INTERSPEECH 2009 BRIGHTON

Development of a Kenyan English Text To Speech System: A Method of Developing a TTS for a previously undefined English Dialect

Mucemi Gakuru

Teknobyte Ltd, Nairobi, KENYA

mucemi@teknobyte.co.ke

10.21437/Interspeech.2009-592

Abstract

This work provides a method that can be used to build an English TTS for a population who speak a dialect which is not defined and for which no resources exist, by showing how a Text to Speech System (TTS) was developed for the English dialect spoken in Kenya. To begin with, the existence of a unique English dialect which had not previously been defined was confirmed from the need by the English speaking Kenyan population to have a TTS in an accent different from the British accent. This dialect is referred to here and has also been branded as Kenyan English®. Given that building a TTS requires language features to be adequately defined, it was necessary to develop the essential features of the dialect such as the phoneset and the lexicon and then verifying their correctness. The paper shows how it was possible to come up with a systematic approach for defining these features through tracing the evolution of the dialect. It also discusses how the TTS was built and tested.

1. Introduction

1.1. The need for a Kenyan English TTS

Most countries include a world language amongst their official languages, and this language is the obvious choice when deploying a speech system which requires TTS. When that language is English, it is tempting to suppose that because of imported films and TV people would be used to hearing US or UK English, and therefore either would be adequate for the output of a speech system.

This hypothesis was tested in Kenya with a pilot voice information line on growing bananas [1][2] that used English female English TTS (Nina) and a Kiswahili TTS system we had previously developed[3]. The system was evaluated by rural Kenyan farmers. It was observed that although they liked to listen in English in preference to Kiswahili, they struggled with the British accent. They faced difficulties in picking up entire phrases due to differences in intonation and pronunciation, especially of the large number of borrowed words from local languages.

Speech was then played from a demonstration Kenyan English TTS system built by simply re-recording the database of a US English TTS system with a Kenyan speaker. It was observed that they were very comfortable with this speech and that in general they found it even clearer than the Kiswahili, even though the pronunciation still needed refining. This suggested that the dialect of English which exists in Kenya is different from British English mainly in pronunciation.

We concluded that the development of a TTS based on a proper representation of the Kenyan English dialect was necessary in order to produce a system whose output would be natural to the Kenyan population and hence more easily understood. The problem, as is so often the case with local dialects, is that the dialect itself is evolving and has never been linguistically defined.

1.2. Language resources for developing the TTS

The language resources needed for building a TTS include a lexicon, tokenisation rules and Part of Speech tagging. Given that the grammar in the Kenyan English is the same as the British English, the tokenisation rules and Part of Speech tagging could therefore be adopted from existing British English systems. However, it was observed that in order to produce a fairly good Kenyan TTS, a Kenyan English lexicon would be needed. This became clear from the demonstration Kenyan English system which had been made from an open-source US English TTS and the OALD lexicon by simply rerecording the US English unit-selection database with a Kenyan speaker then performing automatic labelling using the Hidden Markov Model Toolkit (HTK). Although the demonstration voice was intelligible enough, it had many instances of mispronunciation. It also failed totally to pronounce borrowed words from local languages.

In section 2 we show how it is infeasible to carry out phone mapping from the existing English lexicons such as OALD and Unilex, thereby concluding that a completely new lexicon has to be built. In section 3 we describe the procedure for building the lexicon, including testing its validity. In section 4 we describe how we construct the speech database to ensure good coverage of acoustic features including borrowed words, and in sections 5 & 6 how the speaker was chosen and the system evaluated.

2. Defining the Features of Kenyan English

The evolution of the dialect has been systematic and can be seen to use fairly clear rules in adoption of phones from local languages as well as integration of borrowed words. This change can be attributed to some key factors which made it possible to adopt a consistent approach for developing the features of the dialect as is now described.

2.1 Evolution of the Kenyan English dialect

English language was introduced in Kenya in early 1900s by the British Colonial authorities. It was adopted as an official language at independence in 1963 for it carried with it scientific knowledge, politics (newspapers and publications), business (banking and documentations) as well as matters of the state (registration, laws etc). It also became the language of instruction in schools and thus a large number of Kenyans started teaching in English and also teaching English. Previously the language had been taught by the native British English speakers and it was essentially British English. The language has evolved considerably since independence [4] and this can arguably be traced to some of the following factors.

To begin with, English speaking has passed on from early speakers through many Kenyan generations to the current speakers, with each generation modifying it from the previous

one. Secondly, there are around 42 ethnic groups in Kenya each with its own local language, while Kiswahili is used as the local language in major towns. Most people in Kenya learn to speak in their local language first before moving on to learn English or any of the other local languages. In fact for most people, English is first learnt in school as a written language and the tendency by many is therefore is to map letters or words in English text to phones or phone-streams of their first language which they will already have learnt. Many people in Kenya speak at least three languages: their local language, Kiswahili and English. Each of the languages inevitably borrows heavily from the others. Therefore the Kenyan English dialect contains many borrowed words, such as people's names and names of places that are pronounced more like in their local language.

Officially the grammar and orthography in Kenyan English remains the same as in British English; both use the same alphabet, word spelling and grammar rules, at least as taught in school and as required in all official documents. The Kenyan English dialect mainly differs from British English in pronunciation and inevitably by having very many borrowed words from the local languages. For the Kenyan English very few resources such as a phoneset or lexicon exist as it had not been previously studied, and therefore had to be developed. The easiest method to develop them would have been by simple modification of the British English, say by using one-to-one or many-to-one phone mappings. However this was not possible and a new phoneset and a corresponding lexicon had to be built from scratch.

2.2 The phoneset

2.2.1. Vowels

As indicated in Section 1.2. English in Kenya is first learnt as a written language and therefore vowels are essentially derived from the five Kiswahili phones:

a e i o u

This contrasts with the British English phoneset which include more vowels, as illustrated in Table 1 below:

British English vowels	Example of where used	Kenyan English vowel
^	cut	a
ɑ	father	a
æ	cat	a
ɜ	turn	a
e	met	e
ɪ	hit	ɪ
i:	see	i:
ɒ	hot	o
ɔ:	call	o:
U	put	u
u:	blue	u:

Table 1. *Comparison of British and Kenyan English Vowels*

In the Kenyan English, the vowels and consonants are also not applied by direct substitution of the British ones with the corresponding Kenyan ones above. For example, the word '*where*' which is pronounced as *w eə,* in British English with *h* being silent, is pronounced as *h w ea* in Kenyan English. There is also no equivalent of the British English schwa */ə/* in Kenyan English and instead various vowels are used. For example, the word cinema in Bristish English is pronounced as *s i n ə m ə* while in Kenya English it is pronounced as *s i n e m a.* This completely ruled out the possibility of predicting pronunciation simply by phone mapping.

2.2.2. Consonants

The consonants are almost the same as those of British English, with the only difference being the absence of the sound ʒ (as in pleasure or vision) which is substituted by *'sh'*. Kenyan English also fails to pick the *'dh'* sound (as in father) a fact that can be traced to the tendency (as mentioned above) by many people to map words in English text to phone-streams in Kiswahili. Because *'dh'* is written explicitly in Kiswahili (for example in the word idhaa) instead of using '*th*', many people consistently map all '*th*' in English text to *'th'* the same way it is done in Kiswahili thereby losing the '*dh*' sound in English. When developing the Kenyan English TTS, the *'dh'* sound can therefore be left out without affecting the quality of the system.

2.3. Borrowed words

Kenyan English sentences often borrow a word or two from the local languages or from Kiswahili. Such words include names of places and people which are inherently Kenyan and therefore would not be pronounced correctly in any English dialect. It is also observed that the borrowed words are not pronounced as in the local language but are first given Kiswahili pronunciation and intonation before adoption into the Kenyan English.

To illustrate how borrowed words are systematically carried into the Kenyan English dialect, the following examples from three distinct local languages are given: In the Kisii language the word *Nyaribari* is pronounced as *ny a r i – v a r i,* but it is pronounced using the Kiswahili equivalent *ny a r i – b a r i,* when adopted into the Kenyan English. Similary, in Kikuyu language the word *Gatundu* is pronounced as *g a t ou d ou* but it is pronounced as *g a t u n d u* in Kenyan English. While in the Kamba language the word *Kyulu* is pronounced as *ch u – l u* but changed to *k iu - l u* on adoption to Kenyan English.

3. Building the Lexicon

Building a lexicon is an onerous task, which takes a long time [5]. Furthermore if the wrong phoneset is used, especially for an un-studied language, the whole exercise can end up in futility. It was therefore necessary to come up with an efficient method that would yield a good lexicon in a reasonable time, and also to come up with a way of checking the correctness of such lexicon. The technique developed here employed an iterative approach which began by setting up an initial phoneset informed by the language features described in Section 2. Based on this starting phoneset, a lexicon of 10,000 most commonly used words in the Kenyan English was then developed using a G2P tool, the DictionaryMaker developed at Mareka Institute[6][7][8]. A trial Kenyan English voice was then made using this lexicon and the recordings from the demonstration Kenyan voice described earlier. This was done in order to find out whether the lexicon would yield an improved voice. Adjustments were carried out on the phoneset and the corresponding lexicon repeatedly until we were satisfied that they captured the Kenyan English dialect as best as possible. Having settled on a final phoneset, the lexicon was built up to 60,000 words including 6,000 borrowed words. After every addition of 10,000 words to the lexicon, a new voice would be built from the same recordings and it was

observed that the voices continued to improve. Note that had the speech database been sufficient to build the Kenyan English voice, the dictionary could just have been built from the words in the prompt list instead continued addition of 10,000. However it was clear that a new speech database would be needed and therefore a more comprehensive dictionary had to be built.

The Festival Engine was then used to extract corresponding letter-to-sound rules for the Kenyan English, enabling the lexicon to be shrunk by 85% of its original size. It was these letter-to-sound rules and the residual lexicon that were used for the development of the Kenya English TTS. At this point, with the phoneset and the lexicon having been developed, the Kenyan English had been substantively defined. Authentication of these language definitions would now be carried through building and testing a complete Kenyan English TTS based on these language features.

4. Design of the speech database

The TTS discussed here is based on unit-selection concatenative synthesis [5] which is done by stringing together pre-recorded segments of speech stored in a speech database. This type of synthesis was chosen as it produces the most natural-sounding speech compared to rule-based synthesis, diphone synthesis or HMM-based synthesis. The speech database is therefore crucial as it must contain sufficient example of the speech segments to be stringed together. The building of such a database is now discussed.

4.1. Text corpus collection, normalisation and transcription

It was important to gather the corpus from many different sources so as to capture as many borrowed words as possible and also how they are used in the Kenyan English. Furthermore different writing styles meant different ways of using the borrowed words, which had to be captured. A large text corpus from such diverse sources would ensure that a phonetically balanced speech database could be made to contain sufficient examples of the language features.

A text corpus comprising of 20,000 sentences, was therefore collected from various sources such as novels, newspaper articles, the Kenyan draft constitution, written speeches among others. The novels used were those rich in local names. For example, one novel is a narrative capturing the Luo traditional community and was thus rich in Luo names and words. The other set in the late 1970s was a memoir which had many local names and was enriched with Kikuyu words. Yet another novel used was set in Nairobi and in recent years.

The corpus was then checked for grammatical and typographical errors like improper punctuations and spelling mistakes which would impact on the transcription. A transcribing tool was developed in Scheme to run on the Festival Speech Synthesis System Engine [5][9]. Transcription of the sentences was then carried out using this tool, the Kenyan English lexicon and letter-to-sound rules developed earlier. The transcribed sentences appeared as strings of phones, with the boundaries marked by the various pauses; B_100, B_150 and B_300.

For example the sentence:

What, of course, Kitosh said was, 'Nataka kufa' which means: 'I am about to die, or I am dying.'

When transcribed appeared as follows:

h w o t B_100 o f k oo s B_100 k i t o sh s ey d w o s B_100 n a t a k a k u f a h w i ch m i n s B_150 ae a m a b au t t u d ae B_100 o r ae a m d ae ng B_300

The same tool was used to store the source-sentences and the transcribed-sentences into two files both arranged in such a way that they had one-to-one correspondence in the listing.

4.2. Selection of the sentences

The approach taken here was changed slightly from the one used in design of Kiswahili speech database [3][10] so as to improve on the database. The Kiswahili database was built by first creating a hypothetical and exhaustive list of units comprising of all possible phones and all phone-phone combinations in Kiswahili, then selecting the minimum sentences covering as many of the units as possible. While the resulting database covered all the possible units in the corpus, it did not necessarily cover all units in the hypothetical list. Therefore the process could have been shortened if the actual units in the corpus were determined first and then used to create the units list. Secondly, to improve on co-articulation, phone-phone-phone combinations were added to the units list to be used in sentence selection.

It had also been observed when designing the Kiswahili database that some sentences were very long. Inevitably such sentences tended to be selected first, even when they had low "unit-density", or the number of units divided by the number of words in the sentence, *L*. A weighting factor was therefore introduced to optimise on the length of the selected sentences. This was done by first working out the mean, *M*, and the standard deviation, σ, of number of words in each sentence. If the optimum length of the sentences is O_L then all the sentences were given a weighting factor, *w* such that

$$w = 1, for\ L \le (1 + x\sigma)O_L \qquad (1)$$

and

$$w = e^{-2(L-O_L)/\sigma^2}\ for\ L > (1 + x\sigma)O_L \qquad (2)$$

The optimum length O_L and the range *x* were determined by repeatedly applying this formula to the selection until the lengths of the selected sentences were observed to be within a certain suitable range. A good beginning point for choice of O_L was found to be the mean *M*. For this selection the length O_L turned out to be 0.8*M* with a range *x* of 0.5.

Having determined the weighting factor, the number of unique units, *n*, or the units-length, was worked out for each and every sentence. Each units-length was then weighted to give the weighted units-length u_w so that

$$u_w = w*n \qquad (3)$$

The weighted-units-length of each sentence was taken to represent its phonetic richness. The most phonetically rich sentence was then selected to become the first sentence of the speech database. The units-lengths were then reconstituted by removing from each and every sentence all the units in the selected sentence. The weighted-units-lengths were again worked out and the new richest sentence selected to the speech database. This was done repeatedly until all the units-lengths got to zero.

Out of the 20,000 sentences in the corpus, 1484 sentences were selected as the optimum number to contain all the units in the corpus. Note that the units in the corpus were not initially determined explicitly as has previously indicated. However this method implemented the requirement implicitly and therefore

ensured that the speech database covered all the units in the corpus.

5. Recording the Speech Database

The speaker chosen was Frank Muiruri a professional Kenyan newscaster and recording was professionally done at the Agricultural Information Resource Centre (AIRC) studios. Such an environment is necessary because of availability of proper equipment and absence of noise and echoes, resulting in high quality audio output. The utterances were recorded and stored in digital format.

6. Testing the system

The system has been used in the development of a voice service, the National Farmers Information Service, NAFIS [11]; which is an automated IVR for providing agricultural extension information through telephony. NAFIS is updated with summarised information through the Web by field extension officers and the information then becomes immediately available through the phone. This feature allows farmers to access regional based information which is relevant to their agro-ecological zones. NAFIS generates audio speech automatically in both the Kenyan English discussed here and Kiswahili voice developed earlier, by converting the keyed in text using the Text-to-Speech systems. The generated audios are seamlessly configured into an IVR.

The information used in NAFIS is therefore random in nature as it comes from different people as well as sources. In addition, it also contains many technical words in agriculture and therefore a very good platform for testing both Text to Speech systems. Two types of feedback on the Kenyan English system were possible; one on the diphone coverage by the speech database and the other on the grapheme-to-phoneme conversion by the lexicon and its derived letter-to-sound rules. Poor diphone coverage would result into a poor speech database that will have many instances of missing diphones thereby making the system crash during speech synthesis. However there was only on average 2 instances of system crashing due to missing diphones for every 100 text files that were converted into speech.

There were also very few cases of mispronunciation as would occur if an incorrect phone stream is generated during grapheme to phoneme conversion or if bad segments are chosen during speech synthesis. Given that the lexicon covered most of the words likely to be used in the Kenyan English and the fact that it was possible to compress the lexicon to 15% of its original size (meaning that the letter-to-sound would predict pronunciation correctly for 85% of all the words not covered in the lexicon), then the probability that an incorrect phone stream would be generated was very low. Furthermore all technical words were provided with their pronunciation through an addendum to the lexicon. These factors plus the high quality of speech recording explain why cases of mispronunciation were extremely few.

Many people have been asked to give comments on the quality on the Kenyan English voice and all have said that it was clear and understandable with no intonation issues being raised. Furthermore the respondents have been able to identify the speaker as he is a well known Kenyan newscaster.

7. Conclusion

A Kenyan English TTS system has been developed and tested, and its performance found to be acceptable. It has been shown that a TTS in the dialect was needed as the existing British and American systems were not suitable for the Kenyan Population. In order to build the TTS some language features such as the phoneset and the lexicon had to be developed, and their correctness verified. A phonetically balanced speech database, which contained sufficient examples of the speech features of the dialect, was then developed. A professional speaker was used to record the database and the system was then built and tested.

The work has clearly demonstrated that a distinct and definable dialect exists and its form has been illustrated. It is interesting to note that Thierry Dutoit [12] observes that TTS's can be powerful research tools for linguists. Indeed, it is the making of the Kenyan English TTS that has made it possible to substantively define Kenyan English dialect for the first time. This work therefore provides a method which others, especially linguists, can use to further develop the language features of the dialect. It also provides a method that can be used to build a TTS for a dialect that is not defined and with few resources. Further improvements on this system would include expanding the lexicon and improving on the speech database so that more examples of borrowed words are captured.

8. Acknowledgements

I would like to thank Faith Kamau for recoding the Faith Multisyn voice, Elizabeth King'ori for her contribution in defining the Kenyan English phoneset, Gideon Kimatu for assisting in developing the Kenyan English lexicon and Willys Obande for the collection of common Kamba and Luo names. I would also like to thank Marelie Davel of Mareka Institute for availing the DictionaryMaker and the Local Language Speech Technology Initiative for inspiring the ideas that led to this work.

9. References

[1] P Näsfors. "Efficient Voice Information Services for Developing Countries", Masters Thesis, Department of Information Technology, Uppsala University, Sweden, 2007. Also available from www.llsti.org/documents.htm

[2] R. Tucker, M. Gakuru, P. Nasfors "Kiswahili/English Voice Information Service for Banana growers in Kenya", Workshop on Mobiles and Development: Contribution of Mobile Devices to Development, 16th may 2007, University of Manchester.

[3] Gakuru, M. et al, "Development of a Kiswahili Text To Speech System" System," Interspeech 2005, Lisbon, Portugal, pp. 1481-1484, 2005.

[4] Angelina N. Kioko and Margaret J. Muthwii, "The Demands of a Changing Society: English in Education in Kenya Today," Language, Culture and Curriculum, Volume 14, Issue 3, 2001, Pages 201 – 213.

[5] 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.

[6] M. Davel and E. Barnard, "Bootstrapping for language resource generation", in the Proceedings of the Symposium of the Pattern Recognition Association of South Africa, South Africa, 2003, pp. 97–100.

[7] M. Davel and E. Barnard, "Bootstrapping pronunciation dictionaries: practical issues," Interspeech 2005, Lisbon, Portugal, pp. 1561–1564, September 2005.

[8] M. Davel and M. Peche, "Dictionarymaker user manual, version 2.0 (i)," September 2006, http://dictionarymaker.sourceforge.net/.

[9] A. Black, P. Taylor and R. Caley. The Festival speech synthesis system. http://www.cstr.ed.ac.uk/projects/festival.html, 1998.

[10] Gakuru, M. et al, "Design of Speech Data Base for Unit-Selection in Kiswahili TTS", E-Tech2004, Kenya

[11] R. Tucker and M. Gakuru, "Experience with developing and deploying an agricultural information system using spoken language technology in Kenya," IEEE Spoken Language Technology Workshop, 2008. SLT 2008, Goa, India, pp. 17-20, 15-19 Dec. 2008.

[12] 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 no1, 1996.