Luganda Text-to-Speech Machine

By
Irene Nandutu
Reg.No.JAN15/COMP/0573U
School of Computing and Engineering

Supervisor
Dr. Ernest Mwebaze (PHD)
Lecturer School of Computing and Engineering
UTAMU

A Proposal Submitted to the School of Computing and Engineering in Partial fulfilment of the requirements for the award of Masters of Science in Computing Mobile Option (UTAMU)

September, 2016

1 INTRODUCTION

1.1 Background of the Study

Luganda has a large number of speakers from different language communities in the country. This language brings together people from the different parts of the country through entertainment, education and innovative content to people of a particular rural or urban gathering. The high number of Luganda speakers has enabled more Luganda linguistic writers who have taken the local Luganda text content to digital and this has led to Luganda digital literature being accessed on the internet. Today the digital Luganda literature is published in many areas like health, education, agriculture among others, this content is majorly displayed on social media site or independent websites that display Luganda digital text and a select few digital TV stations broadcasting majorly Luganda media content.

In 1947 an all Baganda conference recommended a standard orthography of Ganda (Empandiika y'Oluganda Entongole) which witnessed successful publications like Luganda Nouns in 1968 by Fredrick Katabazi Kamoga and Earl W. Stevick., Luganda-English dictionary which was meant to provide a basic vocabulary for learners of Luganda with an indication of the class of a noun and the modified stem of a verb (Sseguya, 2015). This kind of literature showed the nature of digital content in Luganda. Before that, the earlier writers had challenges in writing and translating Luganda because it is identified as a tonal language since it bases on pitch to detect the meaning of the words, differences in pitch can alter the meaning of a word for example in Luganda the word oku-kula which means to grow and oku-kuula which means to uproot are determined through the different pitch levels to identify the difference in them.

In 2013, Moses and Eno-Abasi emphasized that most African languages are resource limited with less or no linguistic resources like textbooks, electronic and printed reference materials, corpora, dictionary among others. In Uganda a language like Luganda is not exceptional as a few researchers have done some work on Natural Language Processing (NLP) in some Ugandan languages like software localization (Tushabe, Baryamureeba and Katushemererwe, 2010) which was done to bridge the digital gap in Uganda, Language translation mobile app which helps Baganda children living in diaspora to learn their native language and Google.co.ug translation into Luganda, an indigenous African language interface for Google Web Search.

This research shows a demand for innovative technology products on luganda language linguistics like learning systems by linguistic students, Luganda Audio literature which can be used in software applications like web and mobile to enhance the product functionality. These are the major demand for The Luganda text-to-speech system. The text-to-Speech

system will take input of text characters, process the characters and output them as spoken speech. This system will display information more efficiently and accurately, this system will use MARY (Modular Architecture for Research on speech sYnthesis) an open source speech sythesis platform developed on java.

1.2 Statement of the Problem

Text-to-Speech systems have simplified the interaction between computers and humans in natural languages. Existing text-to-Speech engines like for Microsoft, Google, festival, marytts among others have languages incorporated in them like English, French, Tegulu, Kiswahili among others which enables developers to use their Application Programming Interfaces to develop more various innovative products like learning systems for linguistic students to practice, language translation of text-to-text then text-to-speech systems, enabling reading aloud of text in websites, among others are products that have solved the needs of the people in these communities. These solutions have not been possible with most African languages like Luganda since its phonetics and transcriptions are not added into these text-to-speech engines to enable the speech capability of luganda text. Luganda text-to-speech engine is preferred for creating the Luganda speech that can be used in innovative products like e-learning courses, websites, mobile applications which will envetually help the visually impaired individuals and software developers among others. These text-to-speech systems can be trained to have machines that speak, read, or even carry out dialogs through text analysis (from raw text to identified words and basic utterances), linguistic analysis (finding pronunciations of the words and assigning prosodic structure to them) and waveform generation (from a fully linguistic analysis to generating a waveform).

1.3 General Objective

The main objective of this research is to develop a speech synthesis that transforms Luganda text into Luganda speech using text analysis, linguistic analysis and waveform generation.

1.4 Specific Objectives

This study will be guided by the following objectives;

i To review literature of building models of a text-to-speech machine

ii To collect data for training the Luganda text-to- speech machine

iii To build the luganda text-to-speech machine.

iv To evaluate the developed system.

1.5 Research Questions

i Which models can be used to build a Luganda text-to-Luganda speech machine?

ii How do we prepare tokenized corpus ready for building a Luganda text-to-Luganda speech machine?

iii Which procedures can we follow to develop a text-to-speech machine.

iv How can we evaluate the used models to determine text-to-speech quality?

1.6 Significance of the Study

The research provides the following values and benefits to users, researchers and competitors.

- This luganda tex-to-speech may be the first of its kind in Uganda and might contribute empirical evidence that support some of the previous theoretical developments and insights from qualitative research in speech synthesis, it shall also enhance previous solutions to Natural Language Processing gaps in Uganda
- The solution shall also be of benefit to organisations with Luganda digital literature in the read aloud process as it will ease the text-to-speech process of text from the author to reader and finally the listener
- Innovative products in speech synthesis concerned with luganda linguists will be realised through software developers who will develop solutions on mobile, web and desktop that capture luganda text and generate corresponding luganda speech
- The research can be upgraded to a text-to-text-to-speech ie translation systems or speech to speech system.
- Software developers in the ugandan industry will be motivated to build on open source platforms inorder to have purely innovative products that will solve local problems in language linguistics

1.7 Scope of the Study

The Geographical scope of this research will be Uganda and the timeframe that will be taken to conduct research will be four months, the researcher will use the already available Luganda tokens to build a text-to-speech machine.

The system will be divided into text analysis (from raw text to identified words and basic utterances), linguistic analysis (finding pronunciations of the words and assigning prosodic structure to them) and waveform generation (from a fully linguistic analysis to generating a waveform). The processing will use MARY (Modular Architecture for Research on speech sYnthesis) and in this study we shall use it for Luganda to Luganda speech synthesizer while evaluation the quality of output and propose future research.

2 Literature Review

2.1 Text to Speech Machine

A text-to-speech (TTS) synthesizer is a computer based system that can read text aloud automatically, regardless of whether the text is introduced by a computer input stream or a scanned input submitted to an Optical character recognition (OCR) engine. A speech synthesizer can be implemented by both hardware and software (D.Sasirekha and E.Chasndra , 2012). Speech synthesis that is the synthetic (computer) generation of speech, and text-to-speech or TTS; the process of converting written text into speech. As such it compliments other language technologies such as speech recognition, which aims to convert speech into text, machine translation which converts writing or speech in one language into writing or speech in another (Paul Taylor). The text-to-speech (TTS) synthesis consists of variety of modules as shown in the illustration below.

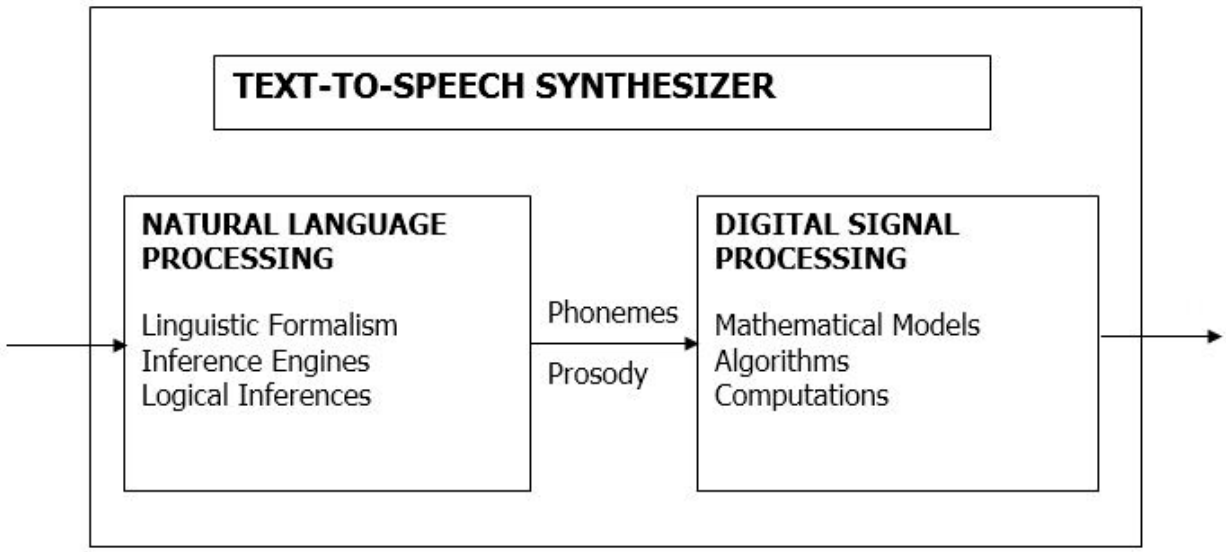


Figure 1: General functional diagram of a text-to-speech system.

2.1.1 Architecture of a Text to Speech System

The text-to-speech conversion is carried out using the following steps; text analysis, phonetic analysis, prosodic analysis and speech synthesis. These are divided into natural language processing as (text analysis, phonetic analysis, prosodic analysis)and digital processing system as speech synthesis or speech generation.

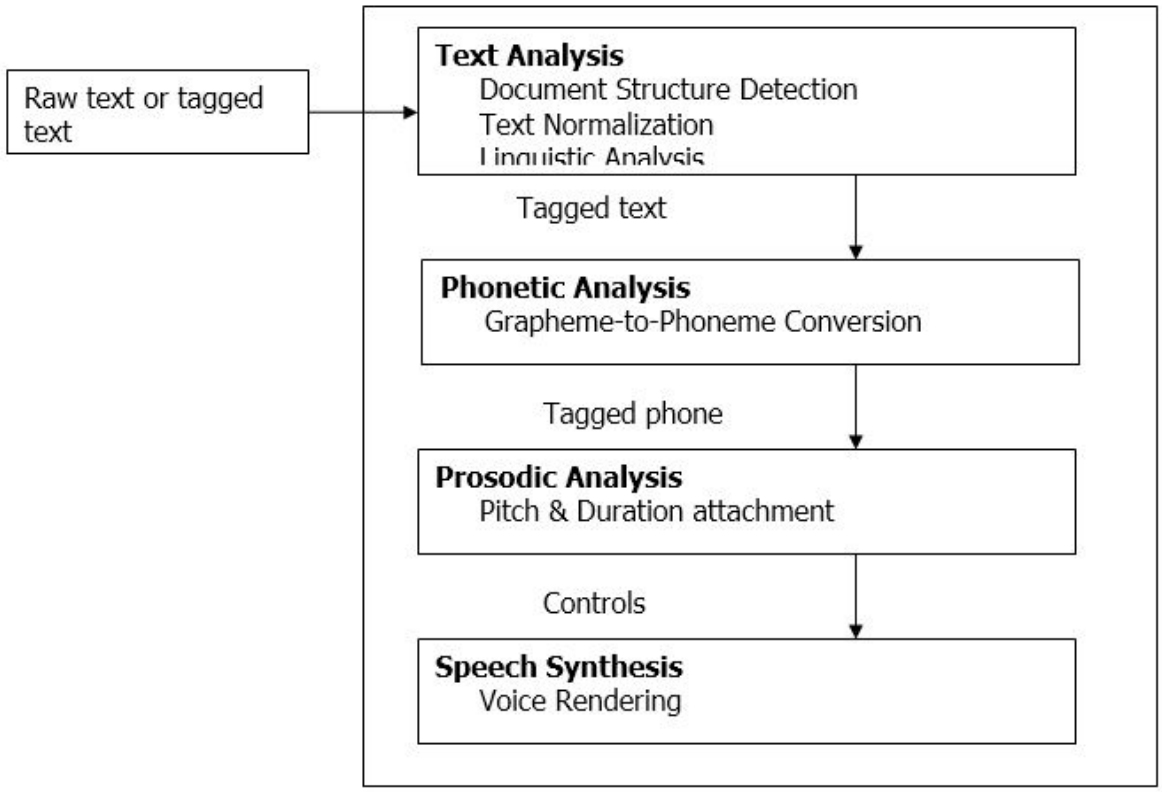


Figure 2: Architecture of a TTS..

- Text Analysis: This step includes Tokenization/Text preprocessing where the inputted text is broken down into smallest unit i.e. tokens, phrases or sentences. Tokenization also includes the expansion of abbreviations.
- Phonetic Analysis: This process includes the combination of the small units to represent them as phonemes that identify the sound associated with each unit. We witness the grapheme to phoneme conversion
- Prosodic Analysis: Are the patterns of stress and intonation in a language.
- Speech Synthesis: This is where the voices are built and rendered

2.2 Human Speech Production System

The main components of the human speech system are: The lungs, tracher(windpipe), larynx, pharyngeal cavity(throat), oral or buccal cavity(mouth), nasal cavity(nose). Normally the pharyngeal and the oral cavity are grouped into one unit called the oral tract. The nasal

cavity is normally called the nasal tract. The exact placement of the main organs is shown in figure

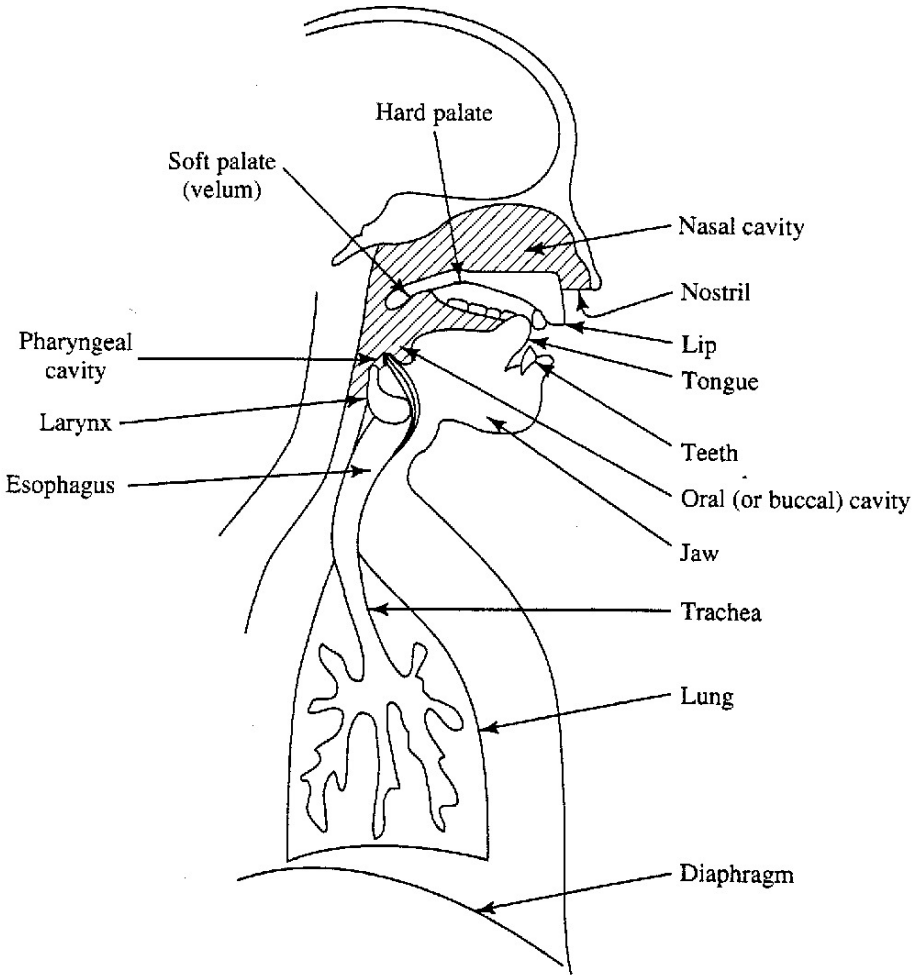


Figure 3: Schematic view of human speech production.

Speech production As seen in figure 3 speech production mechanism follows the following steps.

- Air enters the lungs via breathing.
- Air is expelled from the lungs, through the trachea, and cause to vocal cords to vibrate.
- Air flow is chopped up into quasi-periodic pulses.
- The pulses are frequency-shaped by the oral cavity and the nasal cavity.
- Body parts involved in speech production: lungs, trachea, vocal cords within the larynx, velum (soft palate), hard palate, tongue, teeth, lips, nasal tract.

2.3 Text to IPA Transcription

The orthographic representation of Luganda language text will be converted to International Phonetic Alphabet to be able to represent vocal sounds by signs and written characters. There are number of phonetic sounds that represent a letter of particular language. These phonetic sounds along with their written representation can be combined to generate sound using various speech synthesis tools. The orthographic set for a language consists of letter to sound rules. These rules define the function mapping of sequence of letter to sound segment which are the preliminary level of a text-to-speech. The methodology followed here translates the text to IPA transcription using some defined rules. These rules are used for determining the letter orthographic features. These orthographic features can further be used by the speech synthesis tool to generate sound. The principal vowels and consonants of Luganda will be depicted from the IPA chart for vowels and consonants.

2.3.1 IPA Chart Vowels

The principal vowels are symmetrically distributed on a standard vowel chart: three front vowels, two central vowels, and three back vowels. The three back vowels are rounded

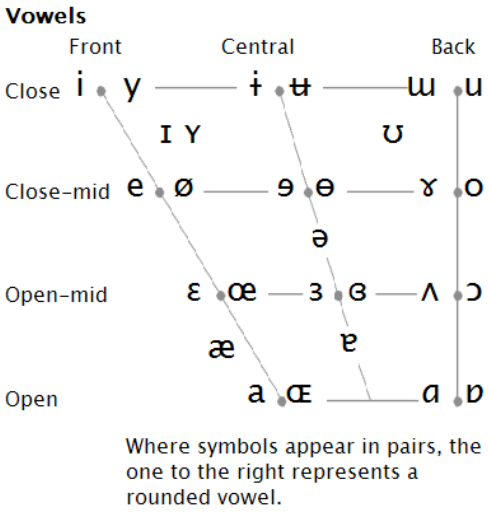


Figure 4: API Chart for Vowels.

2.3.2 IPA Chart Consonants

In the illustration below, Where symbols appear in pairs, the one on the right represents a voiced consonant, while the one on the left is unvoiced. Shaded areas denote articulations judged to be impossible.

International Phonetic Alphabet (IPA) ˌɪntəˈnæʃn̩l fəˈnɛtɪk ˈælfəˌbɛt

Consonants (pulmonic)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Figure 5: API Chart for Consonants.

3 METHODOLOGY

3.1 System Review

The current system supports languages like German, British and American English, French, Italian, Swedish, Russian, Turkish, and Telugu which are integrated in the Marytts. On the server of Marytts interface text is captured in any of these languages and it is detected and responds with spoken audio of the same captured characters, sentences or phrases. Luganda as a language is not incorporated in this engine yet the engine is open source making Luganda have a demand of having a free resource like a speech synthesis that developers could use for various innovations in their projects. This open source system if adopted will be used by websites to read Luganda digital text, Luganda audio bible, documents and also for learning purposes by linguistic students, all this content is to be trained to capture Luganda text to Luganda speech. The proposed system will help Luganda digital content to be trained to enable it have machines that speak, read, or even carry out dialogs in Luganda to Luganda text to speech.

This will replace the manual system where authors are paid by organizations to interpret and read aloud Luganda. The read aloud process is conducted by the authors or writers who extracts or generates verbal message from the meaning and encodes text to written signal to the reader, the reader decodes the text to verbal message and then encodes it to speech hence speech signal, the listener receives the speech signal into a message and interprets its meaning.

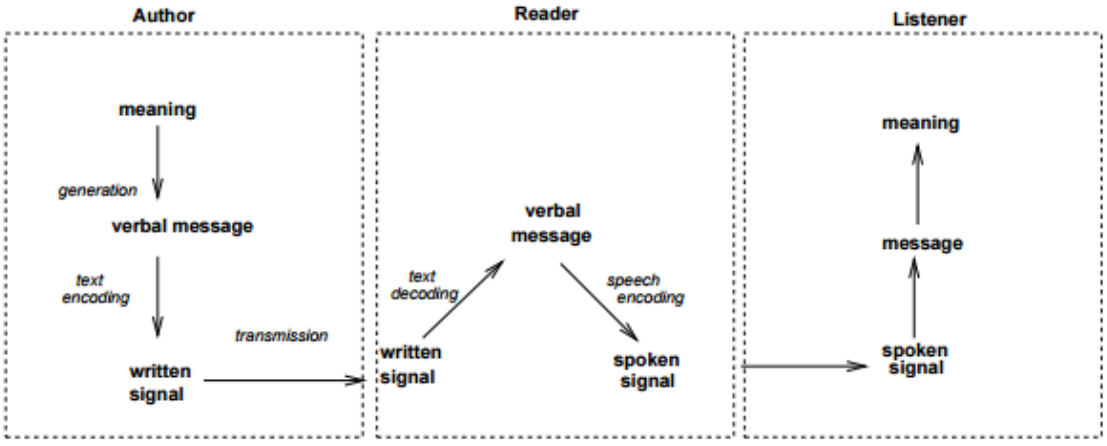


Figure 6: Basic Model of reading aloud, which shows the information flow between the author, reader and author. (Paul Taylor).

3.2 Data Collection

Collecting of observed, recorded, organized, categorized or defined information will be done to enable the success of the proposed system of Luganda text to Luganda speech, this data will be collected from

i Published Luganda literature like bible and newspaper.

ii Manual transcription of text where the researcher will transcribe Luganda words to Luganda phonetics.

iii Recorded Luganda audios or use of audio books like the recent published Luganda audio bible.

iv Internet

3.3 Importance of a Speech Synthesis

The speech synthesis can be applicable in the following situations for example in learning systems where individuals are taught in spelling and pronunciation of different languages, application for the deaf people as the system will enable talk to people who might not know the sign language, the blind as they will be able to respond back after hearing the sound of words from the system and applications for telecommunications and multimedia among others.

3.4 System Design and Implementation

The proposed system of lugnda text-to-speech will use the major procedures of any text-to-speech system as

- Text Analysis: includes such things as dividing the text into words and sentences, assigning syntactic categories to words, grouping the words within a sentence into phrases, identifying and expanding abbreviations, recognizing and analyzing expressions such as dates, fractions, and amounts of money, and so on. There are two reasons for a TTS system to do text analysis. One reason is that word pronunciation sometimes depends on usage.A second, equally important reason for text analysis is that its results will be used to modulate the pitch,timing and amplitude of the speech so as to present the texts message clearly. In other words, we want the program to read as if it were a skilled speaker who had understood the text.

- Linguistic Analysis: includes morphological analysis, part-of-speech tagging and syntactic parsing. To some extent, all these are useful for finding the words. In addition to word identity detection, parsing and other types of linguistic analysis are often seen as being useful for helping with prosody.

- Waveform Generation:This phase of a TTS system uses a detailed phonetic specification to produce time functions of the control parameters for an acoustic or articulatory speech synthesis model, which are then used to calculate the samples of the speech waveform.

In this research, the text-to-speech enginee that will be adopted is the MARY Text-to-Speech System (MaryTTS). MaryTTS is an open-source, multilingual Text-to-Speech Synthesis platform written in Java. It was originally developed as a collaborative project of DFKI's Language Technology Lab and the Institute of Phonetics at Saarland University. It is now maintained by the Multimodal Speech Processing Group in the Cluster of Excellence MMCI and DFKI. As of version 5.2-SNAPSHOT, MaryTTS supports German, British and American English, French, Italian, Swedish, Russian, Turkish, and Telugu; more languages are in preparation. MaryTTS comes with toolkits for quickly adding support for new languages and building unit selection and HHM-based synthesis voices. The MARY text-to-speech (TTS) synthesis consists of variety of modules as shown in the illustration below.(Marc Schrder and Jrgen Trouvain)

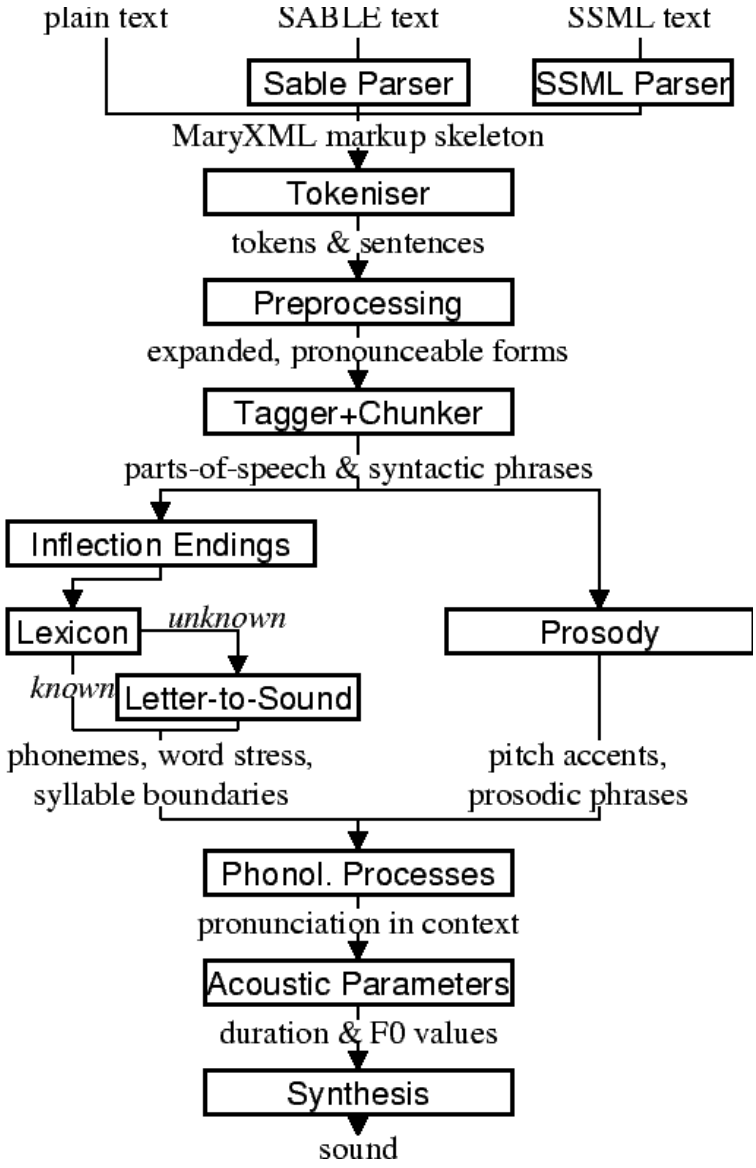


Figure 7: Architecture Structure of a MARY TTS

3.4.1 PlainText

Plain text is the most basic, and maybe most common input format. Nothing is known about the structure or meaning of the text. The text is embedded into a MaryXML document for the following processing steps

3.4.2 SABLE -annotated text and SSML-annotated text

SABLE is a markup language for annotating texts in view of speech synthesis, and SSML is a markup language for annotating texts in view of speech synthesis. It was proposed by the W3C as a standard. Speech synthesis markup languages are useful for providing information about the structure of a document, the meaning of numbers, or the importance of words, so that this information can be appropriately expressed in speech (such as pausing in the right places, pronouncing telephone numbers appropriately, or putting emphasis on the word carrying focus). Such information may be provided by a human user or, more likely, by other processing units such as natural language generators, email processors, or HTML readers.

3.4.3 Optional Markup Parser

The MARY text-to-speech and markup-to-speech system accepts both plain text input and input marked up for speech synthesis with a speech synthesis markup language such as SABLE or SSML. . Both SABLE and SSML are transformed to MaryXML which reflects the modelling capabilities of this particular TTS system. MaryXML is based on XML

3.4.4 Tokenizer

The tokenizer cuts the text into tokens, i.e. words and punctuation marks. It uses a set of rules determined through corpus analysis to label the meaning of dots based on the surrounding context.

3.4.5 Text Normalisation Module

In the preprocessing module, organizes the input sentences into manageable lists of words. It identifies numbers, abbreviations, acronyms and idiomatics and transforms them into full text when needed, those tokens for which spoken form does not entirely correspond to the written form are replaced by a more pronounceable form.

- Numbers: The pronunciation of numbers will highly depend on their meaning. Different number types, such as cardinal and ordinal numbers, currency amounts, or telephone numbers, must be identified as such, either from input markup or from context, and replaced by appropriate token strings.
- Abbreviations: Two main groups of abbreviations are distinguished: Those that are spelled out, such as "USA", and those that need expansion.

3.4.6 Part of Speech Tagger/Chunk Parser

Part-of-speech tagging is performed with the statistical tagger TnT (Brants, 2000), using the Stuttgart-Tbingen Tagset (STTS) (Schiller, 1995), we shall traine on annotated luganda corpus. A chunk parser is used to determine the boundaries of noun phrases, prepositional phrases and adjective phrases.

3.4.7 Phonemisation

The output of the phonemisation component contains the phonemic transcription. The SAMPA phonetic alphabet will be created for luganda and adopted to be used for each token, as well as the source of this transcription (simple lexicon lookup, lexicon lookup with compound analysis, letter-to-sound rules, etc.).

- Inflection endings: This module deals with the ordinals and abbreviations which have been marked during preprocessing as requiring an appropriate inflection ending. The part-of-speech information added by the tagger tells whether the token is an adverb or an adjective. In addition, information about the boundaries of noun phrases has been provided by the chunker, which is relevant for adjectives.
- Lexicons: The pronunciation lexicon contains the graphemic form, a phonemic transcription, a special marking for adjectives, and the inflection information
- Letter-to-sound conversion: Unknown words that cannot be phonemised with the help of the lexicon are analyzed by a "letter-to-sound conversion" algorithm. Letter-to-sound rules are statistically trained on the MARY lexicon.

3.4.8 Prosody Module

The prosody rules were derived through corpus analysis and are mostly based on part-of-speech and punctuation information. Some parts-of-speech, such as nouns and adjectives, always receive an accent; the other parts-of-speech are ranked hierarchically (roughly: full verbs ¿ modal verbs ¿ adverbs), according to their aptitude to receive an accent. This ranking comes into play where the obligatory assignment rules do not place any accent inside some intermediate phrase. According to a GToBI principle, each intermediate phrase should contain at least one pitch accent. In such a case, the token in that intermediate phrase with the highest-ranking part-of-speech receives a pitch accent. After determining the location of prosodic boundaries and pitch accents, the actual tones are assigned according to sentence type (declarative, interrogative-W, interrogative-Yes-No and exclamative). For each sentence type, pitch accent tones, intermediate phrase boundary tones and intonation phrase boundary tones are assigned. The last accent and intonation phrase tone in a sentence is usually different from the rest, in order to account for sentence-final intonation patterns.

3.4.9 Postlexical Phonological Process

Once the words are transcribed in a standard phonemic string including syllable boundaries and lexical stress on the one hand, and the prosody labels for pitch accents and prosodic phrase boundaries are assigned on the other hand, the resulting phonological representation can be re-structured by a number of phonological rules. These rules operate on the basis of phonological context information such as pitch accent, word stress, the phrasal domain or, optionally, requested articulation precision.

3.4.10 Calculation of Acoustic Parameters

This module performs the translation from the symbolic to the physical domain. The output produced by this module is a list containing the individual segments with their durations as well as F0 targets. This format is compatible with the MBROLA .pho input files.

3.4.11 Synthesis

At present, MBROLA is used for synthesizing the utterance based on the output of the preceding module. Due to the modular architecture of the M ARY system, any synthesis module with a similar interface could easily be employed instead or in addition.

3.5 Project Requirements

- Speech Engine; which the core of this research, we shall adopt MARYTTS engine to help in reading aloud luganda text to Luganda speech.
- Computer System; the computer will be used by the researcher as a platform to run the project application and any other components associated to it.
- Luganda corpus from both the internet and the published literature .
- Speakers to output the speech signal.

3.6 System Testing

The system will be tested to identify and resolve any errors and below are the major areas of focus;

- User Interface; this will enable users to interact with the system and their feedback used to rectify any bugs in the system
- Speech detection; this will be tested to detect the phonetic transcriptions of incoming text, it will be achieved by inputting data into the program, using the trained data the input data will be studied to determine the level of accuracy in the system.
- Measuring Quality of the system; The luganda phonetics and transcriptions will be created from the luganda text and aligned to corresponding luganda recorded phrases. To ensure quality of the speech we shall use global amplitude scaling to control the maximum amplitude of recorded files, remove low-frequency noise, trim initial and final sentences to avoid training long pause duration models, finally use of a binary sox to convert sampling rate among others to help us build a luganda voice.

References

[1] D.Sasirekha and E.Chandra (2012). TEXT TO SPEECH: A SIMPLE TUTORIAL. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1.

[2] K. Ngugi, W. Okelo-Odongo and P. W. Wagacha (2015). African Journal of Science and Technology (AJST) Science and Engineering Series Vol. 6, No. 1, pp. 80 89

[3] Paul Taylor. Text-to-Speech Synthesis, University of Cambridge

[4] Marc Schróder and Júrgen Trouvain The German Text-to-Speech Synthesis System M ARY. A Tool for Research, Development and Teaching

[5] Tushabe, F., Baryamureeba, V., and Katushemererwe, F. (2010). TRANSLATION OF THE GOOGLE INTERFACE INTO RUNYAKITARA.

[6] Jonathan Gosier, Nabireeba James and James Olweny (2015). How volunteer translators impact local communities: A Ugandan case study, from http://google-africa.blogspot.ug/2009/07/how-volunteer-translators-impact-local.html Monday, 17/July/2009.

[7] Leah Sternefeld and Sseguya Francis Nickshere (2015). Luganda-English Dictionary

[8] Moses Ekpenyonga,1, Eno-Abasi Uruab , Oliver Wattsc , Simon Kingc and Junichi Yamagish (2013). Statistical parametric speech synthesis for Ibibio.

[9] Ambrose Awici-Rasmussen (2015). Language Translation App, Safarini Translator from http://www.newvision.co.ug/new-vision/news/1325519/ugandan-develops-language-translating-app , 7th/May/2015.

[10] Zohre Owji, M.A(2013). Translation Strategies, A Review and Comparison of Theories,http://translationjournal.net/journal/63theory.htm Last updated on: 05/20/2014 03:40:03.

[11] Aras Ahmed Mhamad, Jutyar Zhazhlaiy and Lona Mariwany (2015). The Endless Challenges of Translation. Culture—Interview Last Updated 24/Oct/2015 from http://www.fairobserver.com/culture/endless-challenges-of-translation-75098/

[12] Er. Sheilly Padda and Ms. Rupinderdeep Kaur (2012), Architecture and Implementation of Punjabi Text to Speech System Using Transcriptions Concept. International Journal of Engineering Research and Development ISSN: 2278-067X, Volume 1, Issue 5 (June 2012), PP.08-11 . www.ijerd.com

[13] Brants, T. (2000). TnT A Statistical Part-of-Speech Tagger. Proc. 6 th Applied Natural Language Processing Conference, Seattle, WA, USA. http://www.coli.uni-sb.de/ thorsten/publications

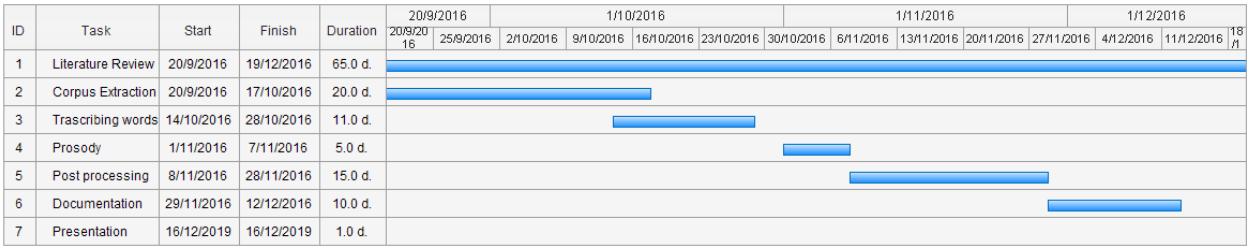
[14] Schiller, A., Teufel, S. and Thielen, C. (1995). Guidelines fr das Tagging deutscher Textkorpora mit STTS. Technical Report, IMS-CL, University Stuttgart. http://www.sfs.nphil.uni-tuebingen.de/Elwis/stts/stts.html

[15] Vincent Colotte and Richard Beaufor. Linguistic features weighting for a Text-To-Speech system without prosody model, LORIA - Speech team, University Henri Poincare, Nancy, France and Speech department, TTS group, Multitel ASBL, Mons, Belgium

[16] Paul Taylor.Text-to-Speech Synthesis. University of Cambridge

APPENDICES

Appendix 1: Time Frame



ID	Task	Start	Finish	Duration
1	Literature Review	20/9/2016	19/12/2016	65.0 d.
2	Corpus Extraction	20/9/2016	17/10/2016	20.0 d.
3	Trascribing words	14/10/2016	28/10/2016	11.0 d.
4	Prosody	1/11/2016	7/11/2016	5.0 d.
5	Post processing	8/11/2016	28/11/2016	15.0 d.
6	Documentation	29/11/2016	12/12/2016	10.0 d.
7	Presentation	16/12/2019	16/12/2019	1.0 d.

Figure 8: Time Frame