

ADDIS ABABA UNIVERSITY

FACULTY OF NATURAL SCIENCE

SCHOOL OF INFORMATION SCIENCE

DIPHONE BASED AMHARIC SPEECH SYNTHESIS SYSTEM USING MARY TTS

By: Bantyder Bahru

Advisor:

Solomon Teferra (PHD)

June, 2019

ADDIS ABABA UNIVERSITY
FACULTY OF NATURAL SCIENCE
SCHOOL OF INFORMATION SCIENCE

DIPHONE BASED AMHARIC SPEECH SYNTHESIS SYSTEM USING MARY TTS

A Thesis submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Masters of Science in Information Science

By:

BANTYDER BAHRU ARAGAE

June, 2019

DIPHONE BASED AMHARIC SPEECH SYNTHESIS SYSTEM

USING MARY TTS

By:

BANTYDER BAHRU ARAGAE

Name and signature of Member of the Examining Board

Name	**Title**	**Signature**	**Date**
_____	Advisor	_____	_____
_____	Examiner	_____	_____
_____	Examiner	_____	_____

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university. The complete work is the result of my own investigation. All relevant sources are acknowledged through citations with explicit references at the end of the report.

Declared by:

Name: _____

Signature: _____

Date: _____

The thesis has been submitted for examination with my approval as university advisor.

Name: _____

Signature: _____

Date: _____

DEDICATION

This work is dedicated to my Dad and Mom!

ACKNOWLEDGEMENT

I would like first to thank my thesis advisor Dr Solomon Tefera (PHD) for his uninterrupted effort to guide me and support me. The door to Dr. Solomon office was always open to whenever I run into a trouble and need guidance.

I would like also to thank Department of Information Science Staff for their uninterrupted support.

I would like to thank my wife Yordanos and my beautiful daughter Edna, my brother and sisters, mom and dad, friends and colleagues for their support in every aspect

Contents

LIST OF TABLES x
LIST OF FIGURES xi
LIST of Equations xii
LIST OF ACRONYMS xiii
Chapter One 1
INTRODUCION 1
1. Background 1
2. Problem Statement 2
3. Research Question 4
4. General Objective 4
5. Specific Objective 4
6. Scope and limitation of the Study 4
7. Significance of Study 5
8. Methodology 5
8.1. Literature Review 5
8.2. Data Collection Methods 6
8.3. Prototype Text to Speech Design and Tool 6
8.4. Evaluation Methods 7
8.5. Approach 7
9. Organization of the Thesis 9
Chapter Two 10
LITERATURE REVIEW 10
1. Human Speech Production 10
2. Historical development of text to speech 11
3. Improvements beyond Intelligibility 13
4. Commercial Text-to-speech development 14
5. Basic Components of a TTS System 15
5.1. TTS Frontend (NLP) 15
5.1.1. Text Analysis 15
5.1.2. Phonetic Analysis 16
5.1.3. Prosodic Analysis 17
5.2. TTS Backend (Speech Synthesis) 18

5.2.1. Articulatory Synthesis 18
5.2.2. Formant Synthesis 19
5.2.3. Series 20
5.2.4. Parallel 21
5.2.5. Concatenate synthesis 22
5.2.6. Concatenation Unit and Selection 24
5.2.7. Dipohne base synthesis 25
5.3. Mary TTS 26
5.3.1. Statistical Model Cost (sCost) 27
5.3.2. sCost Deployment in Mary TTS 27
6. EVALUAITON 28
7. RELATED WORK 30
7.1.1. TTS for Other Local Lanagues(Afan Oromo, Tigrgna and Wolyitgna) 30
7.1.2. Concatenative Amharic TTS System 31
7.1.3. Related work using MARY TTS 31
Chapter Three 33
Amharic language phonology 33
1. Overview of Amharic language 33
2. Phonetics 33
3. Amharic articulatory phonetics 34
3.1. Amharic vowel 35
3.2. Amharic consonants 35
4. Amharic writing system 36
Chapter Four 38
TTS Algorithm 38
1. Introduction 38
2. Pitch Synchronous Overlap-Add Synthesis (PSOLA) 38
2.1. Pitch-synchronous analysis 38
2.2. Pitch-synchronous modifications 39
2.3. Pitch-synchronous overlap-add synthesis 39
3. Time-Domain Pitch Synchronous Overlap Add (TD-PSOLA) 40
4. LPC-Based Synthesis 41
CHAPTER FIVE 42

Experimentation and Evaluation 42
1. Introduction 42
2. Architecture of Synthesizer 42
3. Corpus Preparation 44
4. Natural Langue Processing 45
4.1. Text Analysis 45
4.2. Phonetic Analysis 46
4.3. Transcription and Romanization 47
5. Minimal NLP components 51
6. Digital Signal Processing (DSP) 52
6.1. Recording Selected Sentence 53
7. Diphone Unit Selection Speech Synthesis 55
7.1. Feature Extraction 55
7.2. Feature Extraction from Acoustic Data 55
7.3. Automatic Labelling 56
7.4. Feature Vector Extraction from Text Data 56
7.5. Building acoustic models 57
7.6. Adding a new unit selection voice in MARY 58
8. Phone Unit Selection Synthesis 58
9. Test and Evaluation 58
9.1. Diphone Synthesis 59
9.2. Unit Selection Synthesis(phone) 61
CHAPTER SIX 63
Conclusion and Recommendation 63
1. Conclusion 63
2. Recommendation 64
REFERENCE 65
APPENDIXS 70

LIST OF TABLES

Table 1 : Amharic Consonants (sebsebi Z. et.al 2004) 36

Table 2 Top 10 frequent words from the wikipedia Amharic corpus 46

Table 3 MOS Scale 58

Table 4 Ineligibility Evaluation Result based on MOS scale 59

Table 5 Naturalness Evaluation Result based on MOS scale 60

Table 6 Performance of Amharic TTS 60

Table 7 Ineligibility Evaluation Result based on MOS scale 61

Table 8 Naturalness Evaluation Result based on MOS scale 62

Table 9 Performance of Amharic TTS 62

LIST OF FIGURES

Figure 1 Human Speech Production 11

Figure 2 Speech Synthesis 15

Figure 3 Text Analysis 16

Figure 4 Formant Synthesis 20

Figure 5 Series Formant Synthesis 21

Figure 6 Forman Synthesis Parallel 22

Figure 7 Concatenative Speech Synthesis 24

Figure 8 Amharic Vowels (sebsebi .et.al 2004) 35

Figure 9 TD-PSOLA Process 41

Figure 10 Architecture of MaryTTS (Marc & Jugen, 2003) 43

Figure 11 phonetic conversion of a word 47

Figure 12 Amharic Unicode to Latin character conversion 48

Figure 13 MaryTTS semi-automatic transcription tool 49

Figure 14 Amharic Phone Set based on SAMPA 50

Figure 15 Amharic pronunciation dictionaries 51

Figure 16 Features extracted from reliable sentences 52

Figure 17 RedStart Voice Recording Tool 53

Figure 18 Recorded Voice Display form 54

LIST of Equations

Equation 1 Formant frequency transfer function 20

Equation 2 Transfer Function for Parallel Synthesizer 21

Equation 3 Formant Synthesis with summation 22

Equation 4 Cost Function (S.Pammi & M. Charfuelan, 2013) 27

Equation 5: Overall Cost of Unit Selection (S.Pammi & M. Charfuelan, 2013) 28

Equation 6 Pictch-syncronous analysis (Eric Moulines, 1990) 38

Equation 7 least-square overlap add synthesis scheme (Griffin & Lim, 1984) 40

LIST OF ACRONYMS

CART	Classification and Regression Tree
LPC	Linear Predictive Coding
MOS	Mean Opinion Score
NLP	Natural Language Processing
NSWs	Non-Standard Words
RELP	Residual Excited Linear Predictive
SWs	Standard Words
sCost	Statistical model Cost
TD-PSOLA	Time Domain Pitch Synchronous Overlap and Add
TTS	Text to Speech
VODER	Voice Operating Demonstrator

ABSTRACT

This paper tries to explore possibility of designing Amharic TTS using diphone and phone as unit. Concatenative speech synthesis is now widely adopted due to comparative performance with large volume of corpus. In this regard during the thesis Wikipedia Amharic articles are used as corpus to train letter-to-sound as well as preparation of pronunciation dictionary.

For generation of acoustic unit and features 120 sentences selected for maximum balanced phone coverage and it was recorded. The developed Amharic synthesizer for both diphone and unit selection is then evaluated for its performance, naturalness and intelligibility by five native speakers of the language.

The evaluation results show that 94% of words are correctly identified for diphone based synthesizer and 84 % for unit selection based on Dynamic Rhythmic Test (DRT). The test using MOS levels shows that the synthesizer achieved 3.6 and 2.65 for intelligibility and naturalness respectively for diphone based synthesizer and 2.68 intelligibility and 2.55 naturalness for unit selection synthesizer. The result shows that diphone based synthesizer outperforms the unit selection synthesizer and the result of diphone synthesizer looks promising and with introduction of non-standard word and professional studio level voice recording might increase the performance as well.

Chapter One

INTRODUCION

1. Background

Text-to-speech (TTS) also known, as speech synthesis is computer aided synthetic speech generation. It is also the process of converting written text into speech (acoustic waveforms). Text-to-speech systems have an enormous range of applications. It is often used as an aid to the disabled; their first real use of TTS was in reading systems for the blind, where a system would read some text from a book and convert it into speech (klatt, 1987). Currently TTS have found place to support in day-to-day activity in many hand held devices with desirable promise hands-free, natural, and ubiquitous access to the interacting device (Nadew, 2008).

The main goal of building TTS system is to clearly convey message with human-like voice with good intelligibility and naturalness. One of the main differences of speech synthesis with record and play type system is that TTS is capable of speaking anything regardless of whether the desired message was originally spoken or not. Concatenative TTS and Statistical TTS are main speech synthesis technique. Concatenative TTS requires large databases of speech waveforms in order to synthesize human-like high quality speech from text. (Stan, 2011)

Speech synthesis tool is increasingly becoming important in wide range of application. The importance of TTS technologies can be seen in navigation, tutorial, application for visual impaired persons, etc... Most speech synthesis available currently including MARY TTS includes most researched and widely used languages like English, Germany and French etc. Although application of text to speech is equally important for the local community but there is

no comprehensive speech synthesis tool as per the authors' knowledge. There are a lot of break through research conducted in speech synthesis for local language but there is no speech synthesis tool fully supporting local language. The main motivation of this research is to contribute towards better platform for TTS research for local language specifically for Amharic language. (Million & Alula, 2010; Solomon & Bahiru, 2017)

The study tries to achieve better performing text to speech synthesis system using MARY TTS for Amharic language. MARY TTS is an open source speech synthesis tool with modular step-by-step processing designed to support research on speech synthesis, which enables user to manipulate various stages on speech synthesis (Schroder & Trouvain, 2003). The successful integration on Amharic language on MARYTTS will help usage and further research on Amharic language TTS. This intern helps to create commercial TTS application in Amharic language in different area including IVR, supporting people with disability and as well as hands free Amharic content reading. The baseline speech synthesis integration will be based on diphone text to speech synthesis for Amharic language.

2. Problem Statement

Previous Amharic speech synthesis research attempted to build TTS based on open source tools. Bahiru Demise tries to experimentally demonstrating syllable based acoustic models for developing Amharic speech synthesis system using HMM and ASR corpus using HTS (Solomon & Bahiru, 2017) . Bahiru reported the performance of TTS system 75.56% for syllable base and 77.78% for phone based; on MOS evaluation Bahiru found the naturalness on MOS evaluation 3.13 and 2.8 for syllable based and phone-based TTS respectively. He concluded that that phone-based TTS system outperforms syllable based. On the other hand, Alula Terefe tried to develop a prototype Text-to-Speech synthesis system for Amharic language that converts standard and

non-standard words encountered in real-world text reading. (Million & Alula, 2010). To generate natural sounding and intelligible speech Allula reported overall performance of pronunciation the designed speech synthesis tool is 73.35%. Evaluation of synthetic speech produced by the system using MOS score is 3 and 2.83 intelligibility and naturalness, respectively. Used Festival-LPC. Eyob also developed Amharic synthesizer following concatenate synthesis approach to increases naturalness and intelligibility by multiple recorded phones for addressing context and sentence structure prosody (Eyob, 2011). He evaluated and reported 3.63 and 3.53 for naturalness and intelligibility respectively.

Both Alula and Bahru used festival tool which uses LPC based speech synthesis module, on the other hand Eyobe used Edinburgh Speech Tools which also based on LPC algorithm which is used to represent the spectral envelope of speech in compressed form. Although all previous research has achieved acceptable performance in the research all researchers agreed that there needs to be further research to achieve better performance using different algorithm and approach. One comparative research on Arabic text to speech synthesis shows that MBROLA based speech synthesis module on MARY TTS outperforms LPC based synthesis module on Festival on both intelligibility and naturalness of the speech using DRT and naturalness scoring introduced by Sluijter et al. respectively (M. Z. Rashad, Hazem M. El-Bakry, & Islam R. Ismail, 2010). Arabic is semantic language as Amharic and therefore this study tries to achieve Amharic speech synthesis system with better performance and effectiveness compared to previous Amharic TTS system by applying MBROLA algorithm on speech synthesis on diphone base concatenation.

3. Research Question

This study tries to answer whether diphone base speech synthesis performs with better intelligibility and naturalness compared to phone base speech synthesis.

4. General Objective

The general objective of the study is to develop intelligible and natural sounding diphone base and phone base concatenative speech synthesis for Amharic language with Wikipedia Amharic text corpus.

5. Specific Objective

The following specific objectives are planned to be achieved to fulfill general objectives: -

- Critical literature review will be conducted to adopt techniques
- Preparation of text corpus and text processing to extract features required for NLP
- Create voice in Amharic language
- Prepare dipohne and phone prototype synthesizer module integration of Amharic for MARYTTS
- Analyze and evaluate synthetic voice generated from sample selected from previous Amharic TTS research

6. Scope and limitation of the Study

This research prototype will only focus only on Amharic language TTS. The speech synthesizer is designed to pronounce any standard Amharic word. Part of speech tagger, nonstandard words, word disambiguation and controlling intonation level is not included in this speech synthesis. Due to financial constraint voice was recorded in house not on professional studio. The Amharic

synthesizer also currently accepts only Romanized text due to limitation of MaryTTS .

7. Significance of Study

Text to speech synthesis have found application in helping visual impaired, in language education, in telecommunication, Interactive voice response and announcement system. Most of tools using that application are currently supports western languages like English, French, …etc. But the above-mentioned application is equally important in local context and there is a lot of effort to embrace that feature and make available for our society which needs those technologies at most. Designing good text to speech module for Amharic language will help technology creators to use research results to integrate and support Amharic in their product which intern will help Amharic speakers in different area. This research also tries to prototype Amharic TTS system with better naturalness and intelligibility that could contribute towards the successfully implementation publicly available TTS tool. This research also contributes towards the wealth of knowledge in the field of speech synthesis for local language.

8. Methodology

8.1. Literature Review

Before starting the actual research and at various stages of the research relevant journals, books, scholarly articles and conference proceeding on TTS, Amharic linguistic features, Amharic language TTS and different speech synthesis algorithms and techniques was consulted. Different speech processing techniques was also studied to select suitable procession method for this research.

8.2. Data Collection Methods

In this research text corpus was collected from Wikipedia Amharic language articles and processed for linguistic and voice feature extraction. In addition, one male voice was recorded and used for di-phone and phone processing. Part of the collected sample data was used to test the prototyped Amharic TTS system on MARY TTS.

8.3. Prototype Text to Speech Design and Tool

In this thesis diphone and phone base Amharic speech synthesis prototype using MARY TTS was developed. MARY (Modular Architecture for Speech Synthesis Research) is a platform for text-to-speech synthesis studies, growth and learning. MARY speech synthesis module presents statistical model Cost. Originally created for German, some TTS modules from the FreeTTS project (Walker, Lamere, & Kwok, 2005) have been expanded to US English. MARY uses an XML-based representation format for its data, which makes it possible to access intermediate processing states, and to connect it to other XML-based processing components. The latest version (5.2) of MaryTTS supports German, British and American English, French, Italian, Luxembourgish, Russian, Swedish, Telugu and Turkish. (Marc & Jugen, 2003). Prototype designing started by performing TTS front-end which include text analysis, phonetic analysis and prosodic analysis on the selected corpus. The result of TTS front-end is used to design minimal NLP component compatible to MaryTTS. In addition, 120 sentences were recorded as part of preparation for TTS back-end, which was used for acoustic feature extraction. Different opensource tool embedded in MaryTTS voice import tools was used to process voice. Finally, the output of TTS front-end and back-end was used to build Amharic TTS module for MaryTTS platform.

8.4. Evaluation Methods

Synthetic Amharic speech generated from the system was evaluated using test data for its intelligibility and naturalness. The intelligibility of the speech means whether or not the synthesized speech could be understood by human listener, on the other hand naturalness of the speech mean whether synthesized speech sounds like the human speech or not. There are different techniques to perform evaluation on TTS system. This research used widely adopted Mean Opinion Score (MOS) and Diagnostic Rhyme Test (DRT). MOS is the five-point scale in which speech quality is normally evaluated. The quality speech measured against five points scale namely bad (1), poor (2), fair (3), good (4) and excellent (5). When quality scores reach 3.5 or higher, it shows high level naturalness and intelligibility (Lemmetty, 1999). DRT which measures the intelligibility using a list of word pairs that differ only in a single consonant are uttered and the listeners are asked to mark on an answer sheet which pair of the words they think is correct.

8.5. Approach

To add new language support in MaryTTS requires building a basic set of Natural Language Processing (NLP) components for the new language including tokenization, phonemic transcription, and the creation of the voice models in the new language. Two main tasks were done to add Amharic language on MARY TTS.

1. Construct Minimal Set of NLP

    1.1. Corpus Selection

    Wikipedia Amharic language text dump was used as a corpus for the research. The corpus contains 400 sentences and 100,00 unique words.

    1.2. Preprocessing or text normalization

Collected corpus was cleaned from any non-Amharic words and characters, tokenized and resulting 200 sentences with 80,000 words.

1.3. Natural language processing doing linguistic analysis and annotation;

Speech-relevant data like diphone symbols and intonation labels was extracted out from preprocessed text data. The following NLP step were performed

- Transcription of most frequent words
- Generated Letter-to-sound (LTS) rules for automatic training new word
- Part of speech labeling and shallow parsing is performed.
- Lexicon lookup is performed in the pronunciation lexicon;
- Unknown tokens are morphologically decomposed and phonemised by grapheme to phoneme (letter to sound) rules.
- Symbols for the intonation and phrase structure are assigned by rule, using punctuation, part of speech info, and the local syntactic info provided by the chunkier.
- Post lexical phonological rules are applied, modifying the phone symbols and/or the intonation symbols as a function of their context.

1.4. Calculation of acoustic parameters, which translates the linguistically annotated symbolic structure into a table containing only physically relevant parameters by applying a model for duration and for intonation.

1.5. Synthesis transforming the phoneme parameter table into a sound file

2. Creation of a voice in Amharic language

9. Organization of the Thesis

This thesis is organized as the follows: Chapter two presents detail literature review and related works on human speech production system, speech synthesis. In addition, related works on speech synthesis system on different national language using different speech units, techniques tried by researchers are reviewed. Chapter three presents phonology aspect of Amharic language, syllable structure, linguistic and prosodic feature of speech synthesis in Amharic and novelty of the research is discussed. Chapter four discussed different speech procession algorithms. Chapter five is design prototypes Amharic speech synthesizer, experimental results and evaluation are discussed. in Chapter five. In Chapter six, conclusion and future work directions are described.

Chapter Two

LITERATURE REVIEW

1. Human Speech Production

Human speech production is generation of voice through articulatory organs and parameters characterizing it. Sound produced by expiration of air from lung through vibrating vocal cords call sound. The generated voice passes through chest and vocal organs like throat and mouth cavity process before it audible. The quality of the voice is determined by resonance and the manner in which the vocal cords vibrate, which in turn can be characterized in terms of pitch, quality and intensity (Louw, 2004) .

The produced voice becomes speech as it goes through the vocal organs as shown in Figure 1 (Louw, 2004). The sound energy is generated mainly from the lungs together with the diaphragm. During speech air flow is forced through the three main oral and nasal cavities (1) of the vocal tract called glottis (10), vocal cords (12) and the larynx and pharynx (9). Air flow forced to go out through the nose from nasal cavity and mouth from the oral cavities. Glottis is V shaped opening in the middle of the vocal cords call glottis is the most important sound source in the vocal system. To modulate the air flow the vocal cords may act in different ways by rapidly opening and closing in order to generate buzzing sound form which voiced consonants and vowels are formed (Louw,2004). With respect to this, the generated sounds can be seen as voiced and unvoiced sound based on whether the vocal cord has vibrated or not when the sound is generated (Louw, 2004).

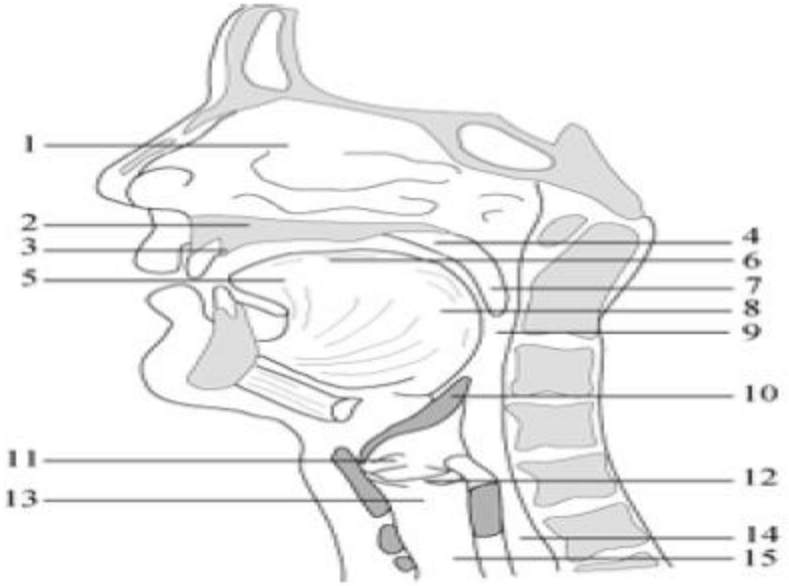


Figure 1 Human Speech Production

The label for each vocal organ: 1. Nasal cavity, 2. Hard palate, 3. Alveolar ridge, 4. Soft palate (Velum), 5. Tip of the tongue (Apex), 6. Dorsum, 7. Uvula, 8. Radix, 9. Pharynx, 10. Glottis, 11. False vocal cords, 12. Vocal cord, 13. Larynx, 14. Esophagus, and 15. Trachea

2. Historical development of text to speech

The first successful recorded attempt to speech synthesis was made by Professor Christian Kratzenstein in 1769 in Russia. When he used mechanical human vocal tract model to construct acoustic resonators. Constructed resonators were capable of reproducing steady state vowels sound a, e, I, o, u by activating with vibrating reeds like in music. (Flanagan, 1972)(Flanagan 1972, Flanagan et al. 1973, Schroeder 1993). But the pneumatic synthesizer capable of producing recognizable single sounds and some sound combinations which was built by Wolfgang von Kempelen in 1791 was recorded as successful connected speech synthesizer

(klatt, 1987). The machine contains pressure chamber, vibrating reed tract and leather tube functioning as lungs, vocal cords and the vocal tract action respectively. The mechanical speech synthesizer research was also used for speech synthesis in late 1930 (Flanagan, 1972) .

With electronic resonators and analogs of acoustic resonators developed it can be easily controlled compared to mechanical or pneumatic synthesizer, using this advantage the bell laboratories produced the first electrical connected speech synthesizer Vocoder. Inspired by bell vocoder, vodER (Voice Operating Demonstrator) is one on the earliest electronic speech synthesizer developer by Dudley in late thirties. Voder is constructed with ten electronic resonators arranged as contiguous band pass either controlling keyboard, writs pedaled foot pedal which required experienced operator to produce speech like piano and considered as no practical use at the time (Flanagan, 1972; klatt, 1987). The development of VODER demonstrated that intelligible speech can be produced artificially and creates interest in the scientist community to engage in speech synthesis research.

The progress of speech synthesis tool reached in the development of various devices which can read spectrogram or simplified hand-painted spectrogram. The most famous spectrogram developed in 1950 was the pattern play-back machine developed by Cooper, Liberman and Borst (Klatt 1987, Flanagan et al. 1973). This machine reads continuous roll of spectrogram by illuminating the spectrogram with light spot the intensity of the light passed through the spectrogram is detected by photocells which will be converted to electric signals which are amplified and original sound is produced.

The first formant synthesizer, PAT (Parametric Artificial Talker), is created by Walter Lawrence in 1953 which consists of spectrogram like patterns painted onto a continuous sheet of plastic which was rotated around a mangle .The pattern on the plastic was divided up into six (later eight) concentric bands, with each band painted in electrically conductive silver to represent single parameter to control voicing amplitude , fundamental frequency and noise amplitude fundamental formant frequencies. A current would be passed through each of these continuous silver lines and a set of electric contacts (one for each parameter) would sense the position of each of the lines values which would be translated into continuously varying voltages that can be translated electronics capable of controlling the synthesizer's various filters etc.

In early days research in speech synthesis their focus was in generating speech with high *intelligibility* regardless of human listener understanding (Keller 1994; Holmes and Holmes 2001). Different methods of synthetic speech evaluation were developed which often involved comparison between different systems. In almost all evaluation methods intelligibility was almost always taken to mean segmental intelligibility. Measuring the speech segments which make up words were sufficiently well rendered. It is essential for synthesizers to move away from the basic requirements of minimally converting text-to-speech (Dutoit, 1997) and give more emphasis on naturalness of speech production. This will mean that the earlier synthesis model will necessarily become insufficient as the focus shifts from the reading task to the quality of the synthetic voice.

3. Improvements beyond Intelligibility

From a segmental perspective speech has been fully intelligible for several years but there are areas of naturalness which still needs to be in acceptable level (Keller, 2002). To convince listeners as if they are hearing real speech the speech synthesizers needs to reflect the most

natural aspect of human speech, that is we can say the synthetic speech sounds completely natural. This is one of the major research focus of speech synthesis because it constitutes a near inflexible problem. Many researchers agree that insufficiently natural synthetic speech is the major remaining problem to fully generate acceptable synthetic speech.

4. Commercial Text-to-speech development

In 1968 Japan Electro technical Laboratory designed the first intelligible full text-to-speech system based on articulatory model for English. It was developed by Noriko Umeda and his colleagues (klatt, 1987). In M.I.T MITalk laboratory Allen, Hunnicutt, and Klatt demonstrated text-to-speech synthesis that is latter modified and used by Tele sensory Systems Inc. (TSI) as commercial TTS system in 1979 (klatt, 1987; Allen, Hunnicut, & Klatt, 1987). In 1981 Dennis Klatt used new sophisticated voicing to introduce his famous work of TTS Klattalk system (klatt, 1987). Kurzweil in 1976 introduced first written text reading aid for blind which contains optical scanner (klatt, 1987). Following major speech synthesis research commercial text-to-speech and speech synthesis products were introduced in early 1980's (klatt, 1987). Although different attempt to design efficient speech synthesizer mechanically and pneumatic, the Votrax chip is the first electronic speech synthesis which consisted of cascade formant synthesizer and simple low-pass smoothing circuits. In 1978 Richard Gagnon introduced modified a cheap Votrax-based Type-n-Talk system. Later in 1980, Texas Instruments introduced Speak-n-Spell synthesizer based on linear prediction coding (LPC) and on low-cost linear prediction synthesis chip. Speak-n-Spell was used to aid for children as electronic reading. In 1982 Street Electronics used new Speak-n-Spell chip to design Echo low-cost diphone synthesizer. At the same time Speech Plus Inc. introduced the Prose-2000 text-to-speech system. A year later, first commercial versions of famous DECtalk and Infovox SA-101 synthesizer were introduced (klatt, 1987).

5. Basic Components of a TTS System

TTS system essentially contains two components as shown in **figure 2 below**. The NLP component on the left is TTS frontend and is logically separated from the synthesizer TTS backend on the right. The fronted takes text as input and performs text analysis and creates controlled information as its output. On the other hand, the TTS backend renders speech output tacking controlled information as input.

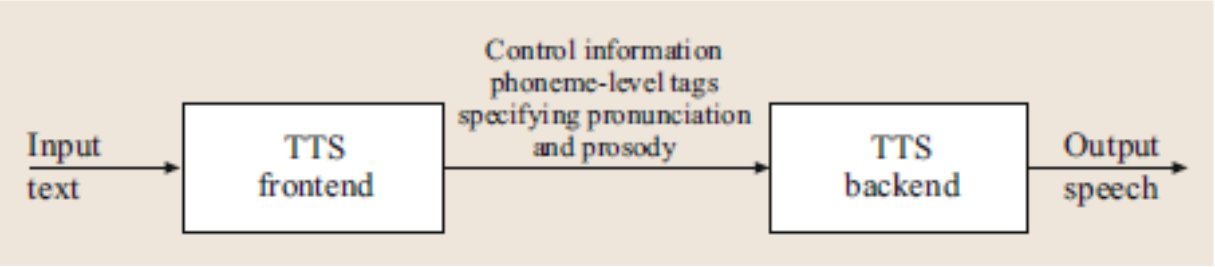


Figure 2 Speech Synthesis

5.1.TTS Frontend (NLP)

The frontend is one of resolving ambiguity from a noisy text by applying text processing, phonetic analysis and prosodic analysis.

5.1.1.Text Analysis

Text processing includes detecting document structure, text normalization and linguistic analysis. Document structure detection includes sentence splitting and sentence tokenization by detecting the presence of white space, punctuation...etc. Tokens are usually encoding of words, numbers, dates, abbreviations and other types. Text normalization handles non-natural language tokens to

be matched with educated reader would read the underling natural language form. To resolve ambiguities in written form of text linguistic analysis which includes morphological analysis is carried out for proper pronunciation.

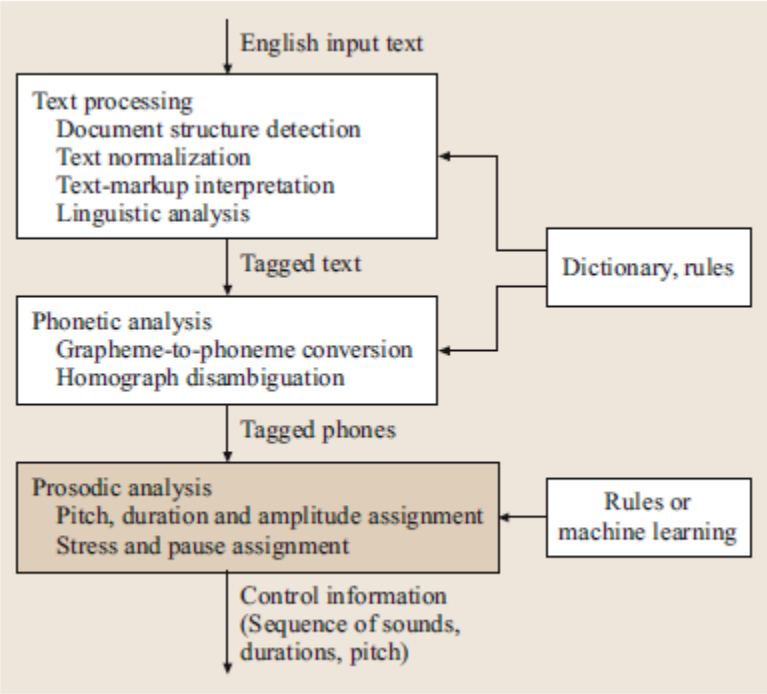


Figure 3 Text Analysis

5.1.2. Phonetic Analysis

In addition to text processing, phonetic analysis is responsible for representing each word with phone and tagging with speaking style and level of emphasis of sound produced and how it is produced. Phonetic analysis performs grapheme-to-phoneme conversion and homograph disambiguation. Grapheme-to-phoneme extracts pronunciation of each word of the input sentence using dictionaries in conjunction with general letter-to-sound rules whereas homograph disambiguating is figuring out what tense of the word is used in an input sentence.

5.1.3.Prosodic Analysis

Prosodic analysis determines the progression of intonation, speaking rate, and loudness across an utterance, which are ultimately represented at the phoneme level as pitch (fundamental frequency), duration, and amplitude. The intonation and timing parts largely determine the rhythm of a synthetic sentence in addition to syllabic stress (expressing the prominence of a syllable relative to its surrounding syllables). Repetition of rhythm and stress patterns across sentences contributes to listener fatigue and low perceptual scores. Therefore, proper prosody facilitates listener focus, highlights important or new content, or even conveys nuances of meaning.

Timing is an important part of prosody assignment, including durations of individual phones. There are several approaches to duration modeling, but two guiding principle approaches exist. The more common rule-based approach is to start with segment intrinsic durations and moderate these durations based on rules that combine a relatively small set of different types of data, extracted by prior analyses either additively or multiplicatively (Klat, 1976). A more-recent data-driven approach is to use large speech corpora and to apply machine learning methods involving a much larger set of control factors (Schweitzer & Moebius, 2004).

Intonation refers to the tonal/melodic parts of prosody. Given a certain speaking style, summarizes the process of intonation/pitch assignment in terms of creating an F0 contour for a given utterance. Starting from a symbolic representation, for example in terms of tone and break indices (ToBI) (K. Silverman, Wightman, Price, Pierrehumbert, & Hirschberg, 1992), pauses and prosodic phrases are determined, followed by setting pitch accents and/or tones. Supra segmental prosodic attributes like pitch range (the difference between the highs and lows in the voice of a

speaker), declination (the gradual lowering of pitch over the course of a sentence), and the expression of prominence are certainly affected by the particular speaker's speaking style.

The final step towards arriving at an F0 contour uses the available symbolic prosody data to create a time series for F0. Stylized or actually observed pitch templates in a corpus might be used. A simpler approach is to use a generative model such as Fujisaki's (Fujisak, 1992) that expands pulse or step-like accent and phrase commands into continuous pitch contours, employing a source–filter model.

Finally, all of the frontend information is then forwarded to the TTS backend for synthesis. Note that TTS frontends have to be somewhat language specific. Pronunciation dictionaries and letter-to-sound rules used in the frontend may even be tailored to match the specific speaker's language accent and speech production peculiarities of the human voice that TTS tries to mimic.

5.2. TTS Backend (Speech Synthesis)

Speech synthesizer heavily depend on NLP component and traditionally distinguished as rule-based which includes articulatory synthesis and formant synthesis versus corpus-based synthesis is also known as concatenative synthesis methods.

5.2.1. Articulatory Synthesis

Articulatory synthesis is way of producing synthetic speech based on vocal tract mechanical and acoustic model by mimicking human speech production. Articulatory synthesis models control parameters that are meaningful to speech production those parameters include geometry and dynamics of articulators (jaw, tongue, lips, velum) and the glottis and forces and timings of all articulatory muscles. Von Kempelen mechanical device is considered as the older articulatory synthesizer (Taylo, 2007). This device is mimicking the human vocal tract using tubes, bellow

and pipes as source and filter to control the produced sound, because it was played like musical instrument the operator can get feedback to control playback, as a result it produces reasonable speech. The two difficulties in articulatory synthesis are firstly, deciding how to generate the control parameters from the specification, and secondly finding the right balances between a highly accurate model that closely follows human physiology and a more pragmatic model than is easy to design and control.

The articulatory static model and dynamic model is developed from dynamic movements of the articulators, the vocal-tract excitation and vocal tract geometry measurement (Coker, 1976; E.L.Saltzman & K.G.Munhall, 1989). Stone describes important detail measurement can be obtained from ultrasound picture of the tong (Stone, 1990) and the movement of the vocal tract can be tracked using nuclear magnetic resonance according to (Baer, Gore, Gracco, & Nye, 1991). A complete articulatory synthesizer would use information such as the forces and timings of all relevant groups of articulatory muscles to drive a realistic mathematical model of glottis and vocal tract. (Stone, 1990) . Articulatory synthesis produces intelligible synthetic speech, but the output is far from natural sounding and due to complexity of the model and computational requirements it is not widely adopted in practical speech (Stone, 1990).

5.2.2.Formant Synthesis

Formant synthesis was the first genuine synthesis technique to be developed and was the dominant technique until the early 1980s. It can produce quite intelligible speech but it doesn't sound natural. Formant synthesis adopts source-filter theory of speech production by encapsulating the vocal tract focus on input output characteristics with formant filters which are excited by a pulse train or with noise. The formant synthesizer makes use of the acoustic tube mode applying control filter of formant frequencies and bandwidths by relating acoustic-

phonetic properties. A typical basic layout of a formant synthesizer in shown in Figure 4 Here we see that the sound is generated from a source and pass through models of nasal cavity and oral cavity in parallel and finally combined and pass through radiation which simulate load and propagation characteristics of the lips and nose.

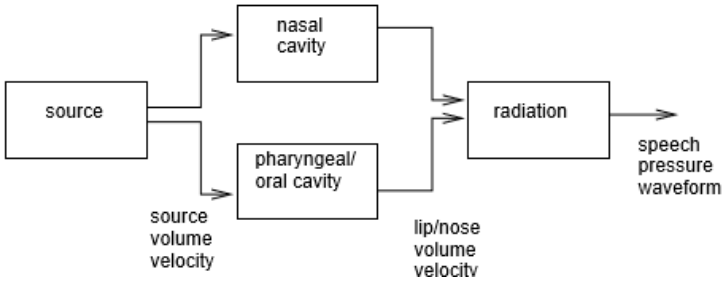


Figure 4 Formant Synthesis

The formant frequencies can be created using second order filter with transfer function in Equation 1.

Equation 1 Formant frequency transfer function

$$H = \frac{U_L}{U_G} = \frac{1}{A - CZ_L}.$$

Where: ZL is radiation impedance at the lips.

5.2.3.Series

The first type of formant synthesis developed is series or cascaded synthesizer with the amplitude of the formant is determined as part of the calculation. The shape of generated spectrum will always seem “natural” but when two formats are close the resultant values measured from the input have some unwanted effects on. Series format synthesizer can is can be represented by the figure shown on figure 5.

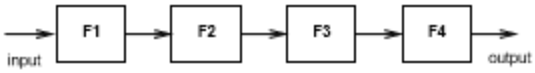


Figure 5 Series Formant Synthesis

A serial synthesizer is a good approach for open, non-nasal vocal tract configurations (e.g., for vowels and liquids). The transfer function for the parallel formant filter can be represented as multiplication of second order filter that can be represented as in equation 2.

Equation 2 Transfer Function for Parallel Synthesizer

$$\frac{U_L}{U_G} = \prod_{i=1}^{p/2} \frac{(1-z_i)(1-z_i^*)}{(1-z_i z^{-1})(1-z_i^* z^{-1})}$$
$$= \prod_{i=1}^{p/2} \frac{1-2e^{-\sigma_i T}\cos(\omega_i T)+e^{-2\sigma_i T}}{1-2e^{-\sigma_i T}\cos(\omega_i T)z^{-1}+e^{-2\sigma_i T}z^{-2}},$$

Where: the i-th root zi = e−σiT e−iωiT, the asterisk (∗) denotes the conjugate complex, and T is the sampling interval. Note that the frequency-independent numerator of each factor (i.e., the filter element) normalizes the gain of that factor to be unity at zero frequency (z=1).

5.2.4.Parallel

In the parallel formant synthesizer source signal pass through separately in each formant and the results from formants are combined at the end. With parallel formant specifying accurate frequency for each format can be taken but on the other side carful measure is required to control the amplitude can be taken as disadvantage.

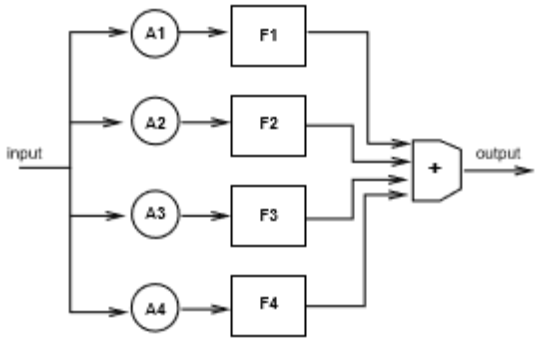


Figure 6 Forman Synthesis Parallel

A parallel formant synthesizer is good to reproduce the speech spectra of obstruent and nasals because control on individual amplitudes, formant frequencies and bandwidths. The transfer function for the parallel formant synthesis is the summation from through partial fractions as shown in equation 3.

Equation 3 Formant Synthesis with summation

$$\approx \sum_{i=1}^{p/2} \frac{A_i}{1 - 2e^{-\sigma_i T}\cos(\omega_i T)z^{-1} + e^{-2\sigma_i T}z^{-2}} .$$

Here, the residual terms Ai (the gain) and Bi (the phasing of the resonance, neglected in most practical realizations) allow for the additional control that we desire.

5.2.5. Concatenate synthesis

Concatenative synthesis woks by joining of recorded speech from a prerecorded inventory (voice database). Concatenation is done carefully to prevent the natural coarticulation, shimmer, jitter and inharmonic content of speech. Concatenative synthesis has two types, unit-selection-based synthesis which stores several instances of each unit to improve the chances of finding a well-matched unit and concatenative synthesis with fixed inventory which contains one sample for each unit and perform prosodic modification to match the required prosody.

A block diagram of a typical concatenative TTS system is shown in Fig.7. The phonetic and prosody targets provided by the frontend act as a fuzzy query in to the voiced at a base those stores all sound units. The sequence of units that optimally match the input query for a to-be-synthesized sentence are retrieved, assembled, and handed off to the speech waveform modification and synthesis module on the right. Unit assembly and waveform synthesis are both backend tasks. Note that changing the language requires changing the frontend, but usually also requires changing the store of sound units. Changing the target speaker (e.g., from male to female) also usually requires changing the store of sound units. The sound store may contain one or more of different types of speech units. For example, using phones (e.g., about 50 for English) leads to unsatisfactory joints between units because of the large coarticulatory effects between adjacent phones. Intuitively, longer units are more likely to result in higher-quality synthesis, given that the rate of concatenations (how many units to-unit transitions occur per second of speech) is lower than in the case of shorter units. For general TTS applications, however, longer units are impractical because of the tremendous multiplicity of possible unit variants (Santen, 1997). Therefore, using whole word units is impossible because of the huge demands on a voice talent that would have to read a few hundreds of thousands of words in a consistent voice and manner but also in different prosodic contexts. Given these contradictory requirements, most TTS implementations until the mid-1990s used one of two types of inventory units: the diphone or the demisyllable.

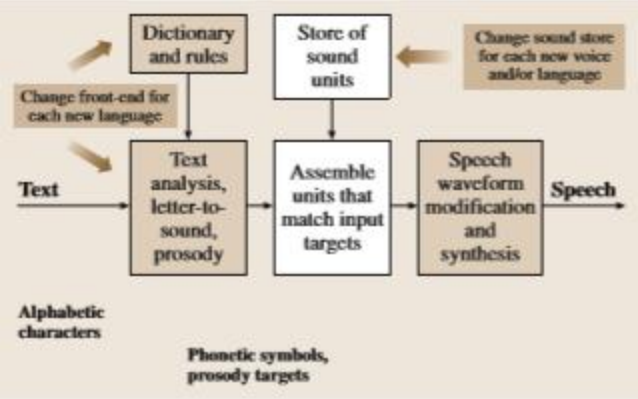


Figure 7 Concatenative Speech Synthesis

5.2.6. Concatenation Unit and Selection

In adopting concatenative speech synthesis selecting smallest unit of voice to create synthetic voice is critical and affects the performance of the synthesizer. Apart from selection of unit there should be an effective algorithm to select the unit from the unit inventory. With a minimum-size inventory, units must be modified by signal processing to match the frontend targets and to smooth over the concatenation points. After recording the voice talent, special care must be spent on selecting the most appropriate instances of each unit to be included in the inventory. Semi-automatic tools can help in finding those instances of units that minimize concatenation problems when used in later synthesis (Olive, Santen, Möbius, & Shih, 1998) . The offline (not carried out at synthesis time) process of selecting optimum units can be improved further by incorporating it in the online synthesis process. The quality of the synthesized speech should increase given that full context information is available at synthesis time, but only general statistical information exists when done offline. This idea originated at advanced telecommunications research (ATR) in Japan in the late 1980s (Y.Sagisaka, 1988).The resulting new subcategory of concatenative synthesis is now called unit-selection synthesis (Hunt & Black, 1996). It was largely enabled by

the ever-increasing power of computers. Unit-selection synthesis adds the problem of how to do fast searches of potentially millions of options to the set of tasks in speech synthesis.

5.2.7.Dipohne base synthesis

A diphone is the segment of speech that starts at the middle of one phone and extends to the middle of the next phone. It brackets exactly one phone-to-phone transition. The cut points at the middle of the phones are in the acoustically most stable region. Note that the average length of a diphone is identical to that of a phone. Because each diphone has two phone identities, the theoretical size of a diphone inventory is quadratic in the number of the phones a language has. However, because of the fact that a language might not use all of these combinations, the actual number of diphones that needs to be in the inventory is significantly smaller. For example, English requires an inventory of at least 1000 diphones (Olive, 1977).

Demisyllables are alternative units for concatenative synthesis (O. Fujimura, 1978). A demisyllable encompasses half a syllable; that is, either the syllable-initial portion up to the first half of the syllablenucleus, or the syllable-final portion starting from the second half of the syllable nucleus. The number of demisyllables in English is roughly the same as the number of diphones. Because demisyllable units are usually longer than diphones and allow for better capture of longer-term coarticulation effects compared to diphones, they should pose fewer concatenation problems.

Diphone synthesis is functionally a reduced form of unit selection synthesis. Rather than using a large database including sentences, words, and syllables, diphone synthesis only uses a database containing every diphone within the language, and therefore every phone transition. This database can have some redundancy in its recordings, but often contains only one recording of each diphone. This results in a

system with a far smaller database size than unit selection synthesis. It is vital that these diphone samples are recorded in such a way that concatenating them will not result in abnormalities in the waveform, or introduce unusual variations in pitch over the course of the sample. Thus, a speaker's voice should remain relatively constant in pitch over the diphone. Sample waveforms should also be at the same point (typically zero) at the beginning and end of the waveform, such that when concatenated the waveform remains continuous. There are many techniques for smoothing the waveform after concatenation, which can further improve the quality of the synthesized speech.

5.3. Mary TTS

Over the past few years, MaryTTS development has increasingly embraced a new, agile build platform, the previous build tools used for MaryTTS development, Gradle uses Java, but unlike them, it is extremely flexible and can be extended with new functionality and easily adapted to custom tasks. The first real proof of Gradle's flexibility and suitability for our purposes was delivered by using it to wrap and replace the Voice import Tools used for creating new MaryTTS voices (Schr¨oder, Pammi, & T¨urk., 2009). For the creation of LTS resources for new (and existing) languages, we can also use Gradle to radically streamline the process. The Wikipedia dump processing can be re-engineered by streaming the compressed dumps and converting them to plain text either in memory or via temporary files in Media Wiki format. The format conversion can nowadays be handled by many freely available libraries and tools; thanks to Gradle's flexibility, this task can also be delegated by executing externally provided programs such as Pandoc6 from within the main build process. The manual transcription of entries in the word list cannot be easily automated, but at least the user can freely choose which program to use; in the end, the list of transcribed words is simply a text file. Once prepared, this file is easy to convert to the LTS resources required for MaryTTS, using Gradle. The basic Java source code required for the NLP components for the new language, is also easy to generate and build

automatically, lifting this burden from the user's shoulders and minimizing the potential for numerous errors. Finally, the actual prompt recording is best done via mature, dedicated digital audio workstation software, and there are numerous professional-quality programs available for all major operating systems, both commercial and free.

5.3.1.Statistical Model Cost (sCost)

Calculation of sCost, labels and dynamic time warping (DTW) are used to align the spectral feature vectors from the extracted and generated cepstrum parameters and calculate an optimum path sCost measure. The criterion for finding the optimal path is the Mahalanobis distance between vectors, where the variance is the phone variance computed on the recorded waveforms. sCost is computed as the sum of the Mahalanobis distance over the optimal path, divided by the number of frames in the recorded segment and in the generated segment.

5.3.2.sCost Deployment in Mary TTS

sCost is used in the MARY TTS unit selection algorithm as follows. As usual unit selection algorithm includes target cost to define how well a unit candidate from the database matches the target unit; and concatenation cost to define how well two selected units can combine at joins. The cost functions can be written as shown in equation 4

Equation 4 Cost Function (S.Pammi & M. Charfuelan, 2013)

$$targetCost(u_i) = \langle w, c(u_i) \rangle$$
$$joinCost(u_i, u_{i-1}) = \langle w, c(u_i, u_{i-1}) \rangle$$

Where: u_i is the i^{th} unit candidate; c is the cost vector; and w is the weight vector for these features. In addition to these two costs, each unit candidate is associated with a pre computed sCost. Then the overall cost for selecting units can be expressed equation 5.

Equation 5: Overall Cost of Unit Selection (S.Pammi & M. Charfuelan, 2013)

$$totalCost(u_i) = W_1^T * \begin{pmatrix} targetCost(u_i) \\ joinCost(u_i, u_{i-1}) \\ sCost(u_i) \end{pmatrix}$$

Thus, at the stage of selecting units, a dynamic programming algorithm finds the best suitable candidates for the target by minimizing the total cost function. Beam search is used to minimize the speed of computation. Weights for the different costs are obtained heuristically (S.Pammi & M. Charfuelan, 2013).

6. EVALUAITON

Evaluation of TTS systems is critical for many reasons. One reason is to gauge the progress in the field; another is to pick the best system for a specific application. Because so many specific interests might come to bear, it is impossible to list recipes for proper evaluation procedures that would fit any scenario. Short of that, we can only give broad guidelines here. There are three quality criteria that one might be interested in: accuracy, intelligibility, and naturalness. They might overlap with each other with respect to which part of a TTS system impacts on which criteria. We define the accuracy of a TTS system as the ability to read a given input text the way a knowledgeable human reader would. For obvious reasons, accuracy is the one criterion solely

homed in the TTS frontend. Contrary to this, unsatisfactory intelligibility or naturalness is much more difficult to tie to any single component of a TTS system. Between the latter two, designers of formant synthesizers primarily aim at maximizing intelligibility because higher naturalness can be achieved with unit selection methods at the cost of higher computational complexity. Vice versa, some of the newer concatenative synthesis systems may overemphasize naturalness with the result of lower intelligibility. Accuracy may be evaluated by running a text corpus of task-relevant acronyms and abbreviations within appropriate carrier sentences through the TTS frontend and judging the generated output text. Moreover, evaluating intelligibility and naturalness requires conducting elaborate listening tests. In intelligibility tests, long known from testing speech coders, word or sentence lists are presented and subjects write down the words they hear (Spiegel, Altom, & Macch, 1990).However, evaluating the intelligibility of a TTS system clearly exposes more issues and aspects than a recovered by standard intelligibility tests that rely on word lists. On a sentence level, for example, incorrect prosody can destroy intelligibility/comprehension. So far, a generally accepted standard intelligibility test for TTS systems is lacking, although some effort has begun to address this problem (Syrdal, 2007).For overall quality evaluation, the International Telecommunication Union(ITU)recommends a specific method (ITU-T, 1994) that seems appropriate for testing naturalness. A five-point (or more) rating scale is employed for testing characteristics such as overall impression, listening effort, comprehension, etc. Another option is to ask listeners to state their preference of one option/system over another (A/B tests). Finally, selecting appropriate test material is also relevant. For example, unit-selection systems may sound very good for short sentences spoken in isolation, but show weaknesses when longer paragraphs of text are being rendered. When having a bake-off of different TTS systems, selecting test material appropriate for the intended

application seems optimal. For more details on quality testing of TTS systems, see, for example (Alvarez & Huckval, 2002).

7. RELATED WORK

As per the knowledge of the researcher, various researches have been conducted on speech synthesis on local languages. Some of related previous researches on local languages will be discussed in the following section.

7.1.1. TTS for Other Local Lanagues(Afan Oromo, Tigrgna and Wolyitgna)

Morka (Morka M., 2003), Tesfay (Tesfay, 2004) and Tewodros (Tewodros, 2009) designed diphone based concatenative speech synthesis. Morka on his paper indicated that native speakers recognized 43.33% utterance of the transcribed phonetic unit and recommended applying smoothing during transition points of diphones with spectral smoothening technique. On the other hand, Tesfaye in his prototype TTS for Tigrgna language used Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) to generate the synthetic speech. Tesfay reported MOS result of 3.05 which is closer to good by evaluating the synthesizer and recommended numeric expansion. Unlike Tesfaye, Tewodros in his prototype TTS for Wolaytta language used Residual LPC to generate speech synthesis and reported MOS value of 3.17 and 2.77 for intelligibility and naturalness respectively. Tesfay reported overall performance of the synthesizer to be 78% and recommended extensive identification and inclusion of non-standard words (NSWs) in the language with efficient representation method.

7.1.2. Concatenative Amharic TTS System

A research work by Henok (Henock, 2003), designed prototype concatenative speech synthesis TTS for Amharic based on diphones & syllables and Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) technique used to generate the synthetic speech. Henock in his research also considered prosodic effects (anger, happiness and emotion). He reported the result of Open Rhyme Test(ORT) 88% for diphone and 75% for syllable based synthesized and based on MOS test result he concluded that syllable based synthesis gives better performance.

Another Amharic language TTS system was done by Sebsibe and colleagues (Sebsibe, Black, & Sangal, 2005). They used transliteration scheme for Amharic scripts in designing a unit selection concatenative speech synthesizer. They have used Festvox with Amharic phone set, syllabification rules, letter to sound rules. Festvox is a voice building framework used for building unit selection voices in a new language Festival speech synthesis system. They have reported performance of 2.9 measured with scale 1-6. Sebsibe and colleagues recommend optimal corpus and proper selection of will give better quality. Although these works have a great deal of contribution in the area speech synthesis on Amharic language, there has not yet enough research like other languages like English. Based on the review made and as to the knowledge of the researcher none of the work has been attempted to design speech synthesizer for Amharic which can be used outside of the research community.

7.1.3. Related work using MARY TTS

Research conducted on designing speech synthesis for Luxembourgish language using MARY TTS shows that how to adopt TTS system such as MaryTTS to a new language. (Schroder & Trouvain, 2003)

M.Z Rashad and colleagues researched Arabic speech synthesis using MARY TTS they have constructed Arabic phonemes and recorded pronunciation of most frequent Arabic words .MARY TTS has been trained for these words, they have conducted two tests and found that DRT is approximately 91% and concluded the result as satisfactory (M. Z. Rashad, Hazem M. El-Bakry, & Islam R. Ismail, 2010).

Chapter Three

Amharic language phonology

This chapter presents the nature of Amharic language linguistic features and how to transcribe to its phoneme sound representation. The first part introduces Amharic language and discusses the general sound production system with special reference to Amharic language. The second and Third sub sections deals about phonetics and Amharic writing system that are related to the development of text to speech synthesis.

1. Overview of Amharic language

Amharic (አማርኛ) is an Afro-Asiatic language of the Semitic branch. Amharic official working language of Ethiopia, spoken by 22 million native speakers and 3 million emigrants outside of Ethiopia is the second most widely spoken Semitic langue in the world next to Arabic (Ethnologue, 2016). It is also used as the official or working language of some of the states within the federal system. Amharic has its own script called Ethiopic Fidel(ፊደል) grew out of the Ge'ez abugida and written left-to-write. The scripts are more or less orthographic representation of the phonemes in the language (Getahun, 2001; Baye, 1997)

2. Phonetics

In many language pronunciations is not directly related to the written text because of this some additional punctuation as indication on how to pronounce. Each language has a unique phonetic letter set and an alternate set of potential phonemes and their mixes. The quantity of phonetic symbol is somewhere in the range of 20 and 60 in each language. A group of phonemes can be identified as the base number of symbols expected to describe every possible word in a language.

In English there are around 40 phonemes. Because of multifaceted nature and diverse sort of definitions, the quantity of phonemes in English and a large portion of different languages can't be characterized precisely (Donovan, 1996).

Phonemes which is abstract unit, speaker's characteristics, emotions and contextual effects guides its pronunciation. During continuous speech, the articulatory movements guided by 30 preceding and following phonemes. The articulators are in various position based on the preceding one and they are adjusting to the following phoneme ahead of time. This causes a few minor departures from how the individual phoneme is articulated. These varieties are called allophones which are the subset of phonemes and the impact is known as co- articulation, a similar phoneme however various allophones and have diverse vocal tract configurations Another motivation behind why the phonetic representation is not flawless is that the phonetic representation is always discrete and speech signal is always continuous. In addition, it is very difficult to do phonetic notation of emotions and speaker characteristics with phonemes. which is usually defined as an acoustic representation of a phoneme (Taylo, 2007).

3. Amharic articulatory phonetics

The phonetic alphabet is usually split in two fundamental categories, vowels and consonants. Vowels are always voiced sounds and they are produced with the vocal cords in vibration, while consonants might be either voiced or unvoiced. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically. Because consonants involve very rapid changes, they are more difficult to synthesize properly. Overview of each of these major categories of Amharic phonemes is given in the following section (Baye, 1997).

3.1. Amharic vowel

Amharic language sounds inventory is composed of 39 phones set, seven vowels and thirty-two consonants. The seven vowels of Amharic can be classified into different categories depending how they are formulated based on position of tongue front or back, constriction position wideness or roundness, place of the tongue high or low, and how open or close the mouth is during articulation (Baye, 1997) .

Amharic vowels (IPA) and their categorization are shown in figure 9.

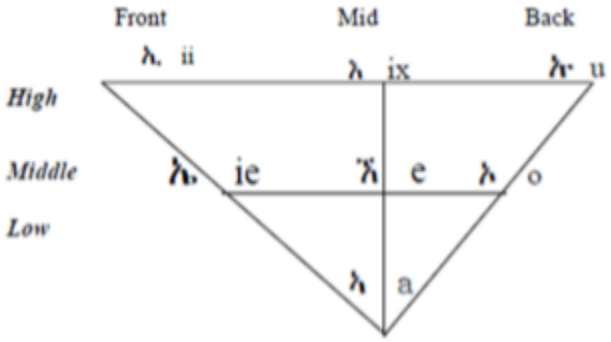


Figure 8 Amharic Vowels (sebsebi .et.al 2004)

3.2. Amharic consonants

Amharic consonants are generally classified as stops, fricatives, nasals, liquids, and semi-vowels Further classification may be made based on the place of articulation as labials (lips), dentals (teeth), alveolar (gums), palatals (palate), velars (soft palate), glottal (glottis), and labiodental (lips and teeth) (Getahun, 2001). Table3.1 shows the phonetic representation of the consonants of Amharic as to their manner of articulation, voicing, and place of articulation. Amharic consonants can be divided as stops, fricatives, nasals, liquids, and semi-vowels into the following categories depending on the place and the manner of articulation (Baye,1997; Sebsebi et al,2005):

Table 1 : Amharic Consonants (sebsebi Z. et.al 2004)

		Labials		*Alveolar*		*Palatals*		*Velars*		*Labio-Velar*		*Glottals*	
Stops	Voiceless	p	ፕ	t	ት			k	ክ	kx	ኵ	ax	ዕ
	Voiced	b	ብ	d	ድ			g	ግ	gx	ጓ		
	Glottalized	px	ጵ	tx	ጥ			q	ቅ	qx	ቋ		
Fricatives	Voiceless	f	ፍ	s	ስ	sx	ሽ					h	ሀ
	Voiced	v	ቭ	z	ዝ	zx	ዥ						
	Glottalized			xx	ጽ							hx	ጓ
Africatives	Voiceless					c	ች						
	Voiced					j	ጅ						
	Glottalized					cx	ጭ						
Nasals	Voiced	m	ም	n	ን	nx	ኝ						
Liquids	Voiced			l	ል								
				r	ራ								
Glides		w	ው			y	ይ						

1. Plosive or stop consonants: /k, p, t, g, b, d/. / ክ ፕ ት ብ ድ ግ ጓ ጵ/ The vocal tract is closed causing stop or attenuated sound. When the tract reopens, it causes burst sound.

2. Fricatives: /f, h, s/ / ፍ ሀ ስ ሽ ዥ ይ/. Amharic fricative is unvoiced and constructed by forcing air flow through narrow channel which is modified by the vocal tract resonances.

3. Nasals: /n, m, ÷/. / ም ን ኝ/The vocal tract is closed but the velum opens a route to the nasal cavity. The generated voiced sound is affected by both vocal and nasal tract.

4. Laterals: /l/ /ር ሌ/. The top of the tongue closes the vocal tract leaving a side-route for the air flow. When synthesizing consonants, better results may be achieved by synthesizing these six consonant groups with separate methods because of different acoustic characteristics they have.

4. Amharic writing system

The Amharic script is an abugida, and the grapheme of the Amharic writing system is called Fidel (Ethnologue, 2016). Each character represents a consonant and vowel sequence, but the

basic shape of each character is determined by the consonant, which is modified for the vowel. Some consonant phonemes are written by more than one series of characters. This is because these Fidel originally represented distinct sounds, but phonological changes merged them (Getahun, 2001).The citation form for each series is the consonant+ ä form, i.e. the first column of the Fidel. The Amharic script is included in Unicode, and glyphs are included in fonts available with major operating systems. present writing system of Amharic is taken from Ge'ez. Ge'ez in turn took its script from the South Arabian mainly attested in inscriptions in the Sabine dialect (Baye, 1997). Amharic did not discriminate in adopting the Ge'ez Fidel; it took all of the symbols and added some new ones that represent sounds not found in Ge'ez. These added alphabetic characters are ቸ ጨ ጀ ኘ ቨ, ሸ, ኸ, and ዠ. Currently, the orthographic representation of the language is organized Table 1. The list of all Amharic letters is depicted in Appendix A. Amharic language's writing system contains 34 base characters each of which occurs in a basic form and six other forms known as orders. The seven orders represent syllable combinations consisting of a consonant following vowel. Out of the seven derivatives six of them are CV (Consonant Vowel) combinations while the sixth is the consonant itself (Baye, 1997). Other symbols representing lateralization, numerals, and punctuation marks are also available. Therefore, having these orthographic variations for each of the 33 core letters, totally the language has more than 230 orthographic symbols; in case of Amharic character ሀ, it is represented by [ha] instead of using [he] unlike other Amharic character sets/orthographies since the sound is the same as its 4th order.

Chapter Four

TTS Algorithm

1. Introduction

In this chapter different algorithm used for modifying prosody is discussed. These algorithms are used to improve TTS systems based on concatenation of basic speech units which includes diphone, demi-syllables or phone units. These TTS systems require large database of acoustical units and totally rely on coding algorithm to compress the size of the database. The synthesis stage involves decoding and reconstructing a waveform of the acoustical unit for their coded version and concatenation of sequenced acoustical units with appropriate modification (Sagisaka, 1988).

2. Pitch Synchronous Overlap-Add Synthesis (PSOLA)

PSOLA synthesis contains three steps: pitch-synchronous analysis, pitch-synchronous modification and pitch-synchronous overlap-add synthesis. The first step produces an intermediate non-parametric representation of the signal, second step modifies intermediate non-parametric representation and last step is responsible for synthesis of modified signal from the intermediate representation.

2.1. Pitch-synchronous analysis

This step produces intermediate representation of the speech waveform $x(n)$ consists of a sequence of short-term signals $x_m(n)$, obtained by multiplying the signal by a sequence of pitch-synchronous analysis windows $h_m(n)$:

Equation 6 Pictch-syncronous analysis (Eric Moulines, 1990)

$$X_m(n) = h_m(t_m - n)d(n)$$

The windows are centered around the successive instants t_m , called pitch-marks , which are set at a pitch-synchronous rate on the voiced portions of the signal and at a constant rate on the unvoiced portions. The windows hm(n) are of Hanning type and they are always longer than one single pitch period, so that neighboring short-term signals (ST-signals) always involve a certain overlap.

2.2. Pitch-synchronous modifications

The stream of analysis ST-signals $X_m(n)$ is converted into a modified stream of synthesis ST-signals Xq(n) synchronized on a new set of synthesis pitch-marks tq. Such a conversion involves three basic operations: a modification of the number of ST-signals, a modification of the delays between the ST-signals, and possibly, a modification of the waveform of each individual ST-signal. The number of synthesis pitch-marks tq depends on the pitch-scale and time-scale modifications factors called β and γ, respectively. The delays $t_q - t_{q-1}$ between two successive pitch-marks must be equal to the local synthesis pitch period. The algorithm works out a mapping $t_q \rightarrow t_m$ between the synthesis and analysis pitch-marks, specifying which analysis ST-signal Xm(n) is to be selected to produce any given synthesis ST-signal .

2.3. Pitch-synchronous overlap-add synthesis

Several overlap-add (OLA) synthesis procedures are available to obtain the final synthetic speech. For instance, the synthetic signal £(n) can be obtained by means of the least-square overlap add synthesis scheme (Griffin & Lim, 1984):

Equation 7 least-square overlap add synthesis scheme (Griffin & Lim, 1984)

$$\bar{x}(n) = \frac{\sum_q \alpha_q \bar{x}_q(n) \bar{h}_q(\bar{t}_q - n)}{\sum_q \bar{h}_q^2(\bar{t}_q - n)},$$

Where $h_q(n)$ denotes the sequence of synthesis windows. The additional normalization factor α_q is introduced to compensate for the energy modifications related to the pitch modification procedure. The spectral interpretation of this synthesis scheme is that it minimizes the quadratic error between the spectra of the synthesis ST-signals $x_q(n)$ and the corresponding short-time spectra of the synthetic speech x(n).

3. Time-Domain Pitch Synchronous Overlap Add (TD-PSOLA)

Time domain pitch synchronous overlap and add or TD-PSOLA is widely regarded as the most popular PSOLA technique and indeed may well be the most popular algorithm overall for pitch and timing adjustment (Moulines & Charpentier, 1990).

The technique works by extracting peach periods, one frame per pitch period from voiced speech signals by identifying epochs with Hamming widow centered at maximum excitation point. Extracted and windowed signal are combined after time-scale and pitch-scale modifications. Time-scale modification to create one frame longer signal can be performed by duplicating and inserting again one frame into the sequence and finally performing overlap and add. Pitch-scale modification is performed by recombining the frames on epochs either by lowering pitch by padding with zero amplitude signals or increasing pitch by shorting the signal. We find that we generated a synthetic waveform of a different pitch. The process of TD-PSOLA is represented in

figure 11. The first panel(a) shows original, the second panel (b) shows extracted and windowed signal and the last panel(c) shows reconstructed signal with a different pitch.

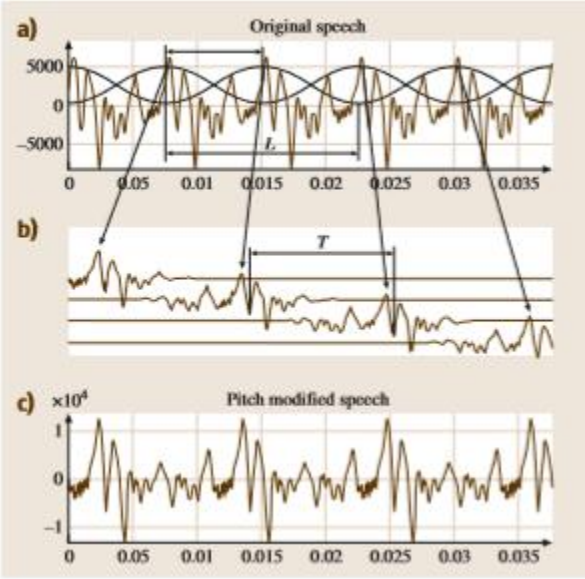


Figure 9 TD-PSOLA Process

4. LPC-Based Synthesis

Linear prediction method (LPC) uses fundamental frequency for synthesis model both reconstructing of waveforms form coded version and concatenation happen in single process.

Linear prediction synthesis adopts the all-pole vocal tract model and uses an impulse and noise source model. In LPC unlike the formant synthesis vocal tract parameters can be determining automatically from the speech itself. The development of liner prediction synthesis was really a by-product of research into speech compression or coding. Linear prediction is of course the basis of linear predictive coding (LPC) which was developed in the late 1960s and early (Stone, 1990).

CHAPTER FIVE

Experimentation and Evaluation

1. Introduction

Text to speech synthesizer using MaryTTS is developed and implemented for Amharic language. Training corpus and test dataset is prepared to implement Amharic TTS and evaluate naturalness and intelligibility based on MOS.

2. Architecture of Synthesizer

The architecture of the MARY TTS system is similar to a typical TTS architecture as described by Dutoit (Dutoit, 1997) shows the individual processing modules, the flow of information and intermediate result. MARY stands for Modular Architecture for Research on speech synthesis and it is a tool used for research, development and teaching in the field of text-to-speech [3]. The modular design of the MARY system has many advantages: it is easy to modify a certain module without affecting other modules; any module with a similar interface can be used instead or in addition to an employed module. To integrate two modules, the only requirement is that the output data types of the first module must match the input data types of the second module. Figure 3 shows the processing modules of the MARY system and the output of each module. As shown in the figure, the MARY TTS accepts two types of input text plain text and markup text (such as SABLE-annotated text or SSML-annotated text). The use of SABLE-annotated text for example as input gives the users the ability to add information to the text that improve the way it is spoken such as pausing at the right places or emphasizing on certain words [4]. This input text

is embedded into MaryXML document for the following processing steps. MaryXML is an internal, low-level markup that reflects the information provided and required by the modules of the MARY system. In other words, MaryXML markup enables the user not only to display intermediate processing results but also to modify this result so that the user can know the influence of a specific piece of information on the output of a given processing step. The processing modules of the MARY TTS can be grouped into four parts: the preprocessing or text normalization, the natural language processing, the calculation of acoustic parameters and the synthesis.

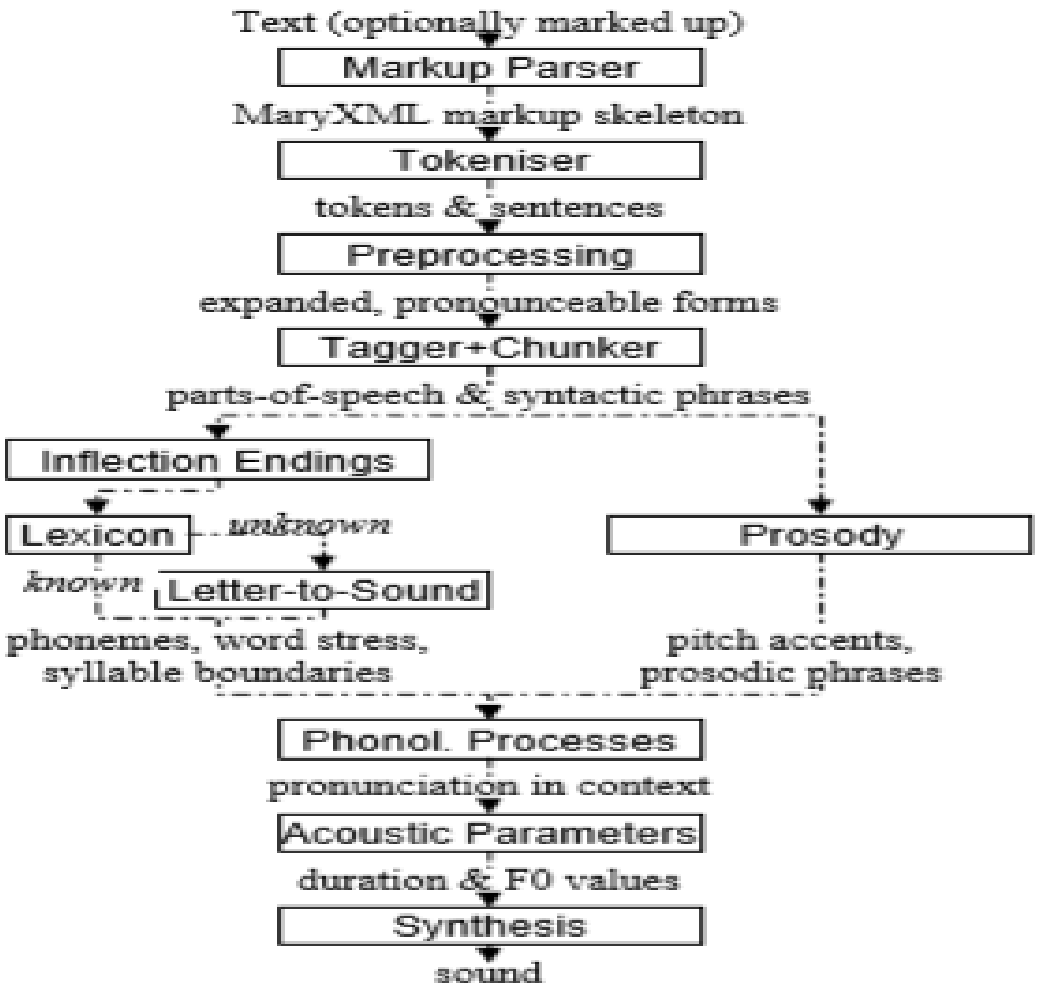


Figure 10 Architecture of MaryTTS (Marc & Jugen, 2003)

3. Corpus Preparation

Corpus preparation is one of crucial step for designing TTS system. There are four general principle during corpus preparation. The principles require the corpus to have good semantic, syntactic, prosodic and word coverage. Although applying all the principles during corpus preparation improves the performance of the TTS, only semantic coverage is mandatory to get good TTS, adding the three will improve the performance. Semantic coverage is about having at least one instance covering the domain concept, syntactic variation is about having rich syntactic variation as it is observed in natural languages, prosodic variation is about covering as many prosodic variation of the same word and finally word coverage is about corpus to cover as many domain word as possible. (M. Theune & E.Klabbers, 2001)

Identifying corpus is one of the major steps that were performed at the beginning of speech synthesis. At this stage the researcher tried to get corpus which have good sematic, syntactic, prosodic and word coverage. Considering the range and diversity of articles posted in Wikipedia the researcher believes that it contains the right semantic, syntactic, prosodic and word coverage, in addition it contains right size of corpus to extract language characteristics as it is required. Therefore, this the study has used Wikipedia Amharic articles as source corpus. To build the corpus extraction of articles from Wikipedia is required and those Wikipedia Amharic articles were downloaded using a command line tool provided by maryTTS. Wikipedia Amharic articles were download as xml dump and requires various stages to be used as corpus. To prepare suitable and usable corpus Wikipedia xml dump requires cleaning of different attributes embedded to characterize the article. To handle this download file splitting in to manageable chunk was mandatory. The conversion from xml dump to manageable chunk was done using Wikipedia dump splitter script. Finally, the usable clean text is extracted by removing Wikipedia

markup and ready to start natural language processing. At this stage we have obtained usable corpus ready to be processed. In the next section we will discuss how the natural language processing is carried out.

4. Natural Langue Processing

4.1. Text Analysis

The first step of the natural language processing is to apply text analysis on prepared corpus. Text analysis is a process of transforming text in to list words which are the smallest unit of sentence-level parser and also serve as base for morphological analyzer. Correct identification of words affects the whole text to speech synthesis process. The process of text analysis includes tokenization, end of sentence detection, recognizing punctuation marks and abbreviation etc. In this study the corpus is process started by tokenizing text in to sequence of words delimited by white space characters as recommended by McAllister ((McAllister, 1989). The text analysis was done by applying Linux script which extracts and store in MySQL database for consequent processing. Additional script was also run on MySQL to extract most frequent words which is required to create phonetic analysis. Some of identified most frequent words are shown in the table below.

NO	Word	Frequency
1	ነው	3063
2	ላይ	2205
3	እና	1986
4	ነበር	1981
5	ወደ	1565

6	ውስጥ	1237
7	ጊዜ	1190
8	ግን	1069
9	ጋር	1066
10	በኋላ	1020

Table 2 Top 10 frequent words from the Wikipedia Amharic corpus

4.2. Phonetic Analysis

The output of the text analysis is used as an input for the phonetic analysis. This analysis is done to extract phonetic representation of the corpus. This phonetic conversion was down based on word processing by transcribing input text to its phonetic representation. First most frequent words are transcribed manually in to their phonetic equivalent and the model is trained to produce phonetic equivalent of all corpus words. The generation of Lexicon for known words transcription is used and letter-to-sound rule is adopted whenever there's unknown word as depicted in the figure 11.

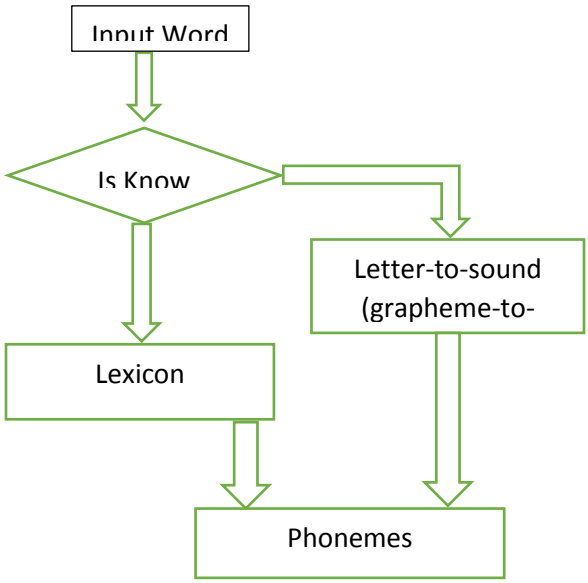


Figure 11 phonetic conversion of a word

4.3. Transcription and Romanization

Transcription of all tokenized corpus was required. The main purpose of transcription is to generate lexicon and letter-to-sound rule to train the model while TTS reading unknown words. Although the prepared corpus of Amharic language is with Amharic orthographic. due to limitation or unknown reason MARYTTS tool to was unable to process Amharic Unicode.

To overcome the limitation of the processing Amharic Unicode by MARYTTS in this research romanization of Amharic word was done. Romanization is the process of converting Amharic text into equivalent Latin character based on predefined rules. The romanization of corpus from Amharic alphabet to Latin character is performed using Unix script shown in annex 2. Sample mapping between the Amharic alphabets to Latin alphabet is shown in figure 12.

unicode2SyllabificationAmhPar.sed

```
s/ሐ/Ha/g
s/ሑ/Hu/g
s/ሀ/Hu/g
s/ሒ/Hi/g
s/ኂ/Hi/g
s/ሓ/HA/g
s/ኻ/HA/g
s/ሐ/HA/g
s/ኀ/HA/g
s/ሔ/HE/g
s/ኼ/HE/g
s/ሖ/Ho/g
s/ዞ/Ho/g
s/ሕ/He/g
s/ዕ/He/g
s/በ/be/g
s/ባ/ba/g
s/ቡ/bu/g
s/ቢ/bi/g
s/ቧ/bA/g
s/ቤ/bE/g
s/ቦ/bo/g
s/ቧ/buA/g
s/ድ/de/g
s/ዳ/da/g
s/ዱ/du/g
s/ዲ/di/g
s/ዳ/dA/g
s/ዴ/dE/g
s/ዶ/do/g
s/ዷ/duA/g
s/ፍ/fe/g
s/ፋ/fa/g
s/ፉ/fu/g
```

Figure 12 Amharic Unicode to Latin character conversion

Transcription process utilized Amharic allophones and Romanized Amharic corpus. This transcription of Romanized text is used as input for subsequent process which are also utilizing Latin character rather than Amharic alphabet. Amharic allophone was prepared based on X-SAMPA as shown in figure 14 from Amharic vowels and consonant based on their voicing characteristics as discussed by Baye (Baye, 1997) and Sebsibe (Sebsibe et.al ,2005) and used as input for transcription.

Transcription is done using semi-automatic graphical transcription tool provided by MARYTTS to train Letter-to-sound (LTS) rules of Amharic language as shown in figure 13. The transcription requires training by manually providing transcription of a word after training it predicts transcription of consequent words. During this process manually reviewing the transcription was done for validation.

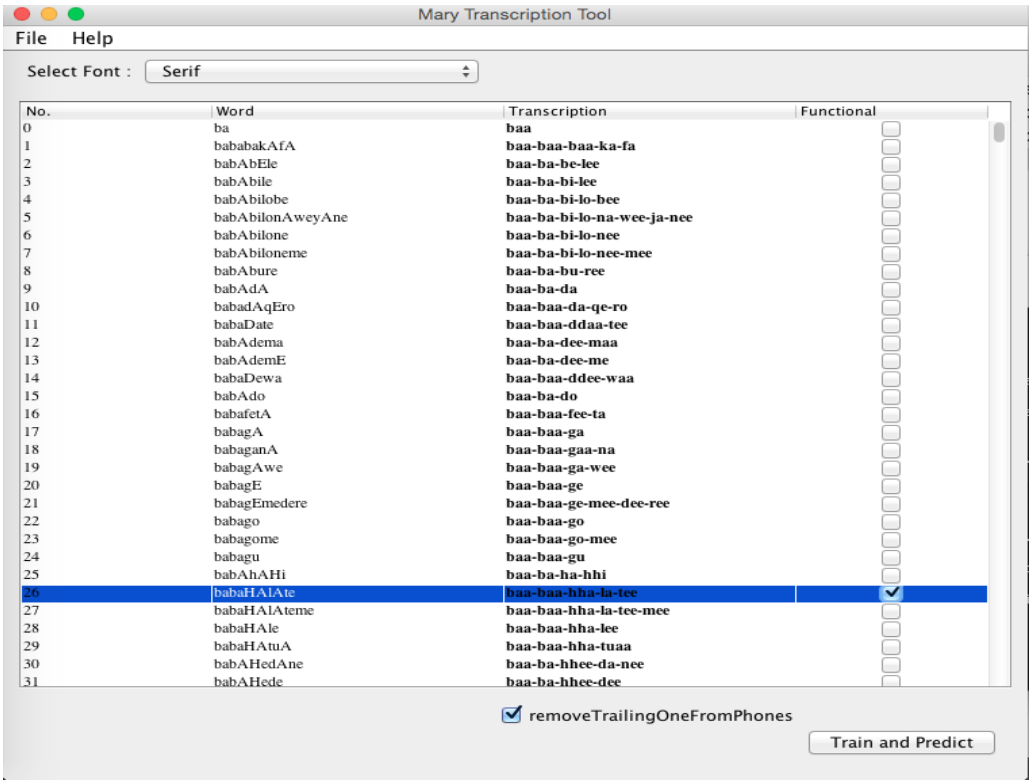


Figure 13 MaryTTS semi-automatic transcription tool

```
<allophones name="sampa" xml:lang="am_ET"
        features="vlng vheight vfront vrnd ctype cplace cvox">
    <silence ph="_"/>

    <vowel ph="a" vlng="l" vheight="2" vfront="2" vrnd="-" />
    <vowel ph="u" vlng="s" vheight="3" vfront="3" vrnd="+" />
    <vowel ph="i" vlng="s" vheight="3" vfront="1" vrnd="-"/>
    <vowel ph="aa" vlng="l" vheight="1" vfront="2" vrnd="-" />

    <vowel ph="e" vlng="l" vheight="2" vfront="1" vrnd="-" />
    <vowel ph="ee"  vlng="l" vheight="2" vfront="3" vrnd="+" />
    <vowel ph="o" vlng="l" vheight="3" vfront="2" vrnd="-" />

    <vowel ph="uee"  vlng="l" vheight="2" vfront="3" vrnd="+" />
    <vowel ph="uaa" vlng="l" vheight="1" vfront="2" vrnd="-" />

    <!-- Stop Consonants -->

    <consonant ph="p" ctype="s" cplace="l" cvox="-"/>
    <consonant ph="b" ctype="s" cplace="l" cvox="+"/>
    <consonant ph="pp" ctype="s" cplace="l" cvox="-"/>
    <consonant ph="t" ctype="s" cplace="a" cvox="-"/>
    <consonant ph="d" ctype="s" cplace="a" cvox="+"/>
    <consonant ph="tt" ctype="s" cplace="a" cvox="-"/>
    <consonant ph="k" ctype="s" cplace="a" cvox="-"/>
    <consonant ph="g" ctype="s" cplace="a" cvox="+"/>
    <consonant ph="q" ctype="s" cplace="a" cvox="-"/>
    <consonant ph="kk" ctype="s" cplace="a" cvox="-"/>
    <consonant ph="hh" ctype="s" cplace="g" cvox="-"/>

  <!-- fractive Consonants -->

  <consonant ph="f" ctype="f" cplace="l" cvox="-"/>
  <consonant ph="v" ctype="f" cplace="l" cvox="+"/>
  <consonant ph="s" ctype="f" cplace="a" cvox="-"/>
  <consonant ph="z" ctype="f" cplace="a" cvox="+"/>
  <consonant ph="x" ctype="f" cplace="a" cvox="-"/>
  <consonant ph="ss" ctype="f" cplace="p" cvox="-"/>
  <consonant ph="zz" ctype="f" cplace="p" cvox="+"/>
  <consonant ph="h" ctype="f" cplace="g" cvox="-"/>

      <!--Afractive Consonants -->
    <consonant ph="c" ctype="a" cplace="p" cvox="-"/>
    <consonant ph="dd" ctype="a" cplace="p" cvox="-"/>
    <consonant ph="cc" ctype="a" cplace="p" cvox="-"/>

    <!-- Glides Consonants -->
  <consonant ph="w" ctype="r" cplace="l" cvox="-"/>
  <consonant ph="j" ctype="r" cplace="p" cvox="-"/>

      <!-- Nasal Consonants -->
    <consonant ph="m" ctype="n" cplace="l" cvox="+"/>
    <consonant ph="n" ctype="n" cplace="a" cvox="+"/>
    <consonant ph="nn" ctype="n" cplace="p" cvox="+"/>

  <!-- Liquids Consonants -->
<consonant ph="l" ctype="l" cplace="a" cvox="+"/>
<consonant ph="r" ctype="l" cplace="a" cvox="+"/>
```

Figure 14 Amharic Phone Set based on SAMPA

The transcription produces lexicon which is used as pronunciation dictionary as shown in figure 15 while generating synthetic sound from word to produce phonetic representation. In addition, it

contains training letter-to-sound rules for generating phonetic representing for unknow word and list of functional word to be use by primitive part of speech tagger.

```
ba|b aa
bababakAfA|b aa - b aa - b aa - k a - f a
babAbEle|b aa - b a - b e - l ee
babAbile|b aa - b a - b i - l ee
babAbilobe|b aa - b a - b i - l o - b ee
babAbilonAweyAne|b aa - b a - b i - l o - n a - w ee - j a - n ee
babAbilone|b aa - b a - b i - l o - n ee
babAbiloneme|b aa - b a - b i - l o - n ee - m ee
babAbure|b aa - b a - b u - r ee
babAdA|b aa - b a - d a
babadAqEro|b aa - b aa - d a - q e - r o
babaDate|b aa - b aa - dd aa - t ee
babAdema|b aa - b a - d ee - m aa
babAdemE|b aa - b a - d ee - m e
babaDewa|b aa - b aa - dd ee - w aa
babAdo|b aa - b a - d o
babafetA|b aa - b aa - f ee - t a
babagA|b aa - b aa - g a
babaganA|b aa - b aa - g aa - n a
babagAwe|b aa - b aa - g a - w ee
babagE|b aa - b aa - g e
babagEmedere|b aa - b aa - g e - m ee - d ee - r ee
babago|b aa - b aa - g o
babagome|b aa - b aa - g o - m ee
babagu|b aa - b aa - g u
babAhAHi|b aa - b a - h a - hh i
babaHAlAte|b aa - b aa - hh a - l a - t ee
babaHAlAteme|b aa - b aa - hh a - l a - t ee - m ee
babaHAle|b aa - b aa - hh a - l ee
babaHAtuA|b aa - b aa - hh a - t uaa
babAHedAne|b aa - b a - hh ee - d a - n ee
babAHede|b aa - b a - hh ee - d ee
babAhelAcawe|b aa - b a - h ee - l a - c aa - w ee
babAhelAcene|b aa - b a - h ee - l a - c ee - n ee
```

Figure 15 Amharic pronunciation dictionaries

5. Minimal NLP components

With all files generated from transcription tool which includes allophones.am.xml, am.lts, am_lexicon.fst and am_pos.fst , we have created Amharic TTS system in MaryTTS platform. The following tools are used to create Amharic minimal NLP module:

1. Operating system used was Ubuntu 14.04
2. Eclipse IDE is used as development platform
3. Maven is used build technology

Default text file is created for Amharic langue, minimal java file AMConfig.java and finally using the maven the project is build and Amharic language NLP module for MaryTTS built. This NLP module will be used in feature extraction from acoustic data during creation of model and in actual Amharic TTS in action.

6. Digital Signal Processing (DSP)

Speech Representation and manipulation started by first dividing the clean text into sentences and labeling the sentence reliable or non-reliable (sentences with unknown words or strange symbols). The cleaning process generated 408 sentences from the selected corpus. In addition, extraction of context features using MARY feature maker to extract phone and next phone from reliable sentences was carried out, those extracted features will be used to identify diphone. The extracted feature shown in the figure 16 will used for the selection of sentence to be recorded.

```
ByteValuedFeatureProcessors
phone 0 C N T _ a ax b bx c cx d e f fx g gx h hx ie ii ix j k kx l m mx n nx o p px
      q qx r rx s sx t tx u v w xx y z zx
next_phone 0 C N T _ a ax b bx c cx d e f fx g gx h hx ie ii ix j k kx l m mx n nx o
           p px q qx r rx s sx t tx u v w xx y z zx
```

Figure 16 Features extracted from reliable sentences

After completing the feature extracting selection of sentences for recording that provides a good diphone and prosodic coverage is required. This task was carried out using MARY voice creation toolkit, an algorithm using greedy methods is used for selection of 120 sentences with optimized coverage of diphone out of 408 sentences. Three parameters are taken into account: the units, coverage definition and stop criteria. Units are defined as vectors consisting of three features: phone, next phone and prosody property. The definition of coverage fixes what kind of units is wanted in the final set; in the current version all diphones and their prosodic variation are used. Other aspects like frequency weights, sentence length, features weight, etc. can be set for

optimizing the coverage. The stop criteria are a combination of number of sentences, maximum diphone coverage and maximum prosody coverage (Hunecke, 2007) . As result of sentence selection 120 phonetically and prosodically balanced sentences were selected and are ready to recorded.

6.1. Recording Selected Sentence

The main objective of recording selected sentence is to extract diphones and get as uniform a set of pronunciations as possible. Because of difficulty to gain access to professional studio recording was conducted in house late night to avoid background noise during recording. Recording of selected voice done using Redstart recording tool shown in figure 17, first by importing selected sentence in to the tool. The sentences are then rendered one by one with option to rerecord with iteration. Native Amharic speaker with BSc degree was selected to for recording and advised to keep emotion during reading to neutral and voice at constant volume to get the maximum natural language during concatenation.

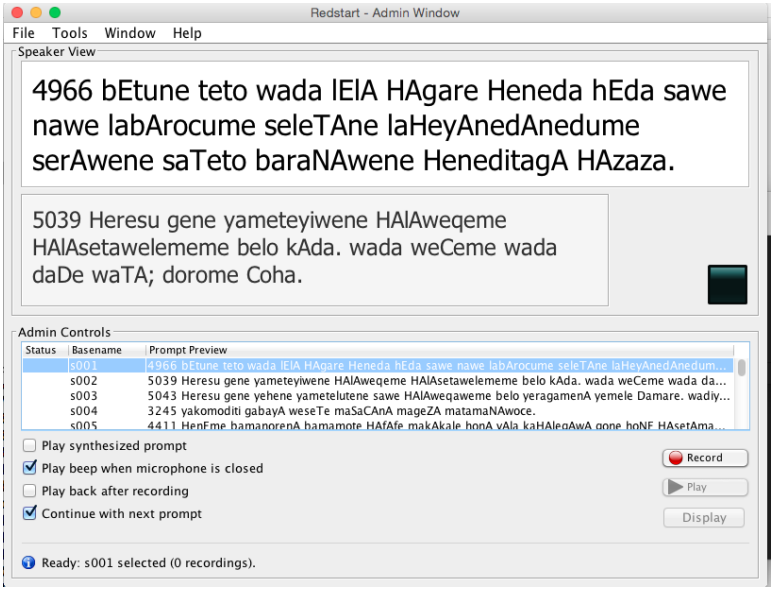


Figure 17 RedStart Voice Recording Tool

After recording is completed, recorded sentence is verified by playing and viewing wave form for speech signal and corresponding pitch and energy contours and rerecord ash shown in the figure 18. If not satisfied with recorded sentence it was re-recorded to get accurate result as possible.

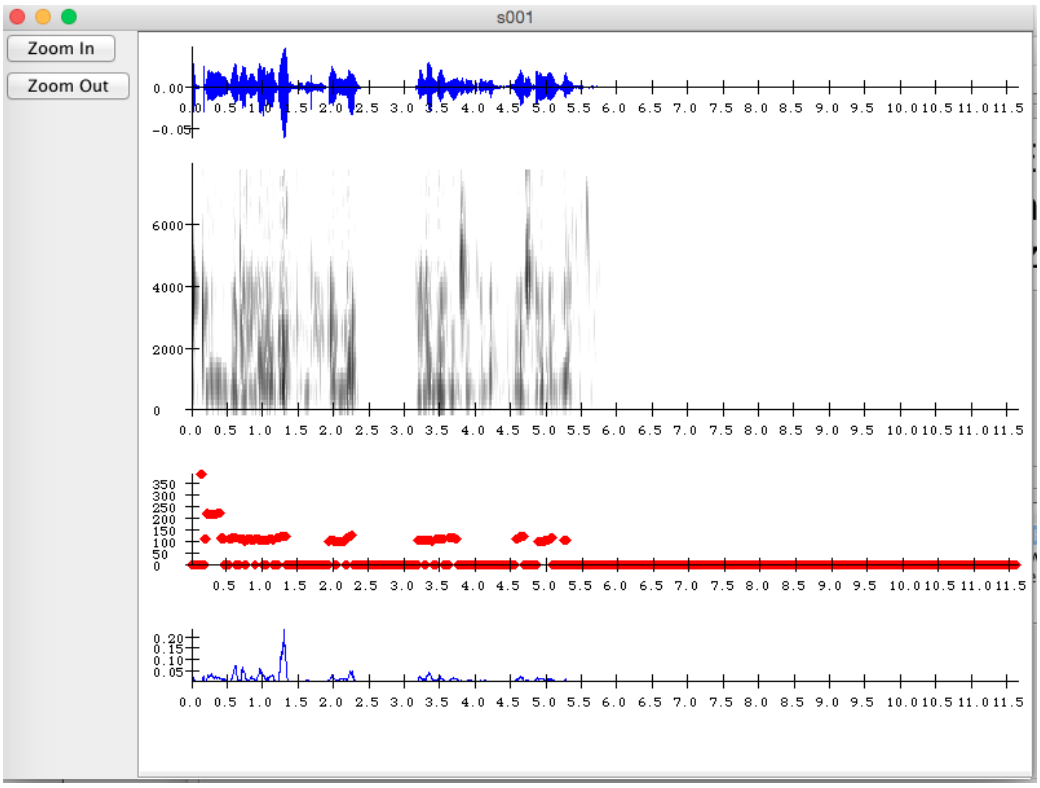


Figure 18 Recorded Voice Display form

Recorded audio is converted to sample rate of 16 khz, converted to mono, before building a synthetic voice global amplitude scaling regardless of recording amplitude is done to control the maximum amplitude recording, in addition Power normalisation across recording sessions attempts to identify recording sessions by the time stamps of files: a pause longer than 10 minutes indicates a session break. For each session separately, a mean energy is computed, and conversion factors for each file are computed such that after the conversion, the average energy for all sessions is the same. The aim behind this processing is to compensate for the case that

from one session to another, there may have been slightly different recording gains or minor differences in the speaker's distance to the microphone.

Noise from recorded voice was removed by applying high-pass FIR filter with a cut-off frequency of 50 Hz and a transition bandwidth of 40 Hz. And finally, trim initial and final silences by applying k-means clustering to identify silence vs. speech portions of the audio file, leaving 0.5 seconds initial and final silence. This is useful to avoid training absurdly long pause duration models.

7. Diphone Unit Selection Speech Synthesis

7.1. Feature Extraction

At the core of the voice building process, the recorded speech data is converted to a feature representation. It is this feature representation which allows the use of machine learning techniques to train models to predict prosody and/or vocoder parameters from text during the actual TTS process in the runtime system. The feature extraction stage of the voice building process yields a combination of frame-wise feature vectors from acoustic analysis of the audio, and time-aligned symbolic features based on linguistic analysis of the corresponding text.

To process feature extraction, MARY voice import tool was used which combine an extensible high quality freely available components specialised in a particular task including r Festvox EHMM for automatic labelling, HTS for training HMM models and Praat for feature extraction etc.

7.2. Feature Extraction from Acoustic Data

Acoustic features include F0, tracked using Praat (Boersma, 2001), and mel-frequency cepstral coefficients (MFCCs), extracted using the Edinburgh Speech Tools (EST) (King et al., 2003). The linguistic features are obtained depending on the MaryTTS language component for the corresponding language. For pitch extraction we use Praat (Boersma, 2001) markers and also for correcting, aligning pitch marks to the nearest Zero Crossings. Praat is configured to use min pitch of 50 and max pitch of 200. For pitch synchronous MFCC extraction from Speech Wave files we use the Edinburgh Speech tools (EST) (King et al., 2003).

7.3. Automatic Labelling

In synthesis is process of finding spoken phone in the voice signal from the sequence of phone at hand. The primary focus is to identify the boundaries of segments. Here we have recorded voice and we will train full acoustic HMM models on the recorded data for automatic labelling. HTK ehmm is used for automatic labelling (Black and Lenzo, 2007) or the Sphinx (Huang et al., 1993; Walker et al., 2004) labelling tool. This process produces label files. Those files are inspected using label file inspector which allows browsing through aligned labels and listening to the corresponding wave file. It is useful for perceptual manual verification on alignment.

7.4. Feature Vector Extraction from Text Data

The linguistic features extracted and assigned to the feature vectors include several related to phonology (e.g., distinctive features, position in the syllable, stress, accent), syntax (e.g., POS, distance to phrase and sentence boundaries), and — optionally — speaking style (Steiner et al., 2010; Charfuelan and Steiner, 2013), information density (Le Maguer et al., 2016), or other high-level context features. Linguistic features were extracted in section 5 processing which includes text analysis, phonetic analysis and prosodic analysis.

Extracted and label feature completed additional process is required to verify if labels and the features vectors aligned correctly. If the alignment fails automatic pause correction is carried out. This is achieved by deleting pauses which are in the label file but not in the feature file and inserting pauses with length zero which are in the feature file but missed in the label file

7.5. Building acoustic models

After completing the both acoustic and linguistic feature extraction the next step is building acoustic models from extracted data. Before running the model, generation trainer phone feature file was prepared using 'phonefearurefielwiter' tool to produce file containing target cost features for the phone. For model generation Classification and Regression Tree (CART) models used by the unit selection run-time system are trained with the 'wagon' tool from Edinburgh Speech Tools Library. Finally, CARTF0Modeller components are used for building the acoustic models for F0 and duration.

The MARY unit selection algorithm combines the usual steps of pre-selecting candidate units, a dynamic programming phase combining weighted join costs and target costs, and a concatenation phase joining the selected units into an output audio stream. In the current version, a very small pre-selection tree is manually specified and can pre-select units, e.g., by their phone or diphone identity (Schroder et al., 2009). A beam search is used in the dynamic programming step to keep processing time low. After successful execution of the unit selection voice building components, an installation component will copy all the necessary unit selection voice files of the new voice to the MARY TTS platform.

7.6. Adding a new unit selection voice in MARY

Voice Compiler compiles the voice to be used in MARY TTS. The default setting values of this component are already fixed. After compiling the voice, it is imported to MARY TTS server.

8. Phone Unit Selection Synthesis

The system performs a selection of phone-sized units, consisting of a CART-based pre-selection of candidates, a dynamic programming phase in which target and join costs determine an optimal unit chain, and a concatenation and signal post-processing stage in which the audio for the selected units is generated, concatenated and post-processed to smooth the F0 contour. In processing unit selection all the process carried out for diphone synthesis are similar expect extraction half phone feature which is used in diphone synthesis full phone feature extraction was carried out. The resulting unit selection synthesizer based on phone is added in to mary tts platform.

9. Test and Evaluation

After integrating the compiled Amharic TTS in to MaryTTS the performance of the synthesizer is evaluated for standard words which can be generated either using letter-to-sound rules trained to do or from lexicon. The performance of the TTS is evaluated using randomly selected 4 sentences from BBC Amharic containing 51 words using MOS. (1), poor (2), fair (3), good (4) and excellent (5).

Table 3 MOS Scale

MOS Scale	Description
5	Excellent

4	Very Good
3	Good
2	Fair
1	Poor

MOS (mean Opinion Score) is five scale measurements as shown in Table 2. To asses intelligibility and naturalness five Amharic language native speakers was selected. Four sentences are played and users are requested to write what they have heard from the synthesizer. Then the number of words correctly identified from the sentence is calculated against the original sentence and converted to the scale of 5 for intelligibility. For measuring naturalness Categorical Estimation (CE) (Kraft V., 1995) is used , the listeners were to rate the synthetic voice they have heard based speed, pronunciation and stress on scale of 5.

Additional test to measure intelligibility is done using Diagnostic Rhyme Test (DRT). In this test, twenty pairs of words that differ only in a single consonant are uttered and the listeners are asked to mark on an answer sheet which word of each pair of the words they think is correct (Maria, 2004).

9.1. Diphone Synthesis

The evaluation result for ineligibility and naturalness is provided in table 2 and table 3 respectively.

Table 4 Ineligibility Evaluation Result based on MOS scale

	Sentence 1	Sentence 2	Sentence 3	Sentence 4	
Evaluator 1	3	3	4	4	3.5

Evaluator 2	4	4	3	4	3.75
Evaluator 3	4	3	4	3	3.5
Evaluator 4	3	4	4	4	3.75
Evaluator 5	4	3	4	3	3.5
Mean Average					3.6

Table 5 Naturalness Evaluation Result based on MOS scale

	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Average
Evaluator 1	3	2	3	3	2.75
Evaluator 2	3	3	2	3	2.75
Evaluator 3	3	2	3	2	2.25
Evaluator 4	2	3	3	3	2.75
Evaluator 5	3	3	2	3	2.75
Mean Average					2.65

Additional test to identify the accuracy of the developed Amharic TTS was carried out on 20 words and 5 evaluators. The result is shown in Table 5 below.

Table 6 Performance of Amharic TTS

	Total No Words	Correctly Identified	% Accuracy
Evaluator 1	10	9	90
Evaluator 2	10	9	90

Evaluator 3	10	10	100
Evaluator 4	10	9	90
Evaluator 5	10	10	100
Mean Accuracy			94

The result shows that generated synthetic speech have intelligibility 3.6 and naturalness 2.65 Table 1 and Table 2 respectively. The value of intelligibility and naturalness shows that the Amharic synthesizer performed very well for both. It is observed that some of the phones have high and low due to recording quality. Insignificant noise is observer during concatenation of diaphones.

9.2. Unit Selection Synthesis(phone)

Unit selection synthesizer performance was measured for the same test sentence used to evaluate the diphone synthesizer. The result of the evaluation is shown in the table 7 for naturalness and table 8 for intelligibility. Additionally, the performance of the synthesizer also measured with similar list of words as it is done for the diphone result is shown in table 9.

Table 7 Ineligibility Evaluation Result based on MOS scale

	Sentence 1	Sentence 2	Sentence 3	Sentence 4	
Evaluator 1	3	3	3	3	3.0
Evaluator 2	3	2	3	3	2.75
Evaluator 3	3	3	3	3	3.0
Evaluator 4	2	2	3	3	2.75
Evaluator 5	3	3	3	3	3.0

Mean Average	2.9

Table 8 Naturalness Evaluation Result based on MOS scale

	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Average
Evaluator 1	3	2	2	3	2.50
Evaluator 2	2	3	2	3	2.50
Evaluator 3	3	2	3	2	2.50
Evaluator 4	2	3	2	3	2.50
Evaluator 5	3	3	2	3	2.75
Mean Average					2.55

Table 9 Performance of Amharic TTS

	Total No Words	Correctly Identified	% Accuracy
Evaluator 1	10	8	80
Evaluator 2	10	9	90
Evaluator 3	10	9	90
Evaluator 4	10	8	80
Evaluator 5	10	8	80
Mean Accuracy			84

CHAPTER SIX

Conclusion and Recommendation

In this thesis Amharic speech synthesis was developed and evaluated considering only standard words. The developed TTS is based on MaryTTS multilingual speech synthesis tool. The major steps of this work include preparation of corpus and text analysis for generation of pronunciation lexicon and to rain letter-to-sound rule, designing minimal NLP component for the Amharic language, recording and processing speech for feature extraction and digital signal processing for concatenation of diphones and concatenation of phones.

1. Conclusion

The evaluation of the diphone TTS system shows intelligibility of 3.6 and naturalness of 2.65 MOS score with overall average accuracy of 94%. In addition, the evaluation of unit selection synthesis shows 2.9 and 2.55 for intelligibility and naturalness respectively with performance of 84%. This clearly shows diphone base speech synthesis out performs unit selection speech synthesis. The performance of the Diphone TTS is promising and with current web access capability of MaryTTs the generated Amharic TTS tool can be used to read articles with standard

words. Improvement with intelligibility and naturalness can be improved with proper professional sound recording environment and device to maintain constant audio throughout the recording session.

2. Recommendation

Developed Amharic TTS system can further be expanded by introducing NLP component to process non-standard words and part of speech tagger, Prosody. etc. Lack of Amharic Grapheme-to-Phoneme processor was one of the issues observed during the research, further works can be done in designing Amharic grapheme-to-phoneme processor so that to include more unknown words in the TTS and to train letter to sound.

REFERENCE

A, H. (2007). *Optimal design of a speech database for unit selection synthesis.* Allgemeine Linguistik: Universit des Saarlandes.

Allen, J., Hunnicut, S. a., & Klatt, D. (1987). *From Text to Speech: the MITalk System.* Cambridge: Cambridge University Press.

Alvarez, Y., & Huckval, M. (2002). The reliability of the ITU-T P.85 standard for the evaluation of text-tospeech systems. *Proc. ICSLP 2002*, (pp. 329–332).

Baer, T., Gore, J., Gracco, L., & Nye, P. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *J. Acoust. Soc. Am. 90*, (pp. 799–828).

Baye, Y. (1997). *የአማርኛሰዋሰውAmharic Grammar.* Addis Ababa, Ethiopia.

Black, A. W., & Lenzo, K. (2007). *Festvox: Building synthetic voices, Version 2.1*. Retrieved March 2010, from http://www.festvox.org/bsv/

Boersma, P. (2001). Praat, a system for doing phonetics by compute. (pp. 5(9/10):341–345). Glot International.

Coker, C. (1976). A model of articulatory dynamics and control. *Proc. IEEE 64*, (pp. 452–459).

Donovan. (1996). Trainable Speech Synthesis‖ PhD Dissertation. University of Cambridge, UK,1996.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis.* Dordrecht: Kluwer.

E.L.Saltzman, & K.G.Munhall. (1989). Adynamical approach to gestural patterning in speech production. *Ecol. Psychol 1(4)*, (pp. 333–382).

Ethnologue. (2016). *Languages of the World*. Retrieved 2004, from Ethnologue: http://www.ethnologue

Flanagan, J. (1972). Speech analysis ,synthesis and perception. *Springer-Verlag*.

Fujisak, H. (1992). The role of quantitative modeling in the study of intonation. *Proc. Int. Symp. Japanese Prosody*, (pp. pp.163–17).

Getahun, A. (2001). Modern Amharic Grammer vol 9. Addis Ababa, Ethiopia: Alpha printing press.

Henock, L. (2003). *"Concatenative Text-to-Speech(TTS) synthesis for Amharic language".* Addis Ababa, Ethiopia: Addis Ababa University.

Huang, X., Alleva, F., Hon, H., Hwang, M., & Lee, K. (1993). The SPHINX-II speech recognition system: an overview. *Computer Speech & Language*, 7(2):137–148.

Hunt, A., & Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *Proc. ICASSP 96*, (pp. 373–376).

ITU-T. (1994). A Method for Subjective Performance Assessment of the Quality of Speech Output Devices (International Telecommunications Union, Geneva 1994). *Recommendation P.85.*

K. Silverman, M. B., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). TOBI: A standard for labeling English prosody. *Proc. ICSLP'92 Banff* , (pp. pp.867–870).

Keller, E. (2002). Toward greater naturalness: future directions of research in speech synthesis. In E. Keller, G. Bailly, A. Monaghan, J. Terken, & M. Huckval, *Improvements in Speech Synthesis. Chichester*. Wiley.

Klat, D. (1976). Linguistic use of segmental duration in English: Acoustic and perceptual evidence. *J. Acoust. Soc. Am. 59*, 1208–1221 .

klatt, D. (1987). Review of text-to-speech conversion for english. *Journal of the Acoustical Society of America*, 82,3.

Lemmetty, S. (1999). *Review of SPeech Synthesis Technology.* Addis Ababa ,Ethiopia: Addis Ababa , University.

Louw, A. (2004). A Short Guide to Pitch-marking in the Festival Speech Synthesis System and Recommendations for Improvements. *Local Language Speech Technology Initiative*. Pretoria, south Africa: CSIR.

M. Z. Rashad, Hazem M. El-Bakry, & Islam R. Ismail. (2010). "Diphone Speech Synthesis System for Arabic Using MARY TTS". *International journal of computer science & information technology(IJCSIT)*, VOL.2,No.4.

Marc, S., & Jugen, T. (2003). The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching. *Intrnation Journal of Speech Technology , 6-4, 365-377 (2003)*(6-4), 365-377.

Meshesha, M., & Tafere, A. (2010, July 1). *A GENERALIZED APPROACH TO AMHARIC TEXT-TO-SPEECH (TTS) SYNTHESIS SYSTEM* . Retrieved NOV 10, 2017, from Addis Ababa University Digital Repository: http://hdl.handle.net/123456789/8362

Morka M. (2003). "Text-To-Speech System for Afan Oromo" MSc Thesis. Addis Ababa,Ethiopia: Addis Ababa University.

Nadew, T. (2008). *Formant based speech synthesis for Amahric vowels.* Addis Ababa, Ethiopia: Addis Ababa University.

Olive, J. (1977). Rule synthesis of speech from diadic units. *Proc. ICASSP 77*, (pp. 568–570).

Olive, J., Santen, J. v., Möbius, B., & Shih, C. (1998). Synthesis. In: Multilingual Text-to-Speech Synthesis – The Bell Labs Approach ed. by. In R. Sproat. Dordrecht: Kluwer Academic.

S.Pammi, & M. Charfuelan. (2013). HMM-based sCost quality control for unit selection speech synthesis. *ISCA Speech Synthesis Workshop.* Barcelona , Spain.

Schr¨oder, M., Pammi, S., & T¨urk., O. (2009). Multilingual MARY TTS participation in the Blizzard Challenge. *In Proc. of the Blizzard Challenge 2009.*

Schroder, M., & Trouvain, J. (2003). "The German text-to-speech synthesis system MARY: A tool for reaserch, developmenent and teaching",. *International Journal of Speech Technology*, Vol.6, pp.365-377.

Schweitzer, A., & Moebius, B. (2004). *Exemplar-based production of prosody: Evidence from segment and syllable durations, Proc. Speech Prosody 2004 (Nara).* (ISCA, Grenoble : Marlien .

Sebsibe, K., Black, K., & Sangal. (2005). Unit Selection Voice for Amharic using Festivox. *5th ISCA Speech Synthesis Workshop*, (pp. page 103-107). Pittsburg.

Spiegel, M., Altom, M., & Macch, M. (1990). Comprehensive assessment of the telephone intelligibility of synthesized and natural speech. *Speech Commun. 9*, (pp. 279–291).

Stan, R. (2011). *hmm-based text-to-speech synthesis with interactive intonation optimization.* Cluj-Napoca: Technical University of Cluj-Napoca.

Stone, M. (1990). A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data. *J. Acoust. Soc. Am. 87*, (pp. 2207–2217).

Syrdal, A. (2007). Development of a standard for the evaluation of intelligibility of text-to-speech synthesis systems by ANSI Accredited Standards Committee S3. *Bioacoustics, working group S3/WG 91, Text-to-Speech Synthesis Systems, Personal communication.*

Taylo, P. (2007). Text-to-Speech Synthesis. University of Cambridge. Book, 2007 .

Teferra, S., & Demessie, B. (2017, June 1). *Syllable-Based Amharic Speech Synthesis (TTS) Using Hmm* . Retrieved November 10, 2017, from Addis Ababa University Digtal Repository: http://hdl.handle.net/123456789/16356

Tesfay, Y. (2004). "Diaphone Based Text-To-Speech Syntheis System for Tigrigna Language", MSc thesis. *MSc thesis.* Addis Ababa, Ethiopia: Addis Ababa University.

Tewodros, A. (2009). *"Text-to-Speech Synthesizer for Wolaytta Language".* Addis Ababa, Ethiopia: Addis Ababa , University.

Walker, W., Lamere, P., & Kwok, P. (2005). *Freetts 1.2*. Retrieved 2019, from http://freetts.sourceforge.net.

Y.Sagisaka. (1988). Speechsynthesisbyruleusinganoptimalselectionofnon-uniformsynthesisunits. *Proc. ICASSP 88*, (pp. 679–682).

APPENDIXS

APPENDIX A: Amharic Alphabets with Their Seven Orders

First	Second	Third	Fourth	Fifth	Sixth	Seventh
ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

APPENDIX B: Sentences Used to Test Amharic TTS

1. ይርጋ ዱባለ የተወለደው ጎንደር ከተማ አርባያ በምትባል መንደር ነው።

2. የሰው ልጅ በአንድነት መኖር ከጀመረበት ጊዜ አንስቶ የግዝፈቱ ደረጃ ይለያይ እንጂ አስተዳደርና መንግስት አለ።

3. በቱንዝያ እና በግብጽ የተነሳው ሕዝባዊ አብዮት የሁለቱን አገሮች አምባገነን ገዥዎች ገለበጠ።

4. በተለያዩ የኮሜዲ ስራዎቹ የሚታወቀው ኮሜዲያን መስከረም በቀለ ከሁለት አስርት አመታት በኋላ ለመጀመሪያ ጊዜ ወደ አገሩ ገባ።

APPENDIX C: Words Used to Test Amharic TTS

No	Word1	Word2
1	በረኛ	በተኛ
2	መተኛ	መገኛ
3	ተገኘ	ተመኘ
4	አደረ	አረረ
5	ዘበኛ	ዘረኛ
6	ትደር	ትመር
7	ገለጠ	ገመጠ
8	ስለላ	ስበላ
9	ስጠጣ	ስመጣ
10	ጫጫታ	ጫዋታ

APPENDIX D: Romanization of Amharic script used in this document

ሸ	Ha	ኀo	o
ኡ	Hu	ኈo	o
ዑ	Hu	ጕui	ui
ኢ	Hi	ጓuA	uA
ዒ	Hi	ጔuE	uE
አ	HA	ጕu	u
ኣ	HA	ሀ	he
ዐ	HA	ሐ	he
ዓ	HA	ኀ	he
ኤ	HE	ኽ	he
ዔ	HE	ኸ	ha
ኦ	Ho	ሁ	hu
ዖ	Ho	ሑ	hu
እ	He	ኁ	hu
ዕ	He	ኹ	hu
ቨ	be	ሂ	hi
በ	ba	ሒ	hi
ቡ	bu	ኂ	hi
ቢ	bi	ኺ	hi
ባ	bA	ሀ	hA
ቤ	bE	ሃ	hA
ቦ	bo	ሐ	hA

ቧ	buA	ሐ	hA
ድ	de	ኀ	hA
ደ	da	ኃ	hA
ዱ	du	ኻ	hA
ዲ	di	ሄ	hE
ዳ	dA	ሔ	hE
ዴ	dE	ኄ	hE
ዶ	do	ኼ	hE
ዷ	duA	ሆ	ho
ፍ	fe	ሖ	ho
ፈ	fa	ኆ	ho
ፉ	fu	ኾ	ho
ፊ	fi	ኼ	ho
ፋ	fA	ኈ	hua
ፌ	fE	ኈ	hui
ፎ	fo	ኼ	hui
ፏ	fuA	ኋ	huA
ግe	e	ኋ	huA
ገa	a	ኻ	huA
ጉu	u	ኌ	huE
ጊi	i	ኼ	huE
ጋA	A	ኈ	hu
ጌE	E	ኼ	hu

ይ	ye	ሎ	lo
የ	ya	ሏ	luA
ዩ	yu	ም	me
ዪ	yi	መ	ma
ያ	yA	ሙ	mu
ዬ	yE	ሚ	mi
ዮ	yo	ማ	mA
ጅ	De	ሜ	mE
ጀ	Da	ሞ	mo
ጁ	Du	ሟ	muA
ጂ	Di	ን	ne
ጃ	DA	ነ	na
ጄ	DE	ኑ	nu
ጆ	Do	ኒ	ni
ጇ	DuA	ና	nA
ከ	ke	ኔ	nE
ከ	ka	ኖ	no
ኩ	ku	ኗ	nuA
ኪ	ki	ኘ	Ne
ካ	kA	ኘ	Na
ኬ	kE	ኙ	Nu
ኮ	ko	ኚ	Ni
ኰ	ko	ኛ	NA

ኩ	kui	ኜ	NE
ኳ	kuA	ኞ	No
ኴ	kuE	ኟ	NuA
ኰ	ku	ፐ	pe
ቅ	qe	ፐ	pa
ቀ	qa	ፑ	pu
ቁ	qu	ፒ	pi
ቂ	qi	ፓ	pA
ቃ	qA	ፔ	pE
ቄ	qE	ፖ	po
ቆ	qo	ፗ	puA
ቈ	qo	ጰ	Pe
ቍ	qui	ጳ	Pa
ቋ	quA	ጱ	Pu
ቌ	quE	ጲ	Pi
ቍ	qu	ጳ	PA
ል	le	ጴ	PE
ለ	la	ጶ	Po
ሉ	lu	ጷ	PuA
ሊ	li	ር	re
ላ	lA	ረ	ra
ሌ	lE	ሩ	ru
ሪ	ri	ተ	ta

ራ	rA	ቱ	tu
ሬ	rE	ቲ	ti
ሮ	ro	ታ	tA
ሯ	ruA	ቴ	tE
ሥ	se	ቶ	to
ስ	se	ቷ	tuA
ሠ	sa	ጥ	Te
ሰ	sa	ጠ	Ta
ሡ	su	ጡ	Tu
ሱ	su	ጢ	Ti
ሢ	si	ጣ	TA
ሲ	si	ጤ	TE
ሣ	sA	ጦ	To
ሳ	sA	ጧ	TuA
ሤ	sE	ጨ	ce
ሴ	sE	ቸ	ca
ሦ	so	ቹ	cu
ሶ	so	ቺ	ci
ሷ	suA	ቻ	cA
ሷ	suA	ቼ	cE
ጽ	xe	ቾ	co
ፅ	xe	ቿ	cuA
ጸ	xa	ኈ	C

ፀ	xa	ጨ	Ca
ጹ	xu	ጩ	Cu
ፁ	xu	ጪ	Ci
ጺ	xi	ጫ	CA
ፂ	xi	ጬ	CE
ጻ	xA	ጮ	Co
ፃ	xA	ጭ	CuA
ጼ	xE	ᦵ	ve
ፄ	xE	ᦶ	va
ጾ	xo	ᦷ	vu
ፆ	xo	ᦸ	vi
ጿ	xuA	ᦹ	vA
ሸ	Se	ᦺ	vE
ሻ	Sa	ᦻ	vo
ሹ	Su	ᦼ	vuA
ሺ	Si	ው	we
ሻ	SA	ወ	wa
ሼ	SE	ዉ	wu
ሽ	So	ዊ	wi
ሿ	SuA	ዋ	wA
ት	te	ዌ	wE
ዎ	wo		
ዝ	ze		

ዘ za

ዙ zu

ዚ zi

ዛ zA

ዜ zE

ዞ zo

ዟ zuA

ᘯ Ze

ᘰ Za

ᘱ Zu

ᘲ Zi

ᘳ ZA

ᘴ ZE

ᘵ Zo

ᘶ ZuA