American Journal of Tourism Management 2017, 6(1): 5-9
DOI: 10.5923/j.tourism.20170601.02

Development of a Mobile Tourist Assistance for a Local Language

Ninan O. Deborah[1,*], Iyanda A. Rhoda[1], Osoba A. Williams[2]

[1]Computer Science and Engineering Department, Obafemi Awolowo University, Ile-Ife, Nigeria
[2]PG Student Computer Science and Engineering Department, Obafemi Awolowo University, Ile-Ife, Nigeria

Abstract Speech is a natural and simplest way of communication amongst people. Speech synthesis is a process of automatic generation of speech for conveying information to a user in a preferred accent, language, and voice. This work develops a speech synthesis system for easy communication in Yoruba Language for tourists and implemented on android application. Text data were gathered from on-site interaction with the native speakers in four common domains and recording of the Yoruba phone-set was done using Praat software. The system was tested for naturalness and clarity among other metrics. The result shows 88.5% for clarity and 80% for naturalness. This system provides effective and efficient communication between locals and tourists in Yoruba Language and a platform to learn simple Yoruba sentences.

Keywords Tourist, Yorùbá, Text-to-speech

1. Introduction

Speech is the natural and most powerful means of communication used by humans to convey or share thoughts, ideas and notions. Speech is talking which is audio communication and a means of information expression that has an advantage of simplicity of use. The clearness of speech and accent are believed to be the important part in conveying the message correctly in verbal communication.

The only channel by which human beings abstract reality is by language. Tourist usually are faced with the difficulty of communicating effectively and efficiently with locals in the country they are touring, they hence would not bother going on a tour or would rather make sure they have a tour guide before embarking on such. Tourists most times would love to visit some domains even after visiting their major point of contact (the Tourist centre or festival) and communicating effectively is usually a problem. Speech enabled applications in public areas such as; railways, airport and tourist information centers might serve customers with answers to their spoken query. It has been observed that better performance in this area still requires further research (Alt, et al. 2008).

The need for Yoruba speech-synthesis system for tourists to communicate effectively in Yoruba language is essential and the primary thesis of this work. This work provides an efficient way of communicating between the locals and tourists; which minimises the need for human translator in communicating with the natives.

Paris in Stabb et al. (2002) observed that various information provided on the web only commonly include travel planning, route descriptions, and advice on sites to visit. Presentation of tourist information, to facilitate understanding and use, must be appropriate and natural. Yorùbá tourist assistance is a text-to-speech application for tourist implemented on android phones. This makes the application a mobile system available to the tourist anywhere and anytime.

The rest of the paper is structured as follows, Section 2 examines and presents a review of related works; Section 3 discusses the tools, data and the methodology employed in the development. Section 4 presents a discourse on the system evaluation and result while the paper concludes in Section 5.

2. Review of Literature

Linguistics is the scientific study of the human language. Linguistics can be broadly broken down into three subfields of study: *language form*, *language meaning* and *language in context. Language form* is a subfield that focuses on the system or rules the speakers or hearers of a language follow. It describes the morphology, which is the formation and composition of words, syntax that involves the formation and composition of phrases and sentences from these words and phonology. Phonetics is a related branch of linguistics concerned with the actual properties of speech sounds and non-speech sounds, how they are produced, and perceived.

The study of language meaning describes how languages

* Corresponding author:
deborah.ninan@gmail.com (Ninan O. Deborah)
Published online at http://journal.sapub.org/tourism

employ logical structures and real world references. It includes semantics that defines how meaning is inferred from words and concepts and pragmatics, which describes how meaning is inferred from context. Yorùbá is a tonal language, meaning that the basic word can have different meaning depending on the tone in which it is said. These tones are very noticeable aspect of both written Yorùbá and spoken Yorùbá pronunciation. There are three basic tones of different pitch levels in Yorùbá: High (marked with an acute accent in written Yorùbá (e.g. á)), Mid (not usually marked (e.g. a)) and Low (marked by a grave accent (e.g. à). In such register tone systems, tone realisation to some extent relies on changes in pitch between consecutive syllables (Van Niekerk, 2014).

There are two vowel types in Yorùbá: oral and nasalized. Oral vowels are produced entirely through the mouth and nasalized ones through the mouth and nose (Schleicher, 1997). Syllables in Yorùbá words end with a vowel or nasal sound, and there are no consonant clusters. It is common in some dialects of Yorùbá to combine the pronunciation of two syllables if one ends in a vowel and the next begins with one. According to Odejobi (2005), the five possible syllable structures of Yorùbá are consonant+oral vowel (CV), consonant+nasalized vowel (CVn), oral vowel (V), nasal vowel (Vn) and syllabic nasal (N). Every syllable bears one of the three basic tones: high, mid and low. Sequences of vowels in Yorùbá are pronounced as separate syllables and dialects differ in the number of vowels they have. Consonant clusters are not allowed in Yorùbá; therefore consonant sounds in the loan words are re-syllabified. The most common method for this is vowel insertion (Akinlabi, 2004).

2.1. Speech Synthesis Techniques

Synthesized speech can be produced by different methods which are usually classified into three groups: *Articulatory synthesis*, which attempts to model the human speech production system directly. *Formant synthesis,* which models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model. *Concatenative synthesis*, which uses different length prerecorded samples derived from natural speech. The formant and concatenative methods were identified as the most commonly used in present synthesis systems (Thakur et al. 2012).

The concatenative method is becoming more and more popular however, several problems which make concatenative synthesis unattractive compared to other methods include distortion from discontinuities in concatenation points; very high memory requirements especially when using long concatenation units such as syllables or words; data collecting and labeling of speech samples are usually time-consuming. The articulatory method has been reported still too complicated for high quality implementations, but may arise as a potential method in the future (Lemmetty, 1999).

Formant synthesis which is based on the source-filter theory of speech production bypasses the complications of modeling articulator movements, and computing the acoustic effects of those movements, by computing spectral properties of the transfer function directly from the linguistic representation using a set of carefully designed rules (Lemmetty, 1999). There are two basic structures in general, parallel and cascade, it has been pointed out that for better performance, some kind of combination of these is usually used. Formant synthesis also provides infinite number of sounds, which makes it more flexible than for example concatenation methods. (Sharma, 2007).

The information and communication technologies (ICT) has reformed the tourism industry especially in the areas of commerce and industry, transportation and interpersonal relationship. The tourist ability to identify his need and the clear communication with the seller was ascertained to be the strategy behind this success. Effective communication is a key factor that eventually leads to better pricing, less wastage and good customer satisfaction hence, this study.

2.2. Machine Translation

In science and technology, the demand for translation has almost exceeded the capacity of translation profession, and these demands are growing rapidly. Several ideas and works have been reported in the use of computers for translating natural languages (Hutchins, 1997). Pnisarn (2002) proposed a machine learning technique which considers the relationship between a word and its context information. Nuno et al. (2002) developed language models for a continuous speech recognition system for the Portuguese language. Savsa (2006) presented corpora for Arabic and French Machine Translation system. Sequence length based models and a pruned dynamic programming search and IBM model were used for sentence alignment.

Àjàdí, O. O. (2007) reported the Quantitative Model of Yorùbá Speech Intonation Using Stem-ML It describes a quantitative approach to modelling intonation in the context of a TTS system for the Standard Yorùbá (SY) language. The model is built and trained on speech data from a native speaker of SY. The resulting model for SY shows similar characteristics when compared to Mandarin and Cantonese intonation models.

3. System Design

Figure 1 shows the system design. The user is able to perform operations such as *Select a Domain*, *Select words to be pronounced*, and the option to *add new words* in the selected domain.

The select a domain option presents a pull down list of objects in the selected domain from which the user makes a selection. A new word and the corresponding speech and the transcription can be added in the *add new words* interface to keep updating the database.

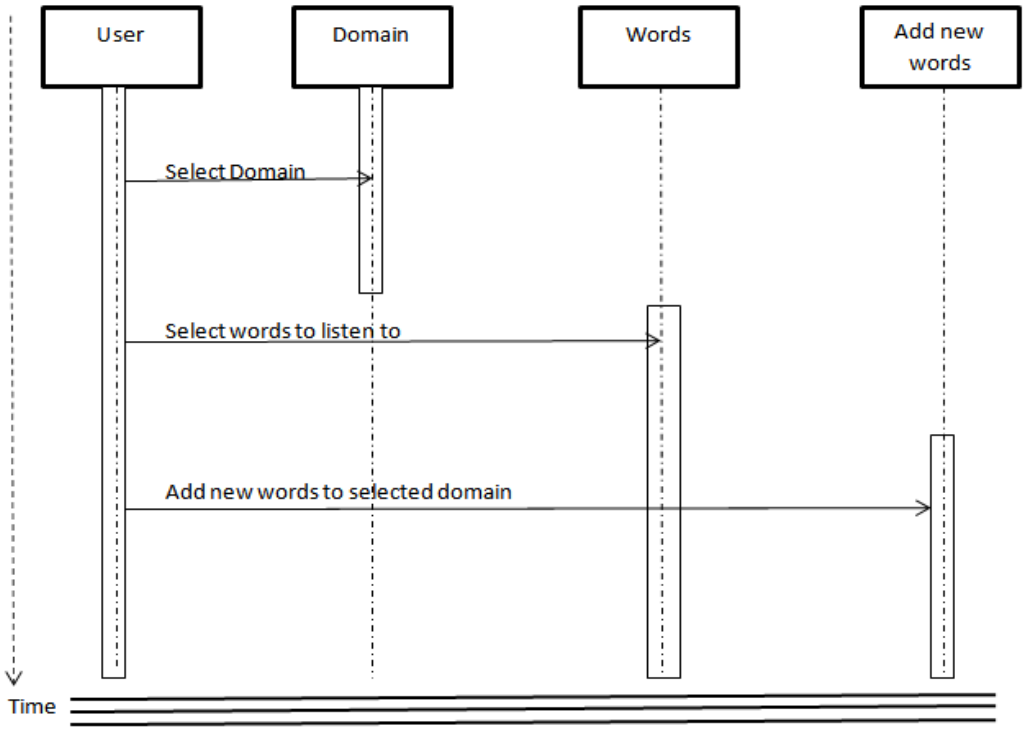

Figure 1. Sequence Diagram

3.1. Data Collection

Text data were collected from on-site interaction with the natives in the following four domains, *Market*, *Hospital, Motor-Park* and *Restaurant.* The sample text data are shown in Table 1 and the corresponding speech data were collected using Praat.

Table 1. Sample Text data from the four domains

Market	Hospital	Park	Restaurant
Banana,	Bed	Conductor	Egg
Beans,	Blood	Driver	Mortar
Chicken	Card	Motor	Pestle
Fish,	Doctor	Passenger	Stew
Maize	Drug	Fare	Plantain
Milk	Injection	Luggage	Chair
Pepper,	Test	Seat	Vegetable
Orange	Urine	Door	Soup
Meat,	Ward		plate
Tomato	Water		Spoon

3.2. Speech Data

Speech data were collected by recording the Yorùbá phoneset on Praat software using a male voice. Examples of recorded speech data (for market and restaurant domain) are shown in Figures 2 and 3. The figures show the digitized speeches which serve as speech input for the system and the formant frequencies (F0- F4) as well as the intensity for each word. The total duration for the recording of ọ̀gẹ̀dẹ̀ (Banana) is approximately 3.75 seconds, and that of ṣíbí (Spoon) is 3.41 seconds. Formant frequencies for each digitized word are shown in Table 2.

Table 2. Sample Formant frequencies for the digitized words

Formant Frequency	Ọ̀gẹ̀dẹ̀ (Banana)	Síbí (spoon)
F_0 (Hz)	98.76	183.10
F_1 (Hz)	624.22	1915.17
F_2 (Hz)	2278.53	2825.25
F_3 (Hz)	3939.47	3458.51
F_4 (Hz)	4864.50	4541.80

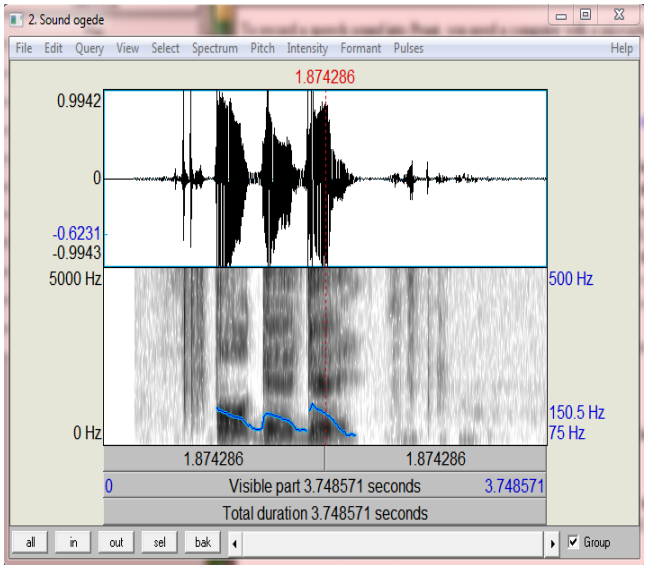

Figure 2. Ọ̀gẹ̀dẹ̀- Yoruba phoneset recording for Banana

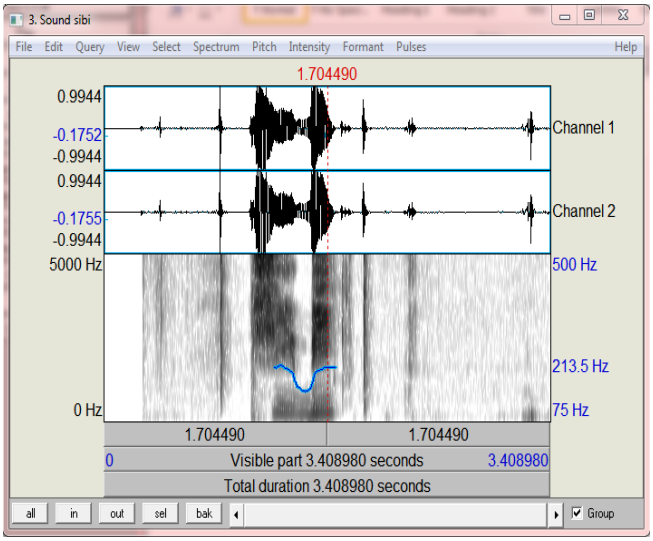

Figure 3. Yoruba phoneset recording for Síbí

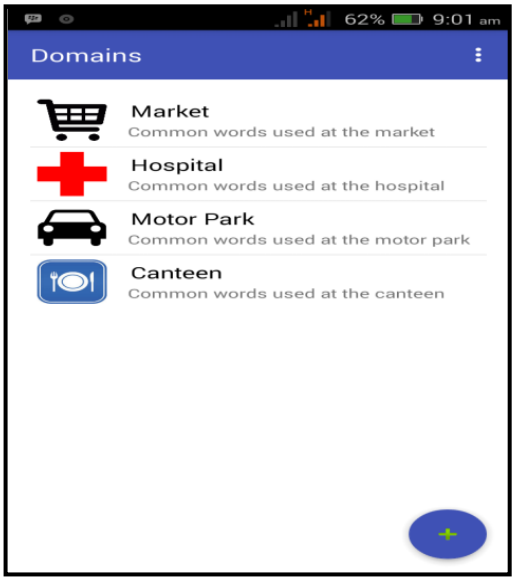

Figure 4. Domain Listing Interface

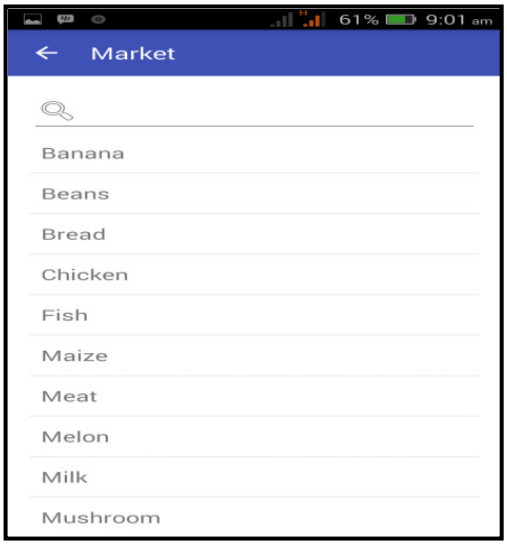

Figure 5. Market Domain Interface

Figure 6. Ògèdè Speech Interface

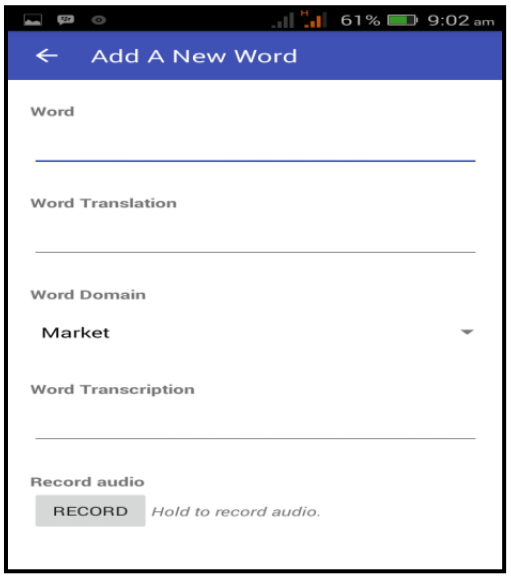

Figure 7. Add New Object Interface

3.3. Implementation

The tool used in the implementation of this application is Android Studio. Figure 4 shows the first interface that presents the four domains of the application. The user selects a domain and chooses an option from the list (Figure 5), then the corresponding audio in Yorùbá of the selected word is produced (Figure 6).

4. Evaluation

It has been reported that the most commonly used criteria for high-quality speech are intelligibility, naturalness and pleasantness (Thakur et al. 2012). These metrics were used for evaluation in this study. Twenty-Five students of Obafemi Awolowo University who are native speakers of Yoruba language were selected for testing the system. Each

listener was required to listen to the speech sounds produced by the system and rated the system on the selected metric using five scales (Very good, Good, Average, Poor to Very poor).

The result of the evaluation in Table 3 shows 28%, 52% and 20% rated the intelligibility of the system as Very good, Good and Average respectively while 64%, 32% and 4% rated the naturalness of the system as Very good, Good and Average respectively. In addition, 60%, 24% and 16% rated the pleasantness of the system as Very good, Good and Average respectively. This shows that the system built provides an effective way for communication between tourists and locals in Yoruba Language.

Table 3. Evaluation Result

Metric Rating	Very Good	Good	Average	Poor	Very Poor
Intelligibility	7	13	5	-	-
Naturalness	16	8	1	-	-
Pleasantness	15	6	4	-	-

5. Conclusions

The developed system provides a simple, effective and efficient way of communicating in Yoruba Language between tourists and locals in the four selected domains, which are usually indispensable. Text data and Speech data were collected and recorded respectively.

Tourists may in most cases have no need of a human translator since they can learn the basic words in the domains using this application. The database can also be updated by recording and adding words directly through the *add new object* interface. This application is expected to encourage tourists visit to tourist centers and other centers of attraction in Yoruba land. The result of this study can also be used for further research on speech synthesis and other speech application for the Yoruba Language and it is as well adaptable to other languages.

REFERENCES

[1] Akinlabi, A. (2004). Yoruba Sound System, understandong life and culture. Africa World Press.

[2] Alt, F. L., Rubinoff, M., and Yovitts, C. (2008). Advances in computers. New York Academic Press, 165-230.

[3] Àjàdí, O. O. (2007). A quantitative model of yorùbá speech intonation using stem-ml. *INFOCOMP Journal of Computer Science*, *6*(3), 47-55.

[4] Hutchins. (1997). From first conception to first demonstration: the nascent years of machine translation, 1947-1954, (pp. 190-260).

[5] Lemmetty, S. (1999). Review of Speech Synthesis T echnology. M.Sc Thesis, Department of Electrical and Electronics Engineering,Helsinki University of Technology.

[6] Nuno, S., Hugo, M., and Joao, N. P. (2002). Building Language Models for Continuous Speech Recognition Systems. Advances in Natural Language Processing Lecture Notes in Computer Science. Volume 2389, 101-110.

[7] Ọdẹ́jọbí, Ọ. À. (2005). A Computational Model for Prosody for Yoruba Text-to-Speech Synthesis. PhD. Thesis, Aston University.

[8] Pnisarn. (2002). Machine Language techniques. Machine Translation. Asia.

[9] Savsa. (2006). Creating a large scale Arabic to French MT system.

[10] Schleicher, N. (1997). Colloquial Yoruba: The Complete Course for Beginners.

[11] Sharma, R. (2007). *Speech Synthesis* (Doctoral dissertation).

[12] Stabb, S., Werther, H., Ricci, F., Zipf, A., Gretzel, U., Fesenmaier, D. R., Paris, C. and Knoblock, C. (2002). Intelligent systems for tourism. *IEEE Intelligent Systems*, *17*(6), 53-66.

[13] Thakur, B. K., Chettri, B., & Shah, K. B. (2012). Current Trends, Frameworks and Techniques Used in Speech Synthesis–A Survey. *International Journal of Soft Computing and Engineering*, *2*(2), 2231-2307.

[14] Van Niekerk, D. R. (2014). *Tone realisation for speech synthesis of Yorubá* (Doctoral dissertation, North West University).