Innovative Systems Design and Engineering
ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online)
Vol.8, No.5, 2017

Development of a Yorúbà Text-to-Speech System Using Festival

Abimbola Rhoda Iyanda* Olufemi Deborah Ninan
Computer Science and Engineering Department, Faculty of Technology, Obafemi Awolowo University

Abstract
This paper presents a Text-to-Speech (TTS) synthesis system for Yorúbà language using the open-source Festival TTS engine. Yorúbà being a resource scarce language like most African languages however presents a major challenge to conventional speech synthesis approaches, which typically require large corpora for the training of such system. Speech data were recorded in a quiet environment with a noise cancelling microphone on a typical multimedia computer system using the Speech Filing System software (SFS), analysed and annotated using PRAAT speech processing software. Evaluation of the system was done using the intelligibility and naturalness metrics through mean opinion score. The result shows that the level of intelligibility and naturalness of the system on word-level is 55.56% and 50% respectively, but the system performs poorly for both intelligibility and naturalness test on sentence level. Hence, there is a need for further research to improve the quality of the synthesized speech.
Keywords: Text-to-Speech, Festival, Yorúbà, Syllable

1. Introduction

Text-to-Speech (TTS) is the process which allows the automatic production of speech, through a grapheme-to-phoneme transcription of the sentences to utter. Graphemes are the letters in a words dictionary listing whilst Phoneme is the smallest unit of speech that differentiates one word from another. Text-to-Speech is an important technology in human-computer interaction (HCI). HCI increases the possibilities of improved man-machine interaction and enhances other information systems.

Speech synthesis and speech recognition are two of the most important technologies deployed in HCI. The speech recognition technology allows a human speaker to give verbal commands to a computer system. Speech synthesis technology, on the other hand, allows a computer system to generate spoken responses to human requests or commands. The combination of these two technologies into the computer user interface provides a number of advantages. One advantage is that their application reduce the amount of formal training required for operating computer systems, which is of importance to African countries in that a large percentage of the population do not have access to formal education (Odejobi, 2007).

At the moment, many speech synthesis systems are available mainly for very few languages. Languages addressed are mainly those with a large population, a high economic power, or for which a high political interest exists. However, the majority of the languages in the world lack such technologies, and researches in the area are quite few. Recently, many localization projects are being undertaken for African languages, but not adequate because of the lack of linguistic resources absence of similar works in the area. This paper describes TTS synthesis for standard Yorúbà language.

1.1 Standard Yorúbà Language

Yorúbà language is one of the three major indigenous languages, along with Hausa and Igbo in Nigeria and it is spoken by over 37 million people (CIA, 2014). Yorúbà is a native language of the Yorúbà people, an ethnic group primarily located in south-western Nigeria (Lagos, Oyo, Ogun, Ondo, Ekiti, Osun and parts of Kwara and Kogi states). Standard Yorúbà (SY) language is used in language education, the mass media and everyday communication. The SY alphabet has 25 letters which is made up of 18 consonants which represented graphemically *by b, d, f, g, gb, h, j, k, l, m, n, p, r, s, ṣ, t, w, y* and seven vowels represented graphemically by *a, e, ẹ, i, o, ọ, u* (Adewole, 1988). It should be noted that the *gb* is a diagraph i.e. a consonant written with two letters. There are five nasalised vowels in the language represented graphemically as *an, ẹn, in, ọn, un* and one syllabic nasals represented graphemically as *n*. It should also be noted that an and *ọn* are phonemically the same.

The pronunciation of the letters without diacritics corresponds more or less to their International Phonetic Alphabet (IPA) equivalents, except for the labialvelar stops [$\widehat{kp}$] (written as < *p* >) and [$\widehat{gb}$] (written as < *gb* >), in which both consonants are articulated simultaneously and not sequentially. The consonant and vowel systems as well as a more detailed description of SY are presented in Iyanda (2014).

1.2 SY Texts

Although, ideally, an SY text suitable for processing in the context of TTS system is written using the orthography discussed in section 1.5, but this is not always the case in practice. In some cases, written text is fully tone marked and all the under-dots are indicated as shown in Figure 1a (scanned from (Onayemi, 2010)) while in other cases, there are texts that are tone marked as well as under-dots at various degree of completeness

as shown in Figure 1b. In Figure 1a, *kọ́* in line 1 should be written as *koọ́*. It is almost not possible to see SY text that is not tone marked at all and also the under-dots are not indicated. In the latter case, it is very difficult to read and understand the meaning of sentences. The contents of the written texts depend solely on the authors' ability to use the language as well as familiarity with the orthography. In other sense, the use of diacritics also depends on the orientation of target readers as well as the purpose of writing (Odejobi, 2005).

The use of diacritics in SY text could pose problems because most word processors (such as Microsoft word) could not help in generating the required tone marks and under-dots, although the software can be manipulated to indicate tone marks on some vowels. For example, CTRL+´ will give low tone (as in *à*) while CTRL+` will produce high tone (as in *á*). This challenge was addressed with the use of 'LATEX' except that it can be laborious for an average user to learn and use in a short while.

To address the issue of ambiguity that can be produced by partially marked and un-marked SY texts, fully tone marked and well under-dotted SY text was used in this study. This enhances the computation of pronunciation.

1.3 Standard Yorúbà Syllable

Syllables are the smallest unit of pronunciation in a language (Akanbi and Odejobi, 2011). In SY language, the syllables are made up of oral vowels (V), nasal vowels (Vn), syllabic nasals (N), combination of consonants and oral vowels (CV) as well as combination of consonants and nasal vowels (CVn) (Odejobi, 2007). The syllables are the tone bearing elements of the SY language. For example, in the statement: *Abímbọ́lá ti lọ sí oko* (*Abímbọ́lá* has gone to the farm), there are ten syllables. This also implies that there exist ten tones (MHMHH-M-M-H-MM) (Iyanda et al., 2014).

2. Speech Sound

In tone languages, tones (characterised by the variation of pitch within syllable) are lexically significant. Therefore, tonal information is essential for speech recognition in such languages (Demeechai and Makemainen, 2000). Tone in Yorúbà is used to distinguish between words having different meanings, but which otherwise is phonemically identical. Words may have the same basic spelling, but the tone with which they are pronounced dictates the meaning of each word and how it sounds. Phonologically, Yorúbà has three contrastive tones (Fajobi, 2000): High (H) which is represented by an acute accent (´); Low (L) which represented by a grave accent (`) and Mid (M) which is unmarked except on syllabic nasals where it is marked with a dash (-) in some cases. The syllable is the domain of tone.

It is important to identify that the understanding of tone is crucial to the understanding of the language. H that occurs in word-initial position only marked consonant-initial words, which reveals an implicit initial vowel when preceded by another word in genitive construction. In standard Yorúbà (SY), most word start with a vowel in which L or M occurs but not H (Akinlabi and Liberman, 2000). Apart from the tone, other set of sounds from which Yorúbà words are built are Yorúbà alphabet (consonants and vowels). The phoneme that occurs in tone languages uses tone to convey differences in lexical meaning. E.g. in SY language, a bisyllabic word *ọkọ́* (hoe), *ọ̀kọ̀* (spear), *ọkọ* (husband), *ọkọ̀* (vehicle) are phonemically identical but correspond to different lexical meaning.

3. Related Work

Odejobi (2007) presented a quantitative model of Yorúbà speech intonation using stem-ML. The model is built and trained on speech data from a native speaker of SY and the resulting model reproduces the data well with its Root Mean Square prediction error (RMSE) of 14 Hz on the test set. It was found that intonation was used to mark sentence and phrase boundaries with beginning syllables as systematically stronger, while ending syllables as systematically weaker than the medial syllables. It also established that M tone has the highest strength followed by the L tone and that the H tone is the weakest. The resulting model for SY shows similar characteristics when compared to Mandarin and Cantonese intonation models.

Alam et al. (2011) presented a Text to Speech (TTS) synthesis system for Bangla language using the opensource Festival TTS engine. Festival is a complete TTS synthesis system, with components supporting front-end processing of the input text, language modeling, and speech synthesis using its signal processing module. The Bangla TTS system proposed here, creates the voice data for festival, and additionally extends festival using its embedded scheme scripting interface to incorporate Bangla language support. Two different concatenative methods: unit selection and multisyn unit selection was used for the implementation. The quality of synthesized speech of the TTS system was evaluated using acceptability and intelligibility metrics of mean opinion score.

Damper et al. (2002) presented a paper on pronunciation analogy module for the festival TTS synthesiser. The paper provided theoretical and empirical motivations for the use of Pronunciation by Analogy (PbA), reviewed approaches to automatic pronunciation generation by analogy as well as the implementation of a PbA module. The system uses the dictionary which provides the primary source of pronunciation of unknown words.

Results showed that the data driven techniques outperformed experts' rules by a significant margin even though the data driven methods required aligned text-phoneme datasets and the alignment process was problematic. The best translation results were obtained with PbA at approximately 86.7% words correct.

Fukada et al. (1999) proposed a model for automatically generating a pronunciation dictionary based on pronunciation neural network using words whose quintphone contexts are the same. The work focused on two techniques: (i) realized pronunciations with likelihoods based on the neural network outputs and (ii) realized pronunciations for word boundary phonemes using word biagram-based language statistics. It was shown that automatically derived pronunciation dictionaries gave higher recognition rates than a conventional dictionary.

Odejobi (2005) presented a computational model of prosody for Yorúbà language in the context of computer text-to-speech synthesis applications. The framework was implemented using the relational tree techniques and the intonation dimension was implemented by developing fuzzy control rules based on data from native speakers of Yorúbà. The approach used provides a flexible and extensible model which as well can be used to implement, study and explain the theory behind the aspects of the phenomena observed in speech prosody.

Ekpeyong et al. (2014) presented a statistical based speech synthesis for Ibibio. This method was found to offer good performance on small corpora since it can directly learn the relationship between acoustic and the available linguistic features and can as well reduce the absence of explicit representation of intermediate linguistic layers such as prosody. This work established that the use of tone marking contributes significantly to the quality of synthetic speech and thereby proposed that the problem of tone assignment be addressed using a dictionary and the building of a prediction module for out-of-vocabulary words. Mean opinion score was used as the evaluation metric.

Kayte et al. (2015) presented a Text to Speech system for Maharashtra Marathi that can convert a Unicode encoded Marathi text into human speech using the open source Festival TTS engine and discussed a few practical applications that use the system. This system is developed using di-phone concatenation approach in its waveform generation phase. Finally, a test was conducted to evaluate the intelligibility of the synthesized speech using modified rhyme test (MRT). Based on the test, the overall intelligibility of the system from 6 listeners is 96.96%. Also, a unit test on text normalizer and G2P converter was performed. The performance of text normalizer is 87% only for ambiguous tokens and that of G2P converter is 89%.

4. G2P System Design

The G2P component of the proposed system was designed using FST. Figure 2 shows the FST general model designed in Java Formal Language and Automata Package (JFLAP) environment and Figure 3 shows the model that provides specific instance of its application. The word *àgbàlagbà* (adult) was passed into the model. At q_0, `a is passed into the model which result into a(L); at $q_2$, *gb* is passed into the model which result into *gb*; at $q_1$, `a is passed into the model which result into a(L); at q_2, l is passed into the model which result into l; at q_1, a is passed into the model which result into a(M); at q_2, *gb* is passed into the model which result into *gb* and finally, at q_1, `a is passed into the model which result into a(L), yielding *a*(L)*gba*(L)*la*(M)*gba*(L).

At q_0, the system goes to q_1 if the input is oral or nasal vowel; q_2 if the input is a consonant; q_3 if the input is a syllabic nasal. At q_1, the system remains in q1 if the input is oral or nasal vowel; goes to q_2 if the input is a consonant; q_3 if the input is a syllabic nasal. At q_2, the system goes to q_1 if the input is oral or nasal vowel. At q_3, the system goes to q_2 if the input is a consonant. In this model, the following abbreviations were used:

V = Oral vowels {*a, e, e., i, o, o., u*}
Vn = Nasal vowels { *an, ẹn, in, ọn, un*}
N = Syllabic nasal {*n*}
Vn = Consonants {*b, d, f, g, gb, h, j, k, l, m, n, p, r, s, s., t, w, y*}

5. Yorùbá Text-to-Speech in Festival

Festival is a multi-lingual TTS engine and a general purpose concatenative TTS system that offers a general framework for building speech synthesis systems as well as supporting all language processing tasks such as document analysis, text analysis and phonological processing. The Festival framework has been used extensively by the research community in speech synthesis. It has been chosen for implementing the Yorúbà TTS system because of its flexible and modular architecture, ease of configuration, and the ability to add new external modules.

The language processing modules in Festival are not adequate for certain languages and the reliance on Scheme as a scripting language makes it difficult for linguists to incorporate the necessary language specific changes within Festival. Thus, new tools or modules to be plugged into Festival are needed. Yorúbà being a new language in this framework requires substantial amount of work, especially in the text processing modules.

Festival is designed as a speech synthesis system for at least three levels of user: (i) those who simply want high quality speech from arbitrary text with minimum effort; (ii) those who are developing language systems and

wish to include synthesis output- in this case, a certain amount of customization is desired, such as different voices, specific phrasing, dialog types etc.; (iii) in developing and testing new synthesis methods (Black et al., 1999).

5.1 Speaker selection and recording of the database

When selecting a speaker, it has been established that professional speakers, who are generally aware of the language features such as phonemes, are often better than non-professional speakers (Black et al., 1998). A good voice is robust to small distortions of synthesized speech and also gives good results in the detailed phonetic annotation. Therefore, professional speakers that are fluent in Yorúbà was used for the recording in a quiet environment. The files were saved as .wav. In the recording of the database, educational level, language background as well as the number of languages spoken by the speakers were considered. Recording time for the whole database was taken.

6. Methodology

Yorúbà phoneset consists of all the combination of vowels and consonants in Yorúbà alphabet using the possible syllable structures.

i. Recording: Yorúbà phoneset was used to create a diphone list of 740 prompts in the diphone database. Each line in the diphone database contains a file id, a prompt, and a diphone name. The file id is used to store the filename for the waveform, label file, and any other parameters files associated with the word. The lists were read by two males and two females who are native speakers of SY with age of the speakers ranging from 22 to 37 years old. Each speaker read the text at their own pace, resulting in the average number of syllable per seconds ranging from two to three. The sounds were prepared by recording at a 44100 Hz sampling rate and 32-bit quantization. After that, they were down-sampled to 16 kHz for analyzing and used in the implementation of the system in Festival TTS engine. All speech units were recorded with normal speaking rate.

ii. Recording equipment: The corresponding speech data for the Yorúbà syllables collected was recorded in a quiet environment with a noise cancelling microphone on a typical multimedia computer system using the Speech Filing System (SFS) software. Adobe Audition 3.0. was used to normalise and to reduce adaptive noise. SY which is the one being used in education, the mass media and everyday communication was chosen for the recording. The recorded speech data was analysed and annotated using PRAAT speech processing software.

iii. Speech file annotation: Each file of the recording was loaded into PRAAT and annotated manually. For the annotation, TextGrid is created in PRAAT for each speech waveform file. The TextGrid and waveform files are selected for annotation and editing. There are two tiers in the annotation: word and syllable. The labelling is done to identify syllabic boundaries. In the annotation of the syllable speech files, only one tier is specified (the syllable tier). The syllable is labelled with its associated tone. In the annotation of the word speech files, two tiers are specified (the syllable and word tiers). Both the spectrogram and the waveform are used in determining syllable and word boundaries.

7. Implementation of Pronunciation System in Festival

Festival Speech Engine was designed to only accept ASCII characters as input. Yorúbà contains a lot of none ASCII characters such as the diacritics (grave, accent and macron). To accept the Yorúbà words as input, a preprocessor was developed in python which accepts the non ASCII characters and converts them to ASCII characters that are absent in the Yorúbà alphabet. The mapping of the characters is shown in the Table 2. ‘v’ represents the grave and ‘c’ represents the accent.

1. Phoneset Definition: The Yorúbà Phoneset was defined in the file ‘oaucisrg yor voice phoneset.scm’ using phone features. For the vowel- vowel height, vowel frontness and lip rounding are used; while for the consonantconsonant type (stop, fricative, affricative, nasal, liquid), place of articulation (labial, alveolar, palatal, labio-dental, dental, velar) as well as consonant voicing are used. The phone can either be vowel (+) or consonant (-); the vowel length can be s, l, d, a, 0 (representing short, long, dipthong, schwa or none); the vowel height can be 1, 2, 3, 0 (representing high, mid, low, or none); the vowel frontness can be 1, 2, 3, 0 (representing front, mid, back, or none); the lip rounding can be +, -, 0 (representing rounded, not rounded or none); the consonant type can be s, f, a, n, l, r, 0 (representing stop, fricative, affricative, nasal, liquid, or none); the place of articulation can be l, a, p, b, d, v, g, 0 (representing labial, alveolar, palatal, labio-dental, dental, velar, or none), the consonant voicing can be +, -, 0 (representing voiced, voiceless, or none). For example, ”a” is a vowel (+), with no vowel length (0), low vowel height (3), mid vowel frontness (2), no lip rounding (-), no consonant type (0), no place of articulation (0) and no consonant voicing (0).

2. Token Processing rule: The token processing rules are defined in ‘oaucisrg yor voice tokenizer’. This maps each token to a list of words. This function is called on each token and returns a list of words. It is set in

the voice selection function as the function for token analysis.

3. Diphone Database: The Diphone database contains 740 prompts which are:
   i. Phonetically balanced
   ii. Targeted toward the intended domain
   iii. Easy to say by the speakers without mistakes iv. Short enough for the speakers to be willing to say it

The text was stored in a file 'txt.done.data' with the corresponding wave files stored as '.wav' with the same file identification no(id). The diphone database contains all possible phone-phone transitions in Yorúbà. This was generated by Festival's scheme interpreter 'yorschema.scm' which contains methods for generating all the set of phone-to-phone transitions in Yorúbà language after given the syllabic structure. The prompts were then presented to the speaker at recording time.

Festival uses the attributes of each entry to synthesize its pronunciation. The large set of lexicon was implemented based on Yorúbà syllabic structure (V, Vn, N, CV, and CVn). An example of lexicon format in Festival is ((“paupau”) (pau t a b a b a pau pau))) which produced tababa. Figure 4 shows a sample spectrogram of the speech samples.

4. Letter-to-Sound Rule: This is defined in 'oaucisrg yor voice lexicon.scm' and it contains 50 Letter-to-Sound rules for Yorúbà. A pronunciation requires not just a list of phones but also a syllabic structure. For Yorúbà, there is a systematic relationship between the written form of a word and its pronunciation i.e. one-to-one correspondence between words and their pronunciation. The basic form of the rules is (LC[alpha]RC = beta) Which is interpreted as alpha, a string of one or more symbols on the input tape is written to beta, a string of zero or more symbols on the output tape, when in the presence of LC, a left context of zero or more input symbols, and RC a right context on zero or more input symbols. Note the input tape and the output tape are different, although the input and output alphabets need not be distinct the left hand side of a rule only can refer to the input tape and never to anything that has been produce by a right hand side.

5. Prosodic Modelling: This involved the following:
   i. Phrasing: This uses a CART tree. A test is made on each word to predict if it is at the end of a prosodic phrase. The basic CART tree returns B or BB. The two levels identify different levels of break, BB being used to denote a bigger break (and end of utterance). The tree is defined in 'phrasing.scm'
   ii. F0 Generation: This involves predicting where accents go (and their types). We also have built an f0 contour based on these. Note intonation is split between accent placement and f0 generation as it is obvious that accent position influences durations and an f0 contour cannot be generated without knowing the durations of the segments the contour is to be generated over.

6. Waveform Synthesis: This involves creating the speech waveform from a complete phonetic and prosodic description. It involves the following:
   i. Speech Labeling: This information helped the system in phoneme segmentation. Two basic methods have commonly been applied to provide the initial boundary estimates (VanNiekerk, 2007):
      a. Hidden Markov Models (HMMs) applied in forced alignment. In this approach, distinct models are trained for each phone in the target language. It is preferred in the speech recognition application domain, where HMMs are generally applied.
      b. Dynamic Time Warping (DTW) of the target signal to match a similar signal of which the boundaries are known. This involves matching segments with similar acoustic properties and it is Popular in the speech synthesis (TTS) application domain, where a reference signal can be synthesised.

DTW, an automatic segmentation was used in this study to achieve phoneme segmentation. It is used for aligning some new recording with some known one. The prompts were labelled using DTW. The aligner uses those prompts to do the aligning. The idea behind the aligner is to take the prompt and the spoken form and derive mel-scale cepstral parameterizations (and their deltas) of the files, then a DTW algorithm is used to find the best alignment between these two sets of features.

7. Extracting the pitchmarks: This is done using Linear-Predictive Coding (LPC) resynthesis. The prompts were recorded with an electroglottograph (EGG, also known as a laryngograph) at the same time as the voice signal. The EGG records electrical activity in the glottis during speech, which makes it easier to get the pitch moments, and so they can be more precisely found. The pitch mark extraction method is defined in 'make pm wave'.

8. Evaluation of the System

The aim of the evaluation is to compare the performance of the developed system with that of experts. The evaluation of the system was done using the intelligibility and naturalness metrics through mean opinion score (MOS) which implements the Turing test for machine intelligence.

Selected group of Yorúbà native speakers were asked to rate how much of the synthetic speech is identifiable and how pleasant are they to their listening. The test data consists of twenty words and nine

statement sentences which are composition of words that were not in our database as shown in Table 3. This is done to see how the developed system can extrapolate between known and unknown data. To evaluate naturalness, a comparative study was made between sound produced by the original speech samples and the approximated speech obtained by using MOS that is expressed as a single number in the range 1 to 5, where 1 is lowest perceived quality and 5 is the highest perceived quality. The MOS is generated by averaging the results of a set of standard subjective tests where a number of listeners rate the perceived audio quality of speech produced by the system. Ten Yorúbà native speakers with age ranging between 22 and 37 years old were selected to listen to the sounds produced by the system and each listener were required to give each sound a rating. The sound produced was also checked for intelligibility (i.e. how meaningful the sound is).

9. Conclusion

The implementation of a Yorúbà TTS system has been described. The speech produced by the system is averagely intelligible, but the measure of naturalness is low. Improvement of intelligibility and naturalness of speech depend on significant amount of work in each phase of the TTS system, and there is a need to address this before there can be a complete and quality TTS system such as those available for many other languages. The Baum-Welch algorithm for HMM can be a better technique for the phoneme segmentation because it has been shown that this algorithm improves the accuracy of automatic phoneme segmentation (Huggins-Daines and Rudnicky, 2006). Further research is thereby proposed on this to address the issues of tone assignment and to improve the quality of the synthesized speech.

References

L. O. Adewole. The Categorical Status and Function of the Yoruba Auxiliary Verb with some Structural Analysis in GPSG. PhD thesis, University of Edinburgh, Edinburgh, 1988.

L. A. Akanbi and O. A. Odejobi. Automatic Recognition of Oral Vowels in Tone` Language: Experiments with Fuzzy Logic and Neural Network Models. Applied Soft Computing, 11(1):1467–1480, 2011.

A. Akinlab´ı and M. Liberman. The Tonal Phonology of Yorúbà Clitics. Most, pages 1–32, 2000.

F. Alam, P. K. Nath, and M. Khan. Text to speech for bangla language using festival. Conference on Human Language Technology for Development (HLDT), pages 128–133, 2011.

A. W Black, P. Taylor, and R. Caley. The festival speech synthesis system. available from http://www.cstr.ed.ac.uk/projects/festival.html. pages 77– 80, 1998.

A. W Black, P. Taylor, and R. Caley. The festival speech synthesis system. available from http://www.cstr.ed.ac.uk/projects/festival.html. 1999.

CIA. Cia world factbook 2014. Available online at http://www.cia.gov/cia/publications/factbook., 2014. Visited April, 2014.

R. I. Damper, C. Z. Stanbridge, and Y. Marchand. A pronunciation-by-analogy module for the festival text-to-speech synthesiser. In in 4th ISCA Workshop on Speech Synthesis, August/September 2001, pages 97–102, 2002.

T. Demeechai and K. Makelainen. Integration of tonal knowledge into phonetic hmms for recognition of speech in tone languages. Signal Processing, 80(10):2241–2247, 2000.

M. Ekpenyong, E. Urua, O. Watts, S. King, and J. Yamagishi. Statistical parametric speech synthesis for Ibibio. Speech Communication, 56:243251, 2014.

E. Fajobi. The nature of Yorúbà intonation: A new experimental study. Linguistics, (1987):1–32, 2000.

T. Fukada, T. Yoshimura, and Y. Sagisaka. Automatic generation of multiple pronunciations based on neural networks. Speech Communication, (27):63–73, 1999.

D. Huggins-Daines and A. I. Rudnicky. A Constrained Baum-Welch Algorithm for Improved Phoneme Segmentation and Efficient Training. In INTERSPEECH. ISCA, 2006.

S. Kayte, M. Mundada, and C. Kayte. A Text-to-Speech Synthesis for Marathi Language using Festival and Festvox. *European Journal of Computer Science and Information Technology*, l.3(5): 30-41, 2015.

A. R. Iyanda. Design and Implementation of a Grapheme-to-Phoneme Conversion System for Yorúbà Text-to-Speech Synthesis. PhD thesis, Obafemi Awolowo University, Ile-Ife, Nigeria, 2014.

A. R. Iyanda, O. A. Odejobi, F. A. Soyoye, and O. O. Akinade. Development of Grapheme-to-Phoneme Conversion System for Yorúbà Text-to-Speech Synthesis. INFOCOMP, 13(2), 2014.

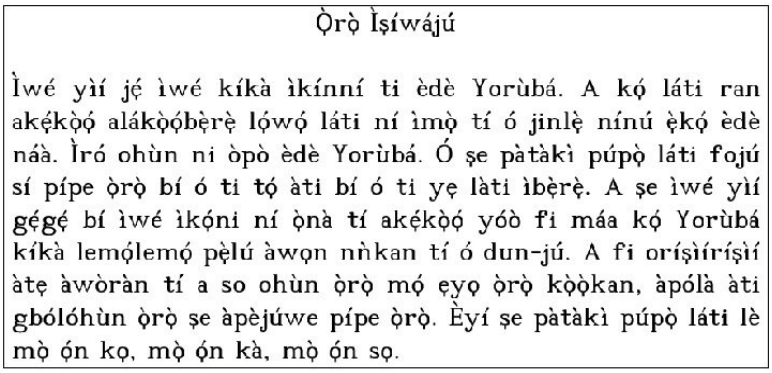
O. A. Odejobi. A Computational Model of Prosody for Yorúbà Text-to-Speech Synthesis. PhD thesis, Aston University, 2005.

O. A. Odejobi. A quantitative model of Yorúbà speech intonation using stem-ml. INFOCOMP Journal of Computer Science, 6(3):47–55, 2007.

A.O. Onayemı. Yorúbà Know How To Write It, Read It, Speak It. Book One, 2nd Edn. Yorúbà Readers' Club International, 2010.

D. R. VanNiekerk. Automatic Approaches to Speech Segmentation. Meraka Institute: African Advanced Insitute

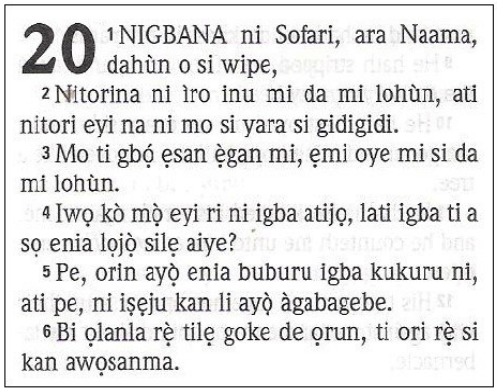
for Information & Communcation Technology, http://www.meraka.org.za/nhn/Members/meraka/student-poster-day-20june-2007/daniel vanniekerk.pdf/download. Visited: June 2014., 2007.



Ọrọ Ìṣíwájú

Ìwé yìí jẹ́ ìwé kíkà ìkínní ti èdè Yorùbá. A kọ́ láti ran akẹ́kọ̀ọ́ alákọ̀ọ́bẹ̀rẹ̀ lọ́wọ́ láti ní ìmọ̀ tí ó jinlẹ̀ nínú ẹ̀kọ́ èdè náà. Ìrò ohùn ni ọ̀pọ̀ èdè Yorùbá. Ó ṣe pàtàkì púpọ̀ láti fojú sí pípe ọ̀rọ̀ bí ó ti tọ́ àti bí ó ti yẹ làti ìbẹ̀rẹ̀. A ṣe ìwé yìí gégẹ́ bí ìwé ìkọ́ni ní ọ̀nà tí akẹ́kọ̀ọ́ yóò fi máa kọ́ Yorùbá kíkà lemọ́lemọ́ pẹ̀lú àwọn nǹkan tí ó dun-jú. A fi orísìírísìí àtẹ àwòràn tí a so ohùn ọ̀rọ̀ mọ́ ẹyọ ọ̀rọ̀ kọ̀ọ̀kan, àpólà àti gbólóhùn ọ̀rọ̀ ṣe àpèjúwe pípe ọ̀rọ̀. Èyí ṣe pàtàkì púpọ̀ láti lè mọ̀ ọ́n kọ, mọ̀ ọ́n kà, mọ̀ ọ́n sọ.

(a) Fully tone marked SY texts



20 1 NIGBANA ni Sofari, ara Naama, dahùn o si wipe,
2 Nitorina ni ìro inu mi da mi lohùn, ati nitori eyi na ni mo si yara si gidigidi.
3 Mo ti gbọ́ ẹsan ẹ̀gan mi, ẹmi oye mi si da mi lohùn.
4 Iwọ kò mọ eyi ri ni igba atijọ, lati igba ti a sọ enia lọjọ́ silẹ aiye?
5 Pe, orin ayọ̀ enia buburu igba kukuru ni, ati pe, ni ìṣẹ̀ju kan li ayọ̀ àgabagebe.
6 Bi ọlanla rẹ̀ tilẹ̀ goke de ọrun, ti ori rẹ̀ si kan awọsanma.

(b) Partially tone marked SY texts

Figure 1. Scanned sample of texts in SY orthography

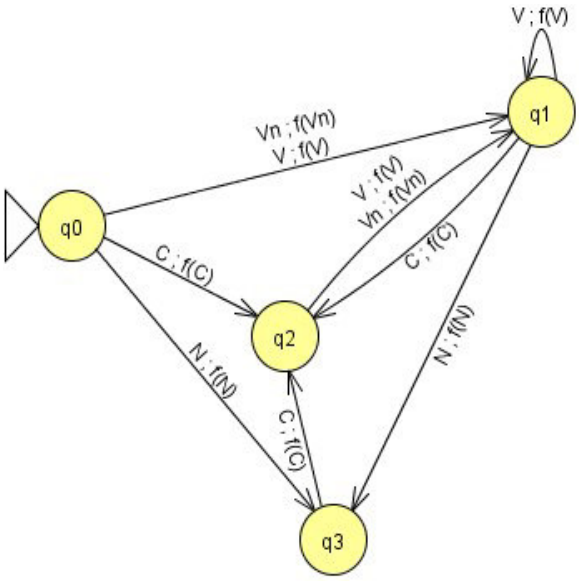


Figure 2. FST model for *Yorúbà* G2P

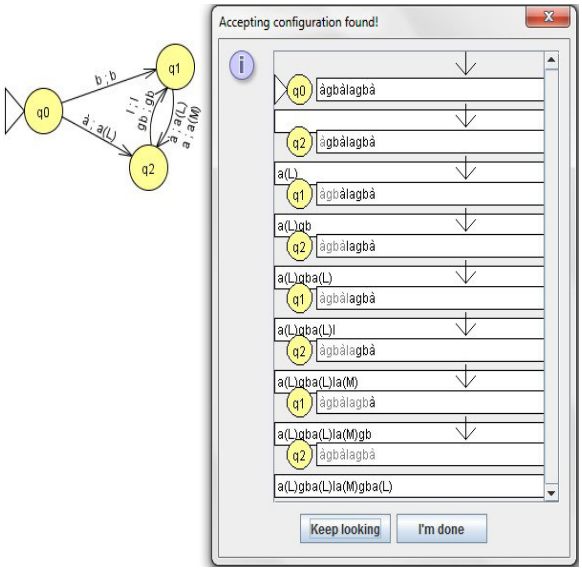


Figure 3. FST model for specific example of *Yorúbà* G2P

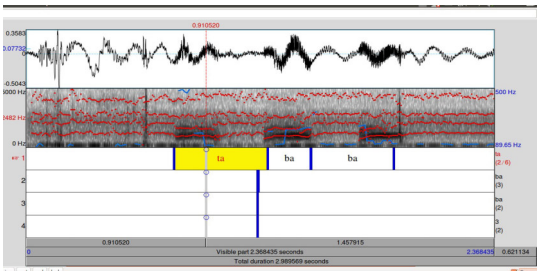


Figure 4**.** Speech sample of the word 'ta ba ba'

Table 1. Description of Yorúbà Grapheme with their Phonemic Transcription

S/N	Grapheme	Sound	IPA	Type	Phoneme Notations
1	a	Consonant	a	Mid	aa
2	b	Oral Vowel	b	Stop	b ih
3	d	Consonant	d	Stop	d ih
4	e	Oral Vowel	e	Front	ey
5	ẹ	Oral Vowel	ɛ	Front	eh
6	f	Consonant	f	Fricative	f ih
7	g	Consonant	g	Stop	g ih
8	gb	Consonant	g͡b	Stop	g b ih
9	h	Consonant	h	Approximant	hh ih
10	i	Oral Vowel	i	Front	iy
11	j	Consonant	d͡ʒ	Stop	jh ih
12	k	Consonant	k͡p	Stop	k ih
13	l	Consonant	l	Approximant	l ih
14	m	Consonant	m	Nasal	m ih
15	n	Consonant	n	Nasal	n ih
16	o	Oral Vowel	o	Back	ow
17	ọ	Oral Vowel	ɔ	Back	ao
18	p	Consonant	p	Stop	p ih
19	r	Consonant	r	Approximant	r ih
20	s	Consonant	s	Fricative	s ih
21	ṣ	Consonant	ʃ	Fricative	sh ih
22	t	Consonant	t	Stop	t ih
23	u	Oral Vowel	u	Back	uw
24	w	Consonant	w	Approximant	w ih
25	y	Consonant	y	Approximant	y ih
26	an	Nasal Vowel	ã	Front	aa n
27	ẹn	Nasal Vowel	ɛ̃	Front	eh n
28	in	Nasal Vowel	ĩ	mid	iy n
29	ọn	Nasal Vowel	ɔ̃	Back	ao n
30	un	Nasal Vowel	ũ	Back	uw n

Table 2. ASCII character map

Character	ASCII equivalent
á	av
à	ac
é	ev
è	ec
ẹ́	xv
ẹ̀	xc
í	iv
ì	ic
ó	ov
ò	oc
ọ́	qv
ọ̀	qc
ú	uv
ù	uc
ṣ	z

Table 3. Selected Yor_ub_a texts for the evaluation

Words			
igi (tree)	aya (wife)	omele (type of drum)	ara (body)
ata (pepper)	ehoro (rabbit)	egungun (bone)	agolo (tin)
igba (200)	igun (edge)	gegere	ike (plastic)
ire (goodness)	baba (old man)	oko (farm)	agara (handicapped)
ọkọ (husband)	irin (iron)	egbin (antelope)	ẹnu (mouth)
Sentences			
	Ọdọ́mọde ni mí	(I am a little child)	
	Tolúlọpẹ́ lorúkọ mi	(Tolúlọpẹ́ is my name)	
	Mo ra bàtà tuntun	(I bought new shoes)	
	Mo gbá Tolú létí	(I slapped Tolú)	
	Mo fẹ́ gbálẹ̀	(I want to sweep the floor)	
	Bàbá mi tidé	(My father has come)	
	Mo lọ sọ́ jà	(I went to the market)	
	Bàbá àgbẹ̀ ta kòkó	(The farmer sold cocoa)	
	Ọmọ re biyán	(Good child)	

Table 4. MOS for intelligibility

Sentences	1	2	3	4	5	6	7	8	9	10	MOS
bàbá mi tidé	3	3	2	1	2	2	1	3	2	3	2.2
ọmọdé ni mí	3	2	2	1	2	2	1	3	2	2	2.0
mo lọ sọ́ jà	3	3	2	3	2	2	3	3	2	3	2.6
bàbá àgbẹ̀ ta kòkó	2	3	2	1	2	2	1	3	2	1	1.9
ọmọ re biyán	3	3	2	1	2	2	1	3	2	3	2.2

Table 5. MOS for Naturalness

Sentences	1	2	3	4	5	6	7	8	9	10	MOS
bàbá mi tidé	2	3	2	1	2	2	3	3	2	3	2.5
ọmọdé ni mí	3	2	2	1	2	2	1	3	1	2	2.1
mo lọ sọ́ jà	2	3	2	3	2	2	3	3	1	3	2.5
bàbá àgbẹ̀ ta kòkó	2	3	2	1	2	2	1	1	2	1	1.7
ọmọ re biyán	2	3	2	1	3	2	1	3	1	3	2.1