices," in *IUI4DR*. ACM, 2011.

[5] R.G. Gordon, B.F. Grimes, and Summer Institute of Linguistics, *Ethnologue: Languages of the world*, vol. 15, SIL International Dallas, TX, 2005.

[6] E.C. Polomé, *Swahili Language Handbook*, Center for Applied Linguistics, Washington, DC, 1967.

[7] M. Guthrie, *Comparative Bantu, (4 vols)*, Farnborough: Gregg International Publishers, 1967-1971.

[8] Lutz Marten, "Swahili," in *The Encyclopedia of Languages and Linguistics, 2nd ed.*, Keith Brown, Ed., vol. 12, pp. 304–308. Oxford: Elsevier, 2006.

[9] W. H. Whiteley, *The rise of a national language*, London: Methuen, 1969.

[10] "Internet world stats, http://www.internetworldstats.com/stats.htm," .

[11] A. Hurskainen, "Swahili language manager: a storehouse for developing multiple computational applications," *Nordic Journal of African Studies*, vol. 13, no. 3, pp. 363–397, 2004.

[12] G. De Pauw, G.M. De Schryver, and P. Wagacha, "Data-driven part-of-speech tagging of kiswahili," in *Text, speech and dialogue*. Springer, 2006, pp. 197–204.

[13] G. De Pauw and G.M. De Schryver, "African language technology: The data-driven perspective," *V. Lyding (eds.)*, pp. 79–96, 2009.

[14] G. De Pauw and G.M. de Schryver, "Improving the computational morphological analysis of a swahili corpus for lexicographic purposes," *Lexikos 18*, 2008.

[15] R. Shah, B. Lin, A. Gershman, and R. Frederking, "Synergy: a named entity recognition system for resource-scarce languages such as swahili using online machine translation," in *Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, 2010, pp. 21–26.

[16] K. Ngugi, W. Okelo-Odongo, and PW Wagacha, "Swahili text-to-speech system," *African Journal of Science and Technology*, vol. 6, no. 1, 2010.

[17] G. De Pauw, P.W. Wagacha, and G.M. De Schryver, "Exploring the sawa corpus: collection and deployment of a parallel corpus english - swahili," *Language resources and evaluation*, pp. 1–14, 2011.

[18] G. De Pauw, P.W. Wagacha, and G.M. de Schryver, "Towards english-swahili machine translation," in *Research Workshop of the Israel Science Foundation*, 2011.

[19] Evans Miriti, *A Kiswahili Dictation System: Implementation of a Prototype*, VDM Verlag Dr. Müller, 2010.

[20] A. Hurskainen, "Hcs 2004–helsinki corpus of swahili," *Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC*, 2004.

[21] K. Getao and Evans Miriti, "Automatic construction of a kiswahili corpus from the world wide web," *Measuring Computing Research Excellence and Vitality*, p. 209, 2006.

[22] V.B. Le, B. Bigi, L. Besacier, and E. Castelli, "Using the web for fast language model construction in minority languages," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[23] R. Sarikaya, K. Kirchhoff, T. Schultz, and D. Hakkani-Tur, "Introduction to the special issue on processing morphologically rich languages," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, 2009.

[24] T. Pellegrini and L. Lamel, "Automatic word decompounding for ASR in a morphologically rich language: Application to Amharic," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 5, pp. 863–873, 2009.

[25] T. Hirsimaki, J. Pylkkonen, and M. Kurimo, "Importance of high-order n-gram models in morph-based speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 724–732, 2009.

[26] M. Creutz and K. Lagus, "Unsupervised morpheme segmentation and morphology induction from text corpora

using morfessor 1.0," Tech. Rep., Computer and Information Science, Report A81, Helsinki University of Technology, 2005.

[27] M.A. Mohamed, *Modern Swahili Grammar*, Nairobi: East Afircan Educational Publishers, 2001.

[28] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*, Wiley-Blackwell, 1996.

[29] H.A. Chang, Y.H. Sung, B. Strope, and F. Beaufays, "Recognizing english queries in mandarin voice search," in *ICASSP*. IEEE, 2011.

[30] Y.H. Sung, M. Jansche, and P.J. Moreno, "Deploying google search by voice in cantonese," in *Interspeech*, 2011.

[31] R. Weide, "The carnegie mellon pronouncing dictionary [cmudict. 0.6]," 2005.

[32] E. Barnard, M. Davel, and C. Heerden, "Asr corpus design for resource-scarce languages," in *Interspeech*, 2009.

[33] M.H. Davel, C. van Heerden, N. Kleynhans, and E. Barnard, "Efficient harvesting of internet audio for resource-scarce asr," in *Interspeech*, 2011.

[34] Thad Hughes, Kaisuke Nakajima, Linne Ha, Atul Vasu, Pedro Moreno, and Mike LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *INTERSPEECH*, 2010.

[35] P. Mayorga, R. Lamy, and L. Besacier, "Recovering of packet loss for distributed speech recognition," in *Proc. Eusipco*. Citeseer, 2002.

[36] C. Barras, L. Lamel, and J.L. Gauvain, "Automatic transcription of compressed broadcast audio," in *ICASSP*. IEEE, 2001, vol. 1, pp. 265–268.

[37] G. Parent and M. Eskenazi, "Speaking to the crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges," in *Interspeech*, 2011.

[38] S. Novotney and C. Callison-Burch, "Cheap, fast and good enough: Automatic speech recognition with non-expert transcription," in *NAACL HLT*. Association for Computational Linguistics, 2010, pp. 207–215.

[39] H. Gelas, S.T. Abate, L. Besacier, and F. Pellegrino, "Evaluation of crowdsourcing transcriptions for african languages," in *HLTD*, 2011.

[40] G. Adda, B. Sagot, K. Fort, and J. Mariani, "Crowdsourcing for language resource development: Critical analysis of amazon mechanical turk overpowering use," in *LTC, 5th Language and Technology Conference*, 2011.

[41] B.C. Roy and D. Roy, "Fast transcription of unstructured audio recordings," in *Interspeech*, 2009.

[42] Martha Yifiru Tachbelie, Solomon Teferra Abate, Laurent Besacier, and Solange Rossato, "Syllable-based and hybrid acoustic models for amharic speech recognition," in *SLTU*, 2012.