Development of Isolated Numeric Speech Corpus for Swahili Language for Development of Automatic Speech Recognition System

Aaron M. Oirere
Department of CS and IT,
Dr. B. A. M. University,
Aurangabad-431004, India
(MS)

Ratnadeep R. Deshmukh
Department of CS and IT,
Dr. B. A. M. University,
Aurangabad-431004, India
(MS)

Pukhraj P. Shrishrimal
Department of CS and IT,
Dr. B. A. M. University,
Aurangabad-431004, India
(MS)

ABSTRACT
Speech corpus being the basic requirement for the development of Automatic speech recognition (ASR) system, it should be done with much accuracy in order to enhance the performance of the system. This paper describes the proposed procedure to abide while collecting the speech corpus of Swahili language from the native and non native speaker for the development of Automatic Speech Recognition system in Swahili language.

General Terms
Speech Database, Speech Recognition, Natural Language Processing, Human Computer Interaction

Keywords
Swahili, Swahili Text corpus, Phonetics, Text Corpus and Speech Corpus, Automatic Speech Recognition

1. INTRODUCTION
Speech is the most prominent and natural form of communication between humans. Communication among the human being is dominated by spoken languages; therefore the researchers are trying to build the speech interfaces for computer [1]. Speech has potential of being used as a mode of interaction with computer. Human beings have long been motivated to create systems that can understand and talk like human. In this direction, researchers have tried to develop system for analysis and classification of the speech signals.

Since, 1960s computer scientists have been researching ways and means to make computer record, interpret and understand human speech. Speech processing has become increasingly important in daily life, as the number of web enabled mobile phone users in rural as well as in urban area is increasing. Most of the researches efforts are in the field are of natural language processing (NLP) for African Language where Swahili is included and has majorly rooted in the rule based paradigm. The rule based approach has some merits as well as demerits. The merit of Swahili is in term of its design transparency and the demerits are that it's highly language dependency and costly to develop as it typically involves a lot of manual effort of experts from Natural Language Processing field.

The system are decidedly competence based which is often tweaked and tuned towards a small sets of ideal sample words or sentences neglecting the real-world language technology application. In Language technologies for many African languages the researchers are getting tired of publication on real-world data or reports.

Currently with the increased need of digital resource usage in the continent of Africa, there is a great need for more empirical approaches such as data driven and corpus based approach for language technologies.

The main advantages of these approaches are: language impendence, development speed, robustness and empiricism. There is scarcity of sources in the sense that the digital text resources are few. The recent effort on the same is handled carefully with selected procedure for Swahili [2, 3]. For language technology applications such as speech recognition system, text-to-speech synthesis, machine aided translation and web related issues there is a great need for translation and usability of the Swahili language. There is a great need of work to be done in semantics and syntactic of Swahili language as the biggest online web Text resources which are available on Google, Yahoo and Wikipedia are not that correct. The major need is the extraction of information which enhances and refocuses on embarking on Swahili as a language, the corpus availability needs to be syntactically & semantically correct.

This paper focuses on the procedure to be followed for the development of isolated numeric speech corpus. The information about Swahili language is described in the Section 2. Section 3 describes the Swahili text Corpus selected. Section 4 describes the procedure to be followed for developing the speech corpus. Section 5 describes the recording procedure to be followed. The Conclusion and the Future work are discussed in Section 6.

2. ABOUT SWAHILI / KISWAHILI LANGUAGE
Swahili language is an agglutinative language with a rich morphology. The language is still under-sourced and much work is on development stage. The basic phone set of Swahili comprises of 5 vowels and 27 consonants [4]. Swahili is currently written in a slightly defective orthography using the Roman alphabet. Swahili has no diphthongs; in vowel combinations, each letter is pronounced separately. The language had previously been written in the Arabic script, unlike adaptations of the Arabic script for other languages, relatively little accommodation was made for Swahili.

Swahili is a Bantu language that serves as a second language to various groups traditionally inhabiting in the parts of the East African coast. Some Swahili vocabulary is derived from

Arabic through more than twelve centuries of contact with Arabic-speaking inhabitants of the coast of Zanj. It has also incorporated Persian, German, Portuguese, English and French words into its vocabulary through contact during the last five centuries. Swahili has become a second language spoken by millions of people in three countries in African Continent, Tanzania, Kenya, and Congo (Democratic Republic of Congo), where it is one of the national languages out of four recognized national languages [5]. The neighboring nation of Uganda made Swahili a required subject in primary schools from 1992, although this mandate has not been well implemented and declared it as an official language in 2005 as preparation for the East African Federation. Swahili, or other closely related languages, are spoken by nearly the entire population of the Comoros and by relatively small numbers of people in Burundi, Rwanda, Malawi, Northern Zambia and Mozambique. The language is still understood in the southern ports of the Red Sea and along the coasts of Southern Arabia and the Persian Gulf in the twentieth century [3]. In the Guthrie non genetic classification of Bantu languages, Swahili is included under Zone G.

The earliest known documents written in Swahili are letters written in Kilwa in 1711, in the Arabic alphabet. They were sent to the Portuguese of Mozambique and their local allies which are now preserved in the Historical Archives of Goa, India [6, 7]. Another ancient written document is an epic poem in the Arabic script titled Utendi wa Tambuka (The History of Tambuka); it is dated back in 1728. The Latin alphabet has become standard under the influence of European colonial powers [8, 9].Swahili is unusual among sub-Saharan languages having lost the features of lexical tone (with the exception of the numerically important Mvita dialect, the dialect of Kenya's second city, the Indian Ocean port of Mombasa).

3. SELECTED TEXT

Text corpus is a very crucial when working in the domain of language technologies such as language modeling, development of Automatic Speech Recognition (ASR), speech synthesis systems and speaker recognition systems. There are few attempts carried out for the development text corpus in Swahili like Helsinki corpus annotated as SALAMA [10], Sawa corpus [11] and other for different purposes such as machine leaning, lexical acquisition and machine translation.

The text corpus that is developed in Swahili language contains sentences mostly from books, journals, short stories, novels, articles from newspapers, magazines and other literature available in Swahili. In the earlier attempt in development of Speech Recognition system they worked on words, sentences from Helsinki Corpus. No work was done for the development of speech recognition system for number from 0 to 9 in Swahili so it decided to use the numbers to be the text corpus for the proposed word. The Table no. 1 shows the number 0-9 and how they are written in Swahili/ Kiswahili.

Table. No. 1. The numbers in Kiswahili/Swahili language

No's	**0**	**1**	**2**	**3**	**4**
Written	Sufuri	Moja	Mbili	Tatu	Inne

No's	**5**	**6**	**7**	**8**	**9**
Written	Tano	Sita	Saba	Nane	Tisa

4. DEVELOPMENT OF SPEECH CORPUS

Speech corpus is very crucial and a basic requirement during the development automatic speech recognition system, speaker recognition system, natural language processing (NLP), speech modeling, speech synthesis system and many other speech related areas. The attempts done in the development of speech corpus for Swahili / Kiswahili were for studying verb morphology for Swahili[12], development of text to speech [13] and a speech corpus for ASR [14]. During the earlier study conducted [15] the speech corpus developed for the Swahili is not large. The earlier speech corpus developed for the development of ASR in Swahili used the words from the Helsinki Text Corpus. The earlier work motivated us to develop a speech corpus of number from0-9 which would be helpful to initiate the research for the resource scarce Swahili language.

The following subsection describes the steps to be followed for developing speech corpora. First step is the selection of speaker for collecting the speech samples. Next subsection describes steps for speech data collection and the last subsection describes the probable statistics of Speech corpus to be developed.

4.1. Selection of Speaker

The speech data will be collected from the native as well as non native speakers of Swahili language. The selected native speakers would be resident of the countries where Swahili/ Kiswahili are one of the recognized national languages. The non native speakers would be those who are comfortable in reading, writing and speaking the English language fluently. The speakers will be selected to cover the complete diversity i.e. age group, gender and language knowledge. For the non native speakers the training will be given to them for the pronunciations.

4.2. Data Collection

The speech data would be collected from 60 speakers in total out of which 30 speakers would be the Native speakers (15 male and 15 female) and 30 non native speaker (15 male and 15 female). Proper attention will be given to select the speakers from different age groups between 20-50 years. The non native speakers would be selected on the basis of the educational qualification and their comfort level with English language.

4.3. Data Collection Statistics

As mentioned in earlier section the speech samples would be collected from 60 speakers 30 native speakers (consisting 15 male, 15 female) and 30 non native speakers (consisting 15 male, 15 female). Each speaker will be asked to speak the numbers from 0-9 with 5 utterance of each number. From every speaker we would be recording 50 samples. After the completion of the recording of speech samples; we will have approximately 300 utterance of each number i.e. total 3000 utterance of the numbers.

5. RECORDING PROCEDURE

The isolated words *i.e.* numbers (from 0 to 9) will be recorded using two different high quality headsets i.e. (Sennheiser PC360 and PC350) and using the PRAAT Software. The data will be recorded in noisy environment. The recording of the Speech samples in such noisy environment will be very useful in future for the development of robust automatic speech recognition system. The speech samples would be recorded in mono mode with a sampling frequency of 16000Hz. A

microphone will be at a distance of about 3 cm from the mouth. The PRAAT software is a freeware tool that is being widely used by the researchers working for the development of Speech technologies.

6. CONCLUSION

This paper presents the procedure that is to be followed for developing the isolated numeric speech database in Kiswahili / Swahili. The proposed database will be useful to cover the basic phonetics for Kiswahili/Swahili language. The earlier study that was carried had motivated for the development of speech corpus. This corpus will be helpful for the development of Automatic Speech Recognition System which could handle the Swahili spoken by native as well as non-native speakers of the Swahili.

This work can further be extended for continuous spoken Swahili/ Kiswahili speech. The current attempt is just for number (0 to 9); in near future the work would be extended for the design and development of speech database and automatic speech recognition system for tourism, health services and for agriculture domain.

7. ACKNOWLEDGMENT

We would like to thank the University Authorities to provide the basic facilities for carrying out the research work. This work is supported by University Grants Commission.

8. REFERENCES

[1] Pukhraj Shrishrimal, R. R. Deshmukh, Vishal Waghmare, 2012 “Indian Language Speech Database: A Review”, International Journal of Computer Application (IJCA), Vol 47, No. 5, (June – 2012), pp. 17-21.

[2] Guy De Pauw and Gilles-Maurice de Schryver,2008 “Improving the Computational Morphological Analysis of a Swahili Corpus for Lexicographic Purposes” Lexikos 18 (AFRILEX-reeks/series 18: 2008): 303-318

[3] G. De Pauw, G. M. de Schryver, and P.W. Wagacha, 2006 “Data-driven part-of-speech tagging of Kiswahili”. In P. Sojka, I. Kopeˇcek, and K. Pala, editors, Proceedings of Text, Speech and Dialogue, 9th International Conference, volume 4188 of Lecture Notes in Computer Science, pages 197–204, Berlin, Germany, Springer Verlag.

[4] Gakuru, Mucemi Iraki, Frederick K. Tucker, Roger Shalonova, Ksenia Ngugi, Kamanda 2005, “Development of a Kiswahili text to speech system”, In INTERSPEECH-2005, 1481-1484.

[5] http://en.wikipedia.org/wiki/Languages_of_the_Democratic_Republic_of_the_Congo dated 27/06/2012

[6] E.A. Alpers, 1975 “Ivory and Slaves in East Central Africa”, London, pp. 98– 99 ;

[7] T. Vernet 2002, “Les cités-Etats Swahili et la puissance omanaise” (1650– 1720), Journal des Africanistes, 72(2), pp. 102–105.

[8] Thomas J. Hinnebusch, 1992 “Ethnologue list of countries where Swahili is spoken”, “Swahili", International Encyclopedia of Linguistics, Oxford, pp. 99–106

[9] David Dalby, 1999/2000, “The Linguasphere Register of the World's Languages and Speech Communities”, Linguasphere Press, Volume Two, pg. 733–735

[10] Arvi Hurskainen, 2004 “Helsinki Corpus of Swahili. Compilers”: Institute for Asian and African Studies (University of Helsinki) and CSC.

[11] Guy De Pauw, Peter Waiganjo Wagacha, Gilles-Maurice de Schryver, 2011 “Exploring the SAWA corpus: collection and deployment of a parallel corpus English—Swahili”, International Journal of Lang Resources & Evaluation, Springer Verlag, vol 45, pp 331-344.

[12] Deen, Kamil Ud 2002 “The acquisition of Swahili verbal morphology”, Palmela, Portugal. Costa, Joao & Freitas, Maria (Eds), in the proceedings to G.A.L.A conference (2002c) pp.41-48.

[13] Gakuru, Mucemi , Frederick K. Iraki, Roger Tucker, Ksenia Shalonova, Kamanda Ngugi, 2005"Development of a Kiswahili text to speech system", In INTERSPEECH-2005, pp1481-1484.

[14] Hadrien Gelas, Laurent Besacier, F. Pellegrino, 2012 “Developments of Swahili resources for an automatic speech recognition system”, SLTU – Workshop on Spoken Language Technologies for Under-Resourced Languages, Cape-Town, South Africa.

[15] Aaron M. Oirere, Ratnadeep R. Deshmukh, Pukhraj P. Shrishrimal, Vishal B. Waghmare, "Swahili Text and Speech Corpus: A Review", Asian Journal of Computer Science and Information Technology, Vol. 2, No. 11, (Nov-2012), pp. 286-290.