

Development of a Kiswahili Text To Speech System

*Mucemi Gakuru¹, Frederick K. Iraki², Roger Tucker³,
Ksenia Shalnova³ and Kamanda Ngugi¹*

¹University of Nairobi, Nairobi, KENYA

²United States International University, Nairobi, KENYA

³Outside Echo Ltd., Chepstow, U.K.

Abstract

This paper discusses how a concatenative Kiswahili Text to Speech System (TTS) was developed based on the Festival Unit Selection Speech Synthesiser. It explains how important Kiswahili linguistic features such as phones, stress and intonation were modelled as inputs to the Festival engine. It also discusses the design, recording and segmentation of the speech database, beginning with text corpus collection and transcription. The choice of the speaker, which is crucial to realising a good TTS is discussed and also how the system was tested.

1. Introduction

Kiswahili has now been recognized as a regional language in the East African region with over 200 million speakers, thereby posing a potential for a large market for Kiswahili based software products. This fact has been recognized by major IT companies such as Microsoft and Google whose products, Microsoft Windows and the Google Search engine respectively, are already available in Kiswahili. Microsoft is also in the process of localizing its MS Office to Kiswahili. The potential for Kiswahili as a market for IT products now appears to have been fully acknowledged.

The work presented here is the development of a Kiswahili Text to Speech System (TTS) based on the Festival Unit Selection Speech Synthesiser [1, 9]. The system is expected to have a wide range of applications, for example, in software aids to visually handicapped people which would enable integration of a large number of blind children into schools. Others would be in mobile phones and server based voice services.

The system developed here is concatenative [1, 3, 4], which means that synthesis is done through creating waveforms by concatenating parts of natural speech recorded from humans. In this type of system all the acoustically and perceptually significant sound variations (allophones or simply phones) in the language are recorded so that they are played back each time the system synthesises speech.

In developing such a system, it is assumed that co-articulation, the mutual influence between adjoining sounds, does not extend beyond phone-phone boundary [1]. Synthesis is carried out using diphones, defined as the mid-point of one phone to mid-point of the next, which are captured within the context of several words in order to take care of prosodic features. When carrying out speech synthesis, the Unit

Selection procedure then employs the optimal selection of the diphones from among the recorded words.

To include all language features, this procedure must therefore use a precisely designed speech database [9] that comprises of phonetically balanced sentences, i.e a limited set of sentences containing all the possible diphones in Kiswahili.

Kiswahili linguistic features needed as inputs in the development of the TTS are defined and presented below.

2. Phone set definition

The phone set was developed from the Kiswahili dialect of Kiunguja, adopted as the standard Kiswahili in East Africa in 1928 [7]. This is the dialect employed in official writing and broadcasting.

2.1. Vowels and Consonants

The basic phone set is well defined [5, 6] and comprises of five vowels:

[a, e, i, o, u]

and 27 consonants:

[b, ch, d, dh, f, g, gh, h, j, k, kh, l, m, n, ng, ng',
ny, p, r, s, sh, t, th, v, w, y, z]

The consonants gh and kh are borrowed from Arabic and Persian. The consonants w and y are semi-vowels, while n and m are the nasals.

2.2. Prosodic features

Prosodic features include stress, pitch and intonation (We discuss pitch and intonation later). Since stress (emphasis of a syllable in word) can either be lexical (word-based) or rhythmical (phrase-based), it was necessary to define a new set of phones to reflect this reality. These are lexical and rhythmical phones which we now describe.

2.3. Lexical phones

A new set of phones were introduced to take care of lexical stress [8]. These phones correspond to the vowels and two nasals (*m* and *n*), which are known to exhibit the lexical stress and were designated as follows:

[a1, e1, i1, o1, u1, m1, n1]

2.4. Rhythmical phones

A further set was also introduced to represent rhythmical stress [8], the emphasis of a syllable in a group of words representing a unit of meaning. These phones correspond to the vowels and nasals (*m* and *n*), which are known to exhibit the rhythmical stress and were designated as follows:

[a2, e2, i2, o2, u2, m2, n2]

In total, 42 phones were defined for the Kiswahili dialect of Kiunguja. These would be the inputs for the subsequent procedures.

3. Kiswahili Letter-to-Sound mapping and Syllabification rules

Kiswahili's letter-to-sound (grapheme to phone) rules are quite systematic and therefore a pronunciation dictionary was not required, except for proper nouns. However, besides the phones, the Festival engine requires Kiswahili syllables. These are well defined [5, 6] and were implemented through syllabification rules developed as follows:

- A consonant or a vowel preceded by a vowel is the start of a syllable
- A consonant (other than a semi-vowel) preceded by a consonant is the start of a syllable (this only happens for *n*, *m* or borrowed words)
- All syllables end at the beginning of the next syllable or at the end of the word.

It therefore follows from these rules that whenever a syllable begins at a vowel, then that vowel is the syllable. For example *mia* (hundred) has two syllables *mi* and *a*.

4. Integrating prosody

4.1. Stress

The two forms of stress implemented were the lexical stress and rhythmical stress [8]. Lexical stress was implemented by substituting all vowels and nasals¹ in the penultimate syllable with the corresponding lexical phones.

Rhythmical stress was implemented at the end of each phrase or sentence (A phrase here refers to a cluster of words that is normally shorter than a full sentence), by substituting the stressed vowels and nasals¹ in the penultimate syllable with the corresponding rhythmical phones.

Stress in Kiswahili is mainly of a durational nature. However, there could be a case for pitch changes (Iraki & Gakuru, forthcoming).

4.2. Pitch and intonation

In a narrow sense, pitch is the melodic height of a phone while intonation describes the patterns of pitch in a language [10]. These two features are crucial in defining a language. In our

¹ for nasals substitution is only done when they appear as syllables

implementation, these features are built intrinsically into the system through the speech database. However, some forms of intonation, e.g interrogatives (?) or exclamations (!) were not implemented in this work.

It is noted here that intonation can be implemented by merely improving the speech database. This was done in this work firstly by introducing the expanded phone set, comprising of the basic phone set and the additional Lexical and Rhythmical phones. Then the speech database was designed to fully embrace a complete set of phonetically balanced sentences (the sentences being phonetically balanced with respect to the expanded phone set).

Many language features were captured because the results of the transcription of a given word depend on the intonation, hence improving on the selection by the Festival engine. This automatically led to overall improvement of the system's intonation.

5. Text corpus collection, normalisation and transcription

A large Kiswahili text corpus was collected comprising of 10,558 sentences from novels, the Quran, the Bible, written speeches, newspaper articles among others. The collection was done from many different sources to make sure that it captured as many language features as possible. The corpus was normalised [3, 4] so that abbreviations, e-mail/URL, digit strings: currencies, dates, telephones numbers, time etc were written out fully in words. This tokenisation method was then implemented into Festival and text data is subjected to similar normalisation every time the system synthesizes speech.

The complete text was then transcribed using a combination of Festival Speech Synthesis System [1] and Kiswahili G2P (grapheme to phoneme) tool developed especially to take care of the rhythmical stress. The transcribed sentences appear as strings of phones, with the sentence boundaries, phrase boundaries, and word boundaries clearly marked with specific symbols.

6. Design of Speech Database

The selection of a set of phonetically balanced sentences was based on the LLSTI Optimal Database selection tool [13]. The tool takes as input the units to be selected and chooses the minimum number of sentences that contain the units by comparing the text corpus and the transcribed text. The units must be in the transcribed text and could be single phones, syllables, phone-phone combinations or indeed any other items deemed important.

It is therefore clear that the choice of these units is critical to realizing the selection of phonetically balanced sentences. Here the units used were: all the phones, all the syllables and all the phone-phone combinations. Further units were derived by considering the position of all the above. That is whether they occur at the beginning of the sentence, end of the sentence, beginning of word, end of word, middle of word, beginning of phrase and at the end of phrase. These combinations yielded a total of 3,725 units.

It is clear however that not all units would be found in the transcribed corpus, as some combinations may not exist in the

language. However, it was necessary to come up with an exhaustive list of units so that every possible language scenario was included in the selection. From the results obtained it is shown here, that this exhaustive approach results in the selection of phonetically balanced sentences.

Of the 3,725 units set up for sentences selection, 1997 units were found in the corpus and 414 sentences were selected as the minimum number of sentences to contain the possible units found in the corpus.

A phone count was then carried out both in the text corpus (10,558 sentences) and in the selected sentences (414 sentences) and an almost 100% correlation was achieved as shown in Fig. 1. In other words, the large corpus collected from several independent sources had almost the same phone distribution as in the one in the 414 sentences. This observation confirmed that the selected sentences were phonetically balanced and therefore representative of the Kiswahili sound system. This set of 414 sentences was the database in the development of the TTS.

The next challenges entailed the selection of a felicitous person to read the constituted database and the recording of the material in appropriate formats.

annotation. The speaker chosen here, Ken Walibora¹ is a professional newscaster who has also published several books in Kiswahili. [11, 12]. He was therefore deemed fluent in the Kiunguja dialect.

The recording was done in a professional studio at the Kenya Broadcasting Corporation (<http://www.kbc.co.ke>). The resultant data was stored in a digital format.

The advantages of using professional speaker and facilities became all too clear when recording was done. The speaker was familiar with using the microphone and how to prompt re-recording of the sentences whenever an error occurred. Recording time was therefore short and it took only 45 minutes to record the whole database.

8. Annotation and Segmentation

The database was hand-labelled to yield the annotated and segmented data, which was used as input to the Festival engine. This is a laborious procedure and only 10-15 sentences can effectively be labelled in a day. The Festival Multisyn engine also includes procedures that use the HTK [9] tool for segmentation and annotation. This tool was tested and was shown to work fairly well based on comparison with the hand-labelled data and the TTS quality made from its labelling.

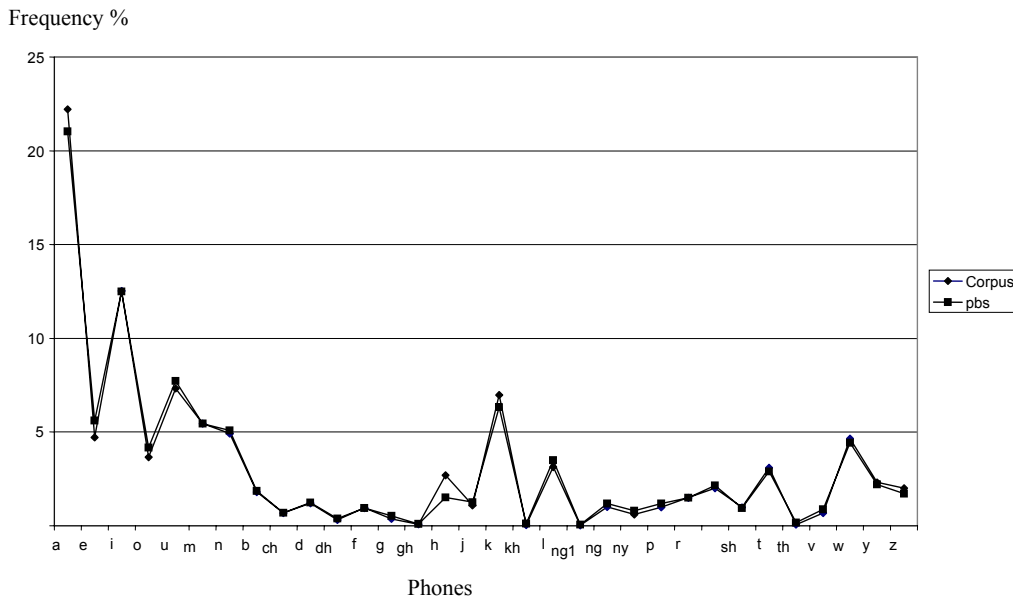


Figure 1: Plot of phone frequency in corpus vs selected sentences

7. Speaker Selection and Recording of the Database

When selecting a speaker, it is argued that professional speakers, who are generally aware of the language features such as phonemes, are often better than non-professional speakers [1]. A good voice is also hard to “spoil”, i.e. a good voice will give good results in the detailed phonetic

Nonetheless, the hand-labelled data was found to offer a much better quality synthesis than the HTK-labelled data. It should be noted that HTK-labelling can yield good results, but will require a large database as used in the *nina* [1] and *awb*[1] voices.

¹ Read Ken Walibora’s story at <http://www.llsti.org>

9. Testing the system

The system was rigorously tested by making it synthesize a large number of randomly configured sentences. Missing diphones usually result in the system crashing. Such cases are taken care of through back-off rules, so that whenever a diphone is missing the next best choice is made to replace it.

There were no cases of the system crashing and this was attributed to the comprehensive design of the speech database. However a limited number of back-off rules were implemented for all the vowels. This was deemed especially important for vowels and nasals with rhythmical stress because they were few in the speech database.

The system has been verified in scientific conferences and in University seminars. The voice is now available for demonstration¹.

10. Conclusion

A Kiswahili TTS system has been developed and tested. Its performance was found to be acceptable, even though many intonation features were not explicitly implemented. This was attributed to the rigorous design of the speech database which tended to compensate for some of them. Implementation of these tonal features like question marks will therefore be subject for further work. Others to be considered for future work include implementation of homographs, though from Kiswahili speech point of view they are limited.

Further improvements on this system would include enriching the database so that there are more examples of “m” “n” and “r”. Also for the semi-vowels which were found hard to isolate during segmentation, new phones can be defined to comprise of all the syllables made from semi-vowels.

11. Acknowledgements

This work was initiated through the LLSTI project, which also helped in getting sponsorship. The funding was by Oneworld as part of a grant from the Vodaphone foundation.

We would like to thank LLSTI contributor Parthar Talukdar for providing the Optimal Text Selection Tool and Joseph Mbutia of Nation Newspapers for helping secure the professional speaker.

12. References

- [1] A. Black, P. Taylor and R. Caley. The Festival speech synthesis system. <http://www.cstr.ed.ac.uk/projects/festival.html>, 1998.
- [2] Carton, F. (1974), Introduction à la phonétique du français, Paris, Bordas.
- [3] Dutoit, Thierry, *An introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht, 1997.
- [4] Dutoit, Thierry, “High Quality text-to-speech synthesis : an overview”, Journal of Electrical & Electronics Engineering, Australia : Special issue on speech recognition and synthesis, vol 17 n°1, 1996.
- [5] Mohammed M.A. *Modern Swahili Grammar*, East African Educational Publishers , Nairobi, 2001.
- [6] Ellen Contini-Morava, “Swahili Phonology”, University of Virginia, <http://rosettaproject.org/work/rosetta>
- [7] Iraki F.K. ”*Lecture pragmatique des morphemes temporels du swahili*”, Phd dissertation, University De Geneve, 2002. www.unige.ch/cyberdocuments
- [8] K. Shalanova and R Tucker, “South Asian Languages in Multilingual TTS-related Database” EACL Workshop on Computational Linguistics for the Languages of South Asia, Budapest, April 2003, pp 57-63.
- [9] Robert Clark, Korin Richmond and Simon King. FESTIVAL 2 – Build your own general purpose Unit Selection Speech Synthesiser. CSTR, The University of Edinburgh, July 2004.
- [10] David Crystal, “*A Dictionary of Linguistics and Phonetics*”, Blackwell Publishers Ltd, 1997.
- [11] Ken Walibora, *Kufa Kuzikana*, Longhorn Kenya, 2003.
- [12] Ken Walibora, *Mgomba Changaraweni*, Phoenix Publishers, 2004
- [13] Talukdar, P., "Optimal Text Selection Module Version 0.2", available from <http://www.llsti.org/downloads-tools.htm>

¹ View the demonstration at <http://www.llsti.org>