Design of a Yoruba Language Speech Corpus for the Purposes of Text-to-Speech (TTS) Synthesis

Théophile K. Dagba[1(✉)], John O.R. Aoga[2], and Codjo C. Fanou[3]

[1] School of Applied Economics and Management,
University of Abomey-Calavi, Cotonou, Benin
theophile.dagba@eneam.uac.bj
[2] Polytechnic School of Abomey-Calavi,
University of Abomey-Calavi, Cotonou, Benin
johnaoga@gmail.com
[3] School of Administration,
University of Abomey-Calavi, Cotonou, Benin
chcodjo@yahoo.fr

Abstract. This paper deals with the design of a speech corpus for a corpus-based Text-To-Speech (TTS) synthesis approach. The purposes are first to provide enough speech to develop Yoruba corpus-based TTS system and second, to provide a simple methodology for other languages corpus design. The paper focuses on text analysis, selection of the reliable sentences, selection of the reader, and sentences recording. The analysis is performed to ensure a good balance of the corpus. Then, 2,415 sentences are gathered (essentially affirmative sentences). Those sentences have been read by a Yoruba language journalist who is a native speaker of the language. There is one speaker for the whole corpus.

Keywords: Yoruba language · Language corpus · TTS · Unit selection

1 Introduction

One of the crucial problems that have to be solved when speech recognition or a speech synthesis system is developed is the availability of a proper speech corpus for the system training and testing. The problem is usually solved in the following way: first, a set of suitable sentences are selected from a database of phonetically transcribed sentences; next the set of selected sentences are read by a group of speakers and, as the last step, the utterances are used to form the training and the test datasets [19]. Several works have been realized in the field of speech technology in general and TTS in particular. Among them we can mention those on Spanish [17], French [10], Czech [13], etc. The main obstacle to African languages in speech applications is the lack of sufficient speech material for the study of speech events and for training, development, and testing of algorithms and systems [10]. A 2013 review on prosody realization in Text-to-Speech applications showed that Yoruba is under-researched in the area of prosody and speech synthesis in general [5]. So far, Yoruba language has not

© Springer-Verlag Berlin Heidelberg 2016
N.T. Nguyen et al. (Eds.): ACIIDS 2016, Part I, LNAI 9621, pp. 161–169, 2016.
DOI: 10.1007/978-3-662-49381-6_16

experienced enough studies in speech corpus design for TTS synthesis. This paper provides an overview of ongoing research on the development of a Yoruba corpus. The goal of our work is to provide base material for the development and evaluation of TTS synthesis systems. We focus on the analysis, selection of text material and reader.

The paper is organized as follows. Sections 2 and 3 presents the Yoruba sound system and related works respectively. In Sect. 4 the methodology for designing Yoruba text corpus is presented. Section 5 deals with the conditions for recording a quality speech corpus. Section 6 is dedicated to our results and discussion. Finally, Sect. 7 contains the conclusion and outlines our future work in this field.

2 Yoruba Sound System

Yoruba is an African language of the family of Niger-Congo languages. It is natively spoken in southwestern Nigeria (the second largest ethnic group in number), Benin and Togo by over 30 million people [11]. There are three sets of sounds which make up Yoruba words: these are vowels, consonants and tones [4]. In Yoruba there are 12 vowels which are classified into 2 types, oral and nasalized vowels. Oral vowels are produced entirely through the mouth and nasalized ones are produced through both the mouth and the nose. Orthographically, nasalized vowels are written with an ‘n’ following an oral vowel. Yoruba has 18 consonants. It is a tonal language. It has three surface tones of different pitch levels. Tones are marked on vowels and syllabic nasals. The tones and their orthographic representations are as in Table 1 with the corresponding musical note[1].

Table 1. Yoruba language tones

Tone	Mark	Corresponding musical note
High	´	mi
Mid	Unmarked	re
Low	`	do

Indeed, a word may have different lexical meanings depending on whether it is said with a high, a mid or a low pitch. This shows the extent to which tones are important in Yoruba. The wrong pronunciation of a word could involve a wrong comprehension as illustrated in Table 2. Then, the tonal information removes the ambiguity in the pronunciation of well written and properly accented Standard Yoruba texts [15].

Table 2. Illustration of tone use on the vowel *o*

Word	Tone on the vowel	Meaning of the word
Kọ́	High	To build
Kọ	Mid	To sing
Kọ̀	Low	To refuse

[1] http://www.africa.uga.edu/Yoruba/phonology.html.

3 Related Work and Motivation

In the past few years, Yoruba TTS study has drawn a wide attention. TTS is then the area of speech technology that attracts more research effort. In 2004, Odéjobi et al. [16] have presented the design and analysis of an intonation model for Text-To-Speech synthesis applications using a combination of Relational Tree and Fuzzy Logic technologies. The model was demonstrated using Standard Yoruba language. In the proposed intonation model, phonological information extracted from text is converted into Relational Tree. Mean opinion Scores of 9.5 and 6.8, on a scale 1–10, was obtained for intelligibility and naturalness respectively. In 2011, a text markup system for text intended as input to standard Yoruba speech synthesis was presented by Odéjobi [15]. In 2012, van Niekerk and Barnard [23] have investigated the acoustic realization of tone in short continuous utterances in Yoruba. Fundamental frequency (F0) contours were extracted for automatically aligned syllables from a speech corpus collected for speech recognition development. Extracted contours were processed and analyzed statistically to describe acoustic properties in different tonal contexts. In 2013, Afolabi and Wahab [2] have focused their research work on the use of E-learning Text-To-Speech to teach Yoruba language online. A database was created for the recorded syllables in the tree tones of Yoruba language. In 2014, Akinadé and Odéjobi [3] examined the process underlying the Yoruba numeral system and described a computational system that is capable of converting cardinal numbers to their equivalent Standard Yoruba number name. In 2015, Adeyemo and Idowu [1] considered the development of TTS in Yoruba to assist Yoruba language speaking people especially the visually impaired users. They therefore created inventory of syllable pronounceable in Yoruba and recorded all of them. In other hand, Dagba et al. [7] investigated the integration of Yoruba into eSpeak[2] system for the purposes of mobile phone applications. They have defined 54 phonemes of Yoruba language by using existing phoneme tables such as Base table, English table and French table. They have built also rules which indicate how to pronounce certain groups of words. They finally have a dictionary file with 70 rules.

As shown by the above review, apart from the work of van Niekerk and Barnard [23] these studies are not corpus oriented or they relied on relatively small samples based on carefully designed corpora.

4 Design of Yoruba Text Corpus

4.1 Background of Corpus Building

The whole corpus building process is diagrammed as shown in Fig. 1 [21]. The criteria in designing speech corpus are size, coverage, domain and quality.

The recently developed corpus-based speech synthesizers tend to rely on large scale database, ranging from a few hours to more than 10 h of speech corpora, to provide

[2] http://espeak.sourceforge.net.

sufficiently natural output speech [12]. But an increase of the corpus size will affect the performance of the used method and slacken the synthesis process.

For unit selection synthesizer, the quality of speech is highly dependent on unit coverage of speech corpus. The corpus database must be phonetically rich. In other words, it must involve as many phonetic combinations as possible, including intra-syllabic and inter-syllabic structures, in a corpus of acceptable size [6]. The corpus words should also include at least one instance of all units [14].

As argued in [18], a system with a good selection module and a high quality speech corpus may yield output speech of extremely high quality, even if the signal processing module is rather simple.

The domain or focus application of a corpus-based Text-To-Speech is very important since a limited domain can reduce the corpus size and yet preserve the quality of synthetic speech. Several projects have been developed in restricted domains such as in weather forecasts and talking clock contexts [9, 14].

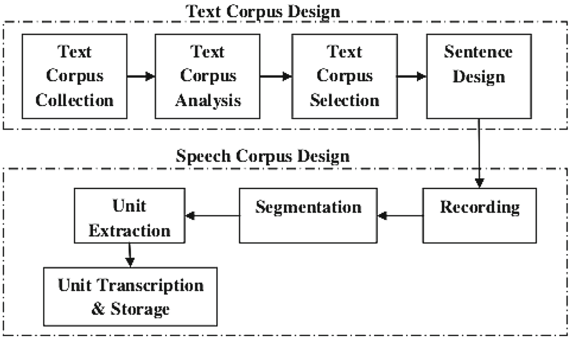

Fig. 1. Corpus building process

Our speech corpus construction requires the collection of texts written in Yoruba with well-spelled words. The reading of each sentence constitutes the audio corpus. The reader must respect the rules of pronunciation, tones and punctuation while adopting a consistent pace in a sound proof environment. Ideally, a recording studio is a suitable environment for this kind of recording. The speech corpus is a set of audio and text corpora, in the same folder, with a link between the text and the corresponding record.

4.2 Text Collection and Preprocessing

The collection of textual data was done from a Yoruba version of the Holy Bible[3]. The first 50 chapters of Genesis were taken into account in the construction of the corpus. However, after processing, some sentences were made entirely of personal names.

[3] http://www.jw.org/yo.

Those sentences have been deleted. Also some sentences of genesis 7, 13 are too long and have been deleted too. Paragraphs are extracted to expect overall consistency of the corpus. Because the meaning of sentences matters here, it is important to have correct sentences for easier reading and to allow the reader to be in a real and coherent context, and to help to enrich the speech corpus in emotions. Illustrations and tables are deleted as well as characters that are not taken into account such as +,−, *, %, etc. It is decided to use sentences as units of the corpus.

4.3 Text Analysis

This step allows the analysis of the text corpus at sentences level. In the first step, the statistics on the size of the corpus of sentences and words (number of words, number of distinct words, the average number of words in sentences, etc.) are produced. Then, the proportion of co-occurrence P(u,v) (see Eq. 1) is computed with f(u) (frequency of word u), f(v) (frequency of word v) and f(u,v) (frequency of word u and v occurring in the same sentence).

$$\mathrm{P(u,v)} = \mathrm{f(u,v)/(f(u)+f(v)-f(u,v))} \tag{1}$$

After that, the existence of different phonemes of Yoruba in every sentence and in the corpus is assessed. It is then ensured that there is no excessive difference between the frequencies of occurrence of different phonemes. In addition, most common contexts of use are represented. It is this tradeoff (between frequencies of phonemes and contexts) which defines the sound balance in the corpus. Finally, a K-means classification is performed based on the frequency of occurrence of words and phonemes to better appreciate the different lexical categories in the corpus. All the above analysis is repeated till acceptable tradeoff between frequencies of phonemes and utilization contexts is achieved.

5 Recording of the Sentences

After the design of text corpus, focus is placed on speech corpus design as illustrated in Fig. 1.

5.1 Choice of the Reader

To find the suitable reader, some criteria are used. First of all, the reader should be someone that practises the language in his/her daily life. A Yoruba language radio journalist who is also a native speaker has been selected. Thus, it is sure to have a voice respecting the rules of pronunciation and tones, but also prosodic parameters such as rhythm, intonation and emphasis. We took into account the playback speed because it is an important factor affecting the proper articulation of words. The recording is done in a recording studio preferably late at night.

5.2 Reading

This step is the recording of the text corpus. It is based on the use of Redstart of MaryTTS [20]. This tool allows us to calibrate the system on the recording settings prior to any playback/recording. Speech audio and timing parameters are given as shown in Table 3.

Table 3. Speech audio and timing parameters

Parameter	Value
Frequency	44,100 Hz
Number of bit	16
Input type	Mono
Timing before reading	2,000 ms
Timing after reading	2,000 ms
Pause	0 ms

The Redstart tool also allows listening to the sound, and re-recording or viewing the signal spectrum, pitch and energy diagram. A meticulous handling of the records is conducted. We listened to each sentence looking at the written version and the spectrum of the signal to verify that the sound is not clipped (Fig. 2) or misread. If one and/or another of the above cases of mistakes occur, the recording is repeated.

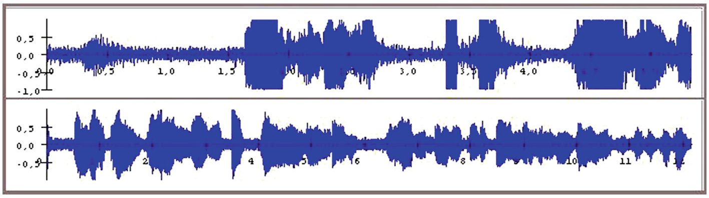

Fig. 2. Spectrum of a clipped signal (top) and a normal spectrum (bottom)

The audio tool convertor of MaryTTS [20] is used for the normalization of recordings in wave format to meet the conditions of the voice synthesis system. This tool also permits to address the overall amplitude and power of the recording sentence by sentence, to filter the noise frequencies below 50 Hz and remove start and end silence of wave sounds.

6 Results and Discussion

6.1 Results

The text corpus collected on the Internet, after analysis and balancing contains 2,415 sentences. Most of these sentences are affirmative sentences (88.65 %) with only 6.05 % of interrogative sentences and 5.30 % of exclamation sentences (see Table 4).

This corpus contains 46,117 words (2,275 distinct words). The average occurrence of words is 20.27 with a standard deviation of 93.22. This standard deviation reflects the unequal distribution of words in the corpus. This state is justified by the fact that words such as pronouns and prepositions appear more than 1,000 times in the corpus while common nouns, verbs, adjectives, and adverbs appear 100 times. Proper nouns and cardinal numbers appear less than 10 times. K-means classification confirmed the three categories of words that we had previously identified and the balance of phonemes.

Table 4. Features of the corpus

Item	Value
Sentences	2,415
Affirmative sentences	88.65 %
Interrogative sentences	6.05 %
Exclamation sentences	5.30 %
Average of word per sentence	11.38
Phonemes	148,823
Frequency of phonemes	2,705.87
Phonemes of high tone	24.48 %
Phonemes of mid tone	16.02 %
Phonemes of low tone	18.60 %
Phonemes of consonant	18.60 %
Size of the corpus	234 mn

6.2 Evaluation of the Corpus

To have an idea about the quality of this corpus, an experimental TTS corpus-based system using “unit selection algorithm” for Yoruba language is built by applying MaryTTS. The Mean Opinion Score (MOS) was used to evaluate the general output of the system. We have got the result as presented in Table 5. In subjective testing, a MOS is the arithmetic mean of all of the individual opinion scores resulting from a single test [8, 24]. Then, 10 native Yoruba speakers were selected. They were between 11 and 30 years old. Each person had listened to 10 synthesis sentences and had given a mark between 0 and 5. The MOS is equal to 2.9. This score is equivalent to a good perception of the voice in the system output. At this stage, we have integrated Yoruba localization into MaryTTS which is available on Gitub branch[4] of this tool. The next version of the tool will merge it with the master branch.

6.3 Discussion

We can first notice that the detailed methodology used to design our speech corpus can be used in similar work for other languages. Second, this corpus can be used in

[4] https://github.com/johnaoga/marytts.

Table 5. Result of MOS evaluation

Appreciation	Mark	Average score in %
Excellent	5	14
Very good	4	30
Good	3	20
Fair	2	14
Poor	1	12
Poor	0	10

synthesis systems. It can also be noticed that the contribution of linguistic analysis to ensure sound balance proved to be very useful. It has helped to know how to present all the phonemes in the corpus. This has also allowed us to better understand the constitution and the features of our corpus. Indeed, the texts of the corpus must be recorded by the same person who must be qualified to do this work. Studies on the possibility of combining heterogeneous voice sources may allow the use of a great mass of heterogeneous data. Those issues were previously mentioned in the literature [14, 22].

7 Conclusion

This paper deals with the design of a speech corpus for corpus-based Text-To-Speech synthesis approach. First, texts have been collected and analyzed. After that, we have proceeded to the recording of the sentences. We have obtained a speech corpus which contains 2,415 sentences with 148,823 phonemes. The corpus has been tested in an experimental TTS system with a good result. Our future work will increase the size of the corpus, and take into account more interrogative and exclamation sentences.

Acknowledgments. The authors acknowledge the contribution of Vincent AWE, radio journalist of Yoruba language at the Office of Radio and Television of Benin (ORTB), for the recording of the corpus.

References

1. Adeyemo, O.O., Idowu, A.: Development and integration of text to speech usability interface for visually impaired users in Yoruba language. Afr. J. Comput. ICT **8**(1), 87–94 (2015)
2. Afolabi, A.O., Wahab, A.S.: Implementation of Yoruba text-to-speech e-learning system. Int. J. Eng. Res. Technol. **2**(11), 1055–1064 (2013)
3. Akinadé, O.O., Ọdẹ́jọbí, O.A.: Computational modelling of Yorùbá numerals in a number-to-text conversion system. J. Lang. Model. **2**(1), 167–211 (2014)
4. Akinlabi, A.: Yorùbá sound system. In: Understanding Yoruba life and culture. Africa world press Inc. pp. 453–468. (2004)
5. Akinwonmi, A.E.: A prosodic text-to-speech system for yorùbá language. In: 8th IEEE International Conference for Internet Technology and Secured Transactions (ICITST), pp. 630–635, London (2013)

6. Chou, F.-C., Tseng, C.-Y., Lee, L.-S.: A set of corpus-based text-to-speech synthesis technologies for mandarin Chinese. IEEE Trans. Speech Audio Process. **10**, 481–494 (2002)
7. Dagba, T.K., Aoga, O.R., Fanou, C.C.: eSpeak support of Yoruba language for the purposes of mobile phone applications. In: 3rd IEEE Pan African Conference on Science Computing and Telecommunication (PACT'2015), pp. 137–141, Kampala (2015)
8. Dagba, T.K., Boco, C.: A text to speech system for Fon language using multisyn algorithm. Procedia Comput. Sci. **35**, 447–455 (2014). Elsevier
9. Fék, M., Pesti, P., Németh, G., Zainkó, C., Olaszy, G.: Corpus-based unit selection TTS for Hungarian. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2006. LNCS (LNAI), vol. 4188, pp. 367–373. Springer, Heidelberg (2006)
10. Gauvain, J.-L., Lamel, L., Eskenazi, M.: Design considerations and text selection for BREF, a large french read-speech corpus. In: 1st International Conference on Speech and Language Processing, vol. 2, pp. 1097–2000 (1990)
11. Igue, A.M.: Grammaire Yorùbá de base abrégée. Center for Advanced Studies of African Society (CASAS), monograph 238 (2009)
12. Kawai, H, Tsuzaki, M.: Study on time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis. In: Proceeding of the IEEE Workshop on Speech Synthesis, pp. 15–18 (2002)
13. Matousek, J., Psutka, J., Kruta, J. : Design of speech corpus for text-to-speech synthesis. In: Interspeech (Eurospeech), pp. 2047–2050 (2001)
14. Nagy, A., Pesti, P., Németh, G., Böhm, T.: Design issues of a corpus-based speech synthesizer. Hung. J. Commun. **6**, 18–24 (2005)
15. Odéjobí, O.A.: Design of a text markup system for Yorùbá text-to-speech synthesis applications. In: Conference on Human Language Technology for Development, pp. 74–80, Alexandria, Egypt (2011)
16. Odéjobí, O.A., Beaumont, A.J., Wong, S.H.S.: A computational model of intonation for Yorùbá text-to-speech synthesis: design and analysis. In: Sojka, P., Kopeček, I., Pala, K. (eds.) TSD 2004. LNCS (LNAI), vol. 3206, pp. 409–416. Springer, Heidelberg (2004)
17. Ortega-Garcia, J., Gonzalez-Rodriguez, J., Marrero-Aguiar, V.: AHUMADA: a large speech corpus in Spanish for speaker characterization and identification. Speech Commun. **31**(2), 255–264 (2000)
18. Piits, L., Mihkla M., Nurk T., Kiissel, I.: Designing a speech corpus for Estonian unit selection synthesis. In: Proceedings of the 16th Nordic Conference of Computational Linguistics NODALIDA-2007, pp. 367–371 (2007)
19. Radová, V., Vopálka, P.: Methods of Sentences Selection for Read-Speech Corpus Design. In: Matoušek, V., Mautner, P., Ocelíková, J., Sojka, P. (eds.) TSD 1999. LNCS (LNAI), vol. 1692, pp. 165–170. Springer, Heidelberg (1999)
20. Schröder, M., Trouvain, J.: The German text-to- speech synthesis system MARY: a tool for research, development and teaching. Int. J. Speech Technol. **6**, 365–377 (2003)
21. Tan, T.-S., Hussain, S.: Scorpus design for Malay corpus-based speech synthesis system. Am. J. Appl. Sci. **6**(4), 696–702 (2009)
22. Taylor, P.: Text-to-Speech Synthesis. Cambridge University Press, Cambridge (2009)
23. Van Niekerk, D.R., Barnard, E.: Tone realisation in a Yoruba speech recognition corpus. In: 2012 Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), Cape Town, South Africa. http://www.mica.edu.vn/sltu2012/files/proceedings/11.pdf
24. Viswanathan, M., Viswanathan, M.: Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. Comput. Speech Lang. **19**(1), 55–83 (2005)