Design of Speech Data Base for Unit-Selection in Kiswahili* TTS

By

Mucemi Gakuru[1], Frederick K. Iraki[2], Roger Tucker[3], Ksenia Shalanova[4] and Kamanda Ngugi[1]

[1] University of Nairobi, P.O. Box 30197, Nairobi, 00100 KENYA
[2] United States International University, P.O. Box 14634, 00800, Nairobi, KENYA
[3] Outside echo Ltd., Beracah House Gloucester Road Tutshill, Chepstow Gwent NP16 7DH, U.K.
[4] HP Labs, Filton Rd, Stoke Gifford, Bristol BS34 8QZ U.K.

1. Introduction

When developing a Concatenative Text to Speech System [1, 3, 4] (i.e. a form of synthesis where waveforms are created by concatenating parts of natural speech recorded from humans) it is necessary that all the acoustically and perceptually significant sound variations (allophones) in the language are recorded so that they are played back each time the system synthesises speech.

Improvement on the system is made by assuming that co-articulation (mutual influence between adjoining sounds) does not extend beyond phone-phone boundary [1]. In this case all possible phone-phone combinations are read and recorded. Each unit of the two-phone combination is referred to as the diphone. Synthesis is then based on concatenation of the diphones thus taking care of the overlap in the phone-phone boundary.

An even better system can be realised when each diphone is captured within the context of several

words and synthesis carried out by using the best selection from the recorded words. It is clear then that this procedure must use proper selection of the sentences from which the diphones are to be captured. In other words, such sentences must be phonetically balanced; implying that they must have the same phone distribution as used entirely in the language.

2. Kiswahili Phoneset [5,6]

To begin with, the following Kiswahili phoneset was established for the standard Kiswahili dialect, Kiugunja [7]:

2.1 Vowels

Vowel	Phoneme	As in
a	/a/	*b**a**ba (father)*
e	/e/	*p**e**mbe (horn)*
i	/i/	*mt**i** (tree)*
o	/o/	*tan**o** (five)*
u	/u/	*d**u**gu (brother)*

2.2 Stressed Vowels

lexical stress	rhythmical stress
a1	a2
e1	e2
i1	i2
o1	o2
u1	u2

2.3 Consonants

Vowel	Phoneme	As in
b	/b/	***B**ata (duck)*
ch	//	***Ch**umba (room)*
d	/d/	***d**eni (debt)*
dh	/ð/	***dh**oruba (storm)*

Vowel	Phoneme	As in
f	/f/	***f**imbo (club*
g	/g/	***g**oti (knee)*
gh**	//	*lu**gh**a (language)*
h	/h/	***h**ewa (air)*
j	//	***j**engo (building)*
k	/k/	***k**azi (work)*
kh**	/x/	*nus**kh**a (duplicate)*
l	/l/	*a**l**ama (mark)*
m	/m/	***m**ama (mother)*
n	/n/	***n**azi (coconut)*
ng	/nŋ/	*kupa**ng**a (arrange)*
ng'	/ŋ/	***ng'**ombe (cow)*
ny	/ /	***ny**umba (house)*
p	/p/	***p**aka (cat)*
r	/r/	***r**adi (thunder)*
s	/s/	***s**amaki (fish)*
sh	/ /	***sh**ule* (school)
t	/t/	***t**angu (since)*
th	//	***th**umni (fifty cents)*
v	/v/	***v**azi (dress)*
w (semi-vowel)	/w/	***w**ali (rice)*
y (semi-vowel	/j/	***y**ai* (egg)
z	/z/	*z**a**wadi (present)*

2.4. Stressed nasals

lexical stress	rhythmical stress
n1 m1	n2 m2

Lexical stress [8] (which is the emphasis of a syllable in a word) was introduced on all vowels and nasals: a1, e1, i1, o1, u1, m1 and n1, whenever they occurred in the penultimate syllable (second last). However at the end of at the end of phrase or sentence, rhythmical stress[8] (which is the emphasis of a syllable in a group of words representing a unit of meaning) was introduced by using the stressed vowels and nasals: a2, e2, i2, o2, u2, m2 and n2, whenever they occurred in the penultimate syllable.

3. Text corpus 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 other others. The corpus was normalised [2, 8] so that abbreviations, e-mail/URL, digit strings: currencies, dates, telephones numbers, time etc were written out fully in words.

The complete text was then transcribed using a combination of Festival Speech Synthesis System [8] 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 boundaries marked as follows:

- Sentence boundary //
- Phrase boundary /
- Word boundary #
- Syllable boundary -

For example the sentence:

Kila mtu ataifurahia nchi yake, wakati itashinda michezo.

transcribed will appear as follows:

//k i1- l a# m1- t u# a- t a- i-f u- r a- h i1- a# n1-ch i# y a2- k e /w a- k a1- t i# i- t a- sh i- n1-d a# m i- ch e2- z o//

4. Selection of phonetically balanced sentences

The approach taken here was prompted by the text selection tool, text_sel [HP Labs, India], which is thus briefly described. 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 or indeed any other items deemed important.

It is therefore clear that the choice of the units is key to realising the selection of phonetically balanced sentences. Here the units considered were:

- all phones denoted here as P
- all syllables (Combinations of Consonant Semivowel Vowel)
- all phone-phone combinations P P

All these were then considered at

- Beginning of the sentence
  //P P- //P- P
- End of the sentence P P//
  P- P//
- Beginning of word P P-
  P- P
- End of word P P#
  P- P#
- Middle of word
  - P P- -P- P
- Beginning of phrase
  /P P- /P- P
- End of phrase
  P P/ P- P/

These combinations yielded a total of 3,725 units which were to be used for sentences selection. 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.

5. Phone count in corpus vs selected sentences.

Of the 3,725 units set up for sentences selection, 1997 units were found in the corpus and therefore covered in the selected sentences. This meant that 1,728 could not be found. Out of the 10,558 sentences in the corpus, 414 sentences were selected as the minimum number of sentences to contain the possible units found in the corpus.

Phone count was then carried out both in the text corpus (10,558 sentences) and in the selected sentences (414 sentences). This is tabulated in Figure 1, while a plot of phone frequency of occurrence (number of times the phone appears in the set) is presented in Figure 2.

Phone	stressed	Count for stressed		% for stressed		Total phone count		% phone count	
		Corpus	Pbs	Corpus	Pbs	Corpus	Pbs	Corpus	Pbs
a	a	148,242	3,209	18.54	16.65	177,711	4,053	22.22	21.03
	a1	24,132	652	3.02	3.38				
	a2	5,337	192	0.67	1				
e	e	25,785	669	3.22	3.47	37,789	1,081	4.72	5.61
	e1	9,854	307	1.23	1.59				
	e2	2,150	105	0.27	0.54				
i	i	69,469	1,569	8.69	8.14	100,213	2,406	12.53	12.49
	i1	25,853	681	3.23	3.53				
	i2	4,891	156	0.61	0.81				
o	o	23,182	588	2.9	3.05	29,308	804	3.66	4.17
	o1	5,050	163	0.63	0.85				
	o2	1,076	53	0.13	0.28				
u	u	44,613	1,038	5.58	5.39	58,573	1,487	7.32	7.72
	u1	11,548	338	1.44	1.75				
	u2	2,412	111	0.3	0.58				
m	m	40,654	966	5.08	5.01	43,610	1,049	5.45	5.44
	m1	2435	68	0.3	0.35				
	m2	521	15	0.07	0.08				
n	n	36,184	874	4.52	4.54	39,346	978	4.92	5.08
	n1	2,666	82	0.33	0.43				
	n2	496	22	0.06	0.11				
b		14,394	356	1.8	1.85	14,394	356	1.8	1.85
ch		5,444	133	0.68	0.69	5,444	133	0.68	0.69
d		9,559	237	1.2	1.23	9,559	237	1.2	1.23
dh		2,513	75	0.31	0.39	2,513	75	0.31	0.39
f		7,605	181	0.95	0.94	7,605	181	0.95	0.94
g		2,954	103	0.37	0.53	2,954	103	0.37	0.53
gh		658	17	0.08	0.09	658	17	0.08	0.09
h		21,515	290	2.69	1.51	21,515	290	2.69	1.51
j		8,811	241	1.1	1.25	8,811	241	1.1	1.25
k		55,723	1,218	6.97	6.32	55,723	1,218	6.97	6.32
kh		200	21	0.03	0.11	200	21	0.03	0.11
l		25,060	672	3.13	3.49	25,060	672	3.13	3.49
ng1		92	11	0.01	0.06	92	11	0.01	0.06
ng		8,071	227	1.01	1.18	8,071	227	1.01	1.18
ny		4,763	154	0.6	0.8	4,763	154	0.6	0.8
p		7,876	228	0.98	1.18	7,876	228	0.98	1.18
r		11,980	288	1.5	1.49	11,980	288	1.5	1.49
s		15,978	415	2	2.15	15,978	415	2	2.15
sh		7,828	182	0.98	0.94	7,828	182	0.98	0.94
t		24,804	558	3.1	2.9	24,804	558	3.1	2.9
th		457	32	0.06	0.17	457	32	0.06	0.17
v		5,359	167	0.67	0.87	5,359	167	0.67	0.87
w		37,095	853	4.64	4.43	37,095	853	4.64	4.43
y		18,448	423	2.31	2.2	18,448	423	2.31	2.2
z		16,046	329	2.01	1.71	16,046	329	2.01	1.71
TOTAL		799,783	19,269	100.00	100.01	799,783	19,269	100.00	100.00

Figure 1: Phone count in Corpus vs the phonetically balanced sentences (pbs)

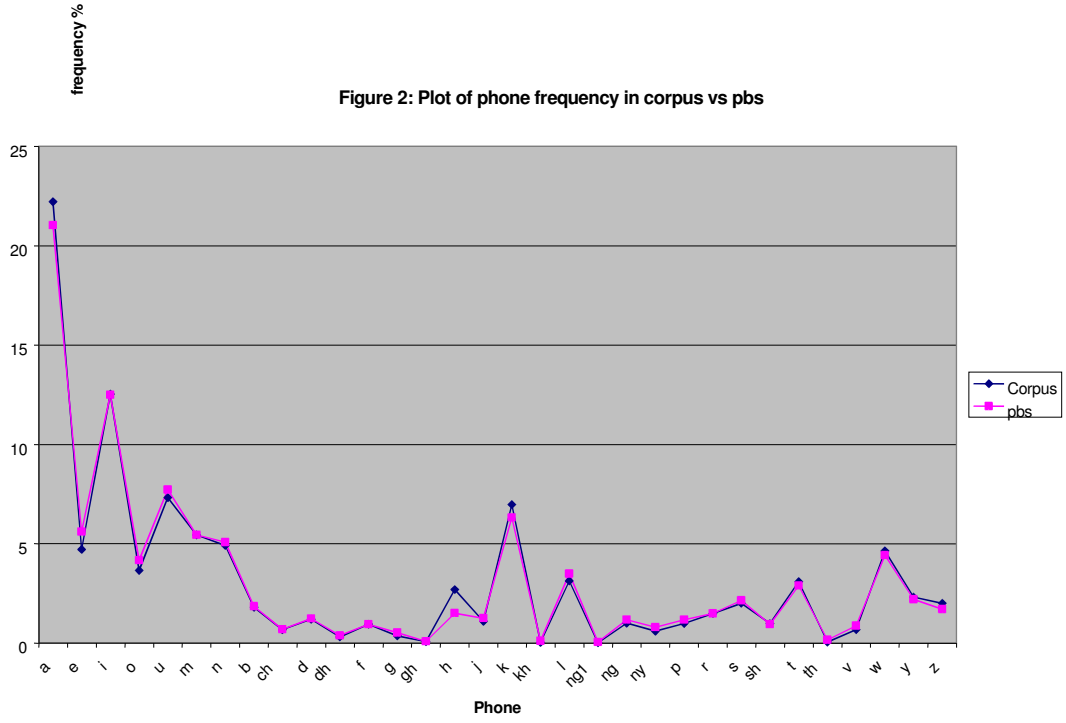

Figure 2: Plot of phone frequency in corpus vs pbs

6. *Results*

It is clear that there is almost 100% correlation between the phone frequencies of occurrence from the corpus and those of the selected sentences. Further the large corpus collected from several independent sources must have the phone distribution as used entirely in Kiswahili and therefore phonetically balanced. In that case the selection method as used yields the minimum number of phonetically balanced sentences that would be needed for setting up the text to speech system.

phones	phone count
Vowels	
a	22.22
i	12.53
u	7.32
e	4.72
o	3.66
TOTAL	50.45

phones	phone count
Consonants	
k	6.97
m	5.45
n	4.92
l	3.13
t	3.10
h	2.69
y	2.31
z	2.01
s	2.00
b	1.80
r	1.50
d	1.20
j	1.10
ng	1.01
p	0.98
sh	0.98
f	0.95
ch	0.68
v	0.67
ny	0.60
g	0.37
dh	0.31
gh	0.08
th	0.06
kh	0.03
ng'	0.01
TOTAL	49.55

Figure 3: Vowel-Consonant order of frequency of occurrence

Another interesting observation made about Kiswahili sounds is that the ratio of the vowels to consonants was 1:1. Considering that vowels are all voiced (vibration of vocal chords) and some consonants too (b, d, g, m, n etc.) the proportion of voiced sounds outweigh that of the unvoiced. As a result, Kiswahili sounds very melodious like Italian or French.

Conclusion

To synthesize speech efficiently, a system requires a phonetically balanced set of basic language sounds. The latter correlates almost perfectly with frequencies of occurrence of each phoneme in a huge corpus. With such findings, the foundation is now laid for efficient speech synthesis.

Further, the statistical data in figures 1, 2 and 3 would be critical in Automatic Speech Recognition (ASR) for Kiswahili, since they would significantly reduce the search space. In a nutshell, the information is vital both in speech synthesis and recognition.

The ratio of 1:1 with respect to vowels and consonants is significant, as it may explain why Kiswahili sounds melodious to the ear, compared to say German or English.

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.