ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

A GENERALIZED APPROACH TO AMHARIC TEXT-TO-SPEECH (TTS) SYNTHESIS SYSTEM

By

ALULA TAFERE ZEGEYE

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

July, 2010

A.A.U

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

A GENERALIZED APPROACH TO AMHARIC TEXT-TO-SPEECH (TTS) SYNTHESIS SYSTEM

By

ALULA TAFERE ZEGEYE

Signature of Board of Examiners for Approval

Million Meshesha (Ph.D) _____

_____ _____

_____ _____

_____ _____

DEDICATION

To Zekawaye and Baba

ACKNOWLEDGEMENT

First and foremost, I would like to express the deepest appreciation to my advisor, Dr. Million for his supervision, advice, and guidance from the very early stage of this research as well as giving me extraordinary experiences throughout the work. Above all and the most needed, he provided me unflinching encouragement and support in various ways. His science intuition has made him as a constant oasis of ideas and passions, which exceptionally inspire and enrich my growth as a student and a researcher want to be. I am indebted to him more than he knows.

I would like to thank my best friend Melkamu Beyene for his valuable assistance and helping me in my research. And I would also thank to my intimate friends, Alemayhu Tilahune, Daniel Beyene, Habtamu Getahune and Mesfin Olika for their assistance in my daily routine.

Many thanks to my colleagues Antenhe Alemu at Adama University, for his relentless cooperation and supporting he has made for me throughout all my studies at University.

Finally, where would I be without my family? My parents deserve special mention for their inseparable support and prayers. My Father, Tafere Zegeye, in the first place is the person who put the fundament my learning character, showing me the joy of intellectual pursuit ever since I was a child. My Mother, Wushen Embiale, is the one who sincerely raised me with her caring and gently love.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

CART	Classification and Regression Tree
DRT	Diagnostic Rhyme Test
DSP	Digital Signal Processing
LPC	Linear Predictive Coding
LMA	Log Magnitude Approximation
MOS	Mean Opinion Score
MRT	Modified Rhyme Test
NLP	Natural Language Processing
NSWs	Non-Standard Words
OCR	Optical Character Recognition
RELP	Residual Excited Linear Predictive
SWs	Standard Words
TD-PSOLA	Time Domain Pitch Synchronous Overlap and Add
TTS	Text to Speech
URL	Uniform Resource Locator
VODER	Voice Operating Demonstrator

ABSTRACT

A text-to-speech (TTS) synthesis converts natural language text into speech. However, written text of a language contains both standard words (SWs) and non-standard words (NSWs) like numbers, abbreviations, synonyms, currency, and dates. These NSWs cannot be detected by an application of "letter-to-sound" rule. This study describes generalized Amharic Text-To-Speech (TTS) synthesis, which attempt to handle both Amharic SWs and NSWs. The system is developed using speech synthesis framework of Festival, based on diphone unit concatenative synthesis by applying RELP coding technique. The model described in this work has two major parts: Natural language processing (NLP) and Digital language processing (DSP). The NLP handles the text analysis (transcription of the input SWs and NSWs) and extraction of the speech parameters. The DSP further enable to generate the artificial speech. Finally, the performance of the system shows that on the average 73.35% words both SWs and NSWs correctly pronounced. In addition, an assessment of intelligibility and naturalness of synthesized speech using MOS testing techniques results a score of 3 and 2.83, respectively. The experiment shows a promising result to design an applicable system that synthesis both SWs and NSWs for unrestricted text of a language. But, still there are areas need further investigations. Thoughtfulness of all type of NSWs and those ambiguities found in NSWs, while in test analysis block, using statistical technique to handle them based on their context. In addition, construction of part of speech POS tag-sets, tagger and tagged corpus for prosody analysis are also some areas that need further devotions.

Keywords: *Diphone concatenation, Speech Synthesis, (NSWs), RELP coding*

CHAPTER ONE

INTRODUCTION

1.1. Background

The development of society and economic system since prehistory time has been paralleled by a growth in man's dependence upon technology. The technologies are highly complex and highly valuable. Specifically, with the advancement in electronic and computer technology there is an explosive growth in the use of computers for processing information [2]. Since speech is one of the most effective means of communication for humans, spoken language is also a preferred method of human-machine interaction [4]. Speech-enabled interfaces are desirable because they promise hands-free, natural, and ubiquitous access to the interacting device [7]. As a result, spoken language system needs to have both speech synthesis and speech recognition capabilities [4].

As Dutoit [16] define, a Text-To-Speech (TTS) synthesizer is a computer-based system that should be able to read any text aloud, whether it was directly introduced in the computer by an operator or scanned and submitted by an Optical Character Recognition (OCR) system. On the other hand, speech recognition system has the capability to convert a given speech to text [4].

Text-to-Speech (TTS) synthesis is a process which artificially produces synthetic speech for various applications such as services over telephone, e-document reading, and speaking system for handicapped people etc [6].

Speech technology has advanced considerably over the last several years [10]. The realization of an idea that a machine could generate speech and such machines has only really been practical over the last 50 years. For instance, one of the first practical application of speech synthesis was in 1936 when the U.K. Telephone Company introduced a speaking clock [10]. The first computer-based speech synthesis systems were created in the late 1950s, and the first complete text-to-speech system was completed in 1968 [9].

The most important qualities of a speech synthesis system are *naturalness* and *Intelligibility* [2][9][13]. Naturalness describes how closely the output sounds like human speech, while intelligibility is the ease with which the output is understood. The ideal speech synthesizer is both natural and intelligible. Speech synthesis systems usually try to maximize both characteristics [9].

Converting text to speech encompasses both natural language processing and digital signal processing [16]. Natural language processing (NLP) is responsible to produce a phonetic transcription of the text read, together with the desired intonation and rhythm (often termed as *prosody*). Those transcriptions are done

under text analysis, phonetic analysis and prosody generation sequential blocks of NLP.

Text analysis block is responsible for analysis of raw of text into pronounceable word [9]. To achieve this, it organizes the input sentences into manageable lists of words and proposes all possible part of speech categories for each word taken individually on the basis of their spelling, and then considers words in their context. This can be achieved through using different techniques such as n-grams, Markov model, neural networks or classification and regression tree (CART) techniques. Finally, examines the remaining search space and finds the text structure (i.e. its organization into clause and phrase-like constituents) [16].

Phonetic analysis is responsible for the automatic determination of the phonetic transcription of the incoming text. Possibly this task can be organize in many ways, often roughly classified into *dictionary-based* and *rule-based* strategies [16]. once phonetic analysis done, the final block of NLP, prosody generation, which is responsible for finding correct intonation, stress, and duration from written text and these features together are called prosodic or suprasegmental features [9].

Digital signal processing (DSP) turns NLP representation into an output signal [3]. Intuitively, the operations involved in the DSP module are the computer

analogue of dynamically controlling the articulatory muscles and the vibratory frequency of the vocal folds so that the output signal matches the input requirements [16].

The broadest subdivision of strategies used to synthesize speech on computers is into system-model and signal-models [14]. While system-model attempts to model the human speech production system, signal-models which attempts to model the resulting speech signal. The system-model approach is known as articulatory synthesis [8][14]. The signal-model approach is perhaps the simpler of the two, and such has been both more thoroughly investigated, and more successful. It can be further subdivided into methods broadly described as rule-based format synthesis, and concatenation synthesis [14].

Concatenation synthesis operates by concatenating appropriate synthesis units to construct the required speech [14]. Concatenative synthesis produces the most natural sounding of synthesized speech. However, it requires a large amount of linguistic resources and generating various speaking style is a challenging task. In general, the amount of work required to build a concatenative system is enormous. Particularly, for languages with limited linguistic resources, it is more difficult to generate the required corpus with various speaking style and tone [6].

On the other hand, rule-based format synthesis systems were the most successful method of synthesizing speech for many years. Format synthesizers use an excitation signal to excite a digital filter constructed from a number of resonances similar to the formats of vocal trace [14]. Formant synthesis method requires small linguistic resources and able to generate various speaking styles. However, this method produce less natural-sounding synthesized speech and the complex rules required to model the prosody is a big problem [6].

There are many spellings in most languages which are pronounced differently based on context. For example in the English sentence, "My latest project is to learn how to better project my voice" contains two pronunciations of "project" [9][10][13]. Semantic representations of most text-to-speech (TTS) systems do not generate for their input texts, it may associate with as; processes for doing so are not reliable, well understood, or computationally effective.

1.2. Statement of Problem and Justification

Speech is one of the vital forms of communication for human being and it is the core activity in our day to day life. Currently, many speech synthesis systems are available mainly for ‘major’ languages such as English, Japanese etc. and successful results are obtained in various application areas. However, thousands of the world’s ‘minor’ languages lack such technologies, and researches in the

area are quite very few. Although recently many localization projects are being undergoing for many languages (such as, Bangle [11], Sinhala [12] and Raramuri [17]), it is quite inadequate and the localization process is not an easy task mainly because of the lack of linguistic resources and absence of similar works in the area. Therefore, there is a strong demand for the development of a speech synthesizer for many of the African minor languages including Amharic [6].

Amharic is the official language of Ethiopia, among 73 languages which are registered in the country [20]. Amharic is a Semitic language that has the greatest number of speakers next to Arabic [6][7][18]. Moreover, Amharic has 30 million speaker as a mother thong and as second language [24]. However, it is one of the least supported and least researched languages in the world [7].

So far some remarkable works have been done on speech synthesis for Amharic and other local languages. As to the knowledge of researcher, Henock [8], Habtame [1], Sebsibe with other abroad researchers [20], Nadew [7], Yibeltal [17] and Tadesse and Takara [6] design text-to-speech synthesis for Amharic language. In addition, Morka [19], Tesfaye [3] and Tewodros [2] attempts for Afaan Oromo, Tigrigna and Wolaytta language, respectively.

Henock [8] tries to develop Amharic language concatenative text to speech synthesis system. Two types of transcriptions (diphones and syllables) had been

implemented and with under consideration of intonation and duration changes. The finding was restricted only to a limited number of diphones and syllables. At last, based on the evaluation result the researcher conclude that, diphone based synthesis seems a better than syllable based synthesis.

Further, Nadew [7], attempts to design Amharic vowel synthesizer. The focus is on how to synthesize Amharic vowels from different contexts. Consonants and semivowels was not being the part of the synthesizer. A system is developed that generates vowels from a given context by best selection of parameters from the decision tree. It was the first attempt towards using formant based speech synthesis method for Amharic vowels. In addition, it enables him to produce vowels from different contexts without a need to store the speech units in the database; rather only speech parameters are stored to synthesize the new speech.

Tadesse and Takara [7], tries to develop Amharic Text-to-Speech system which is parametric and rule-based that employs a cepstral method. The system uses a source filter model for speech production and a Log Magnitude Approximation (LMA) filter as the vocal tract filtera. The system capable of synthesizing intelligible speech with fairly good prosody, syllables producing reasonably natural quality speech and durational modeling. However, the system still lacks

naturalness and needs automatic germination assignment mechanisms for better durational modeling.

The main focus of the above research works have been to design a TTS system for standard words. But close observation of the language shows that texts are full of non-standard terms, such as numbers, punctuations, abbreviations, currency, dates, etc. Non-standard words (NSWs) play significant role in every language text and speech meaning performance. As a result, TTS system also should handle NSWs by keeping in mind their influence in system performance, produced output signal representation of the give text and meaning of converted text. Assume that the given text contains '1992 ዓ.ም' and '1992' which have a difference in representation and meaning, and these input text should be analyzed in different ways in TTS system. Otherwise, discarding or producing as it create problem for listener and system performance.

However, the process of analyzing text is rarely straightforward. This is mainly due to those non-standard terms that all require expansion into a phonetic representation based on their context [10]. Deciding how to convert these non-standard terms are another problem that TTS systems have to address. Because we are not speaking, what we write on text. For instance, numbers occur in many different contexts; "1325" may also be read as "one three two five", "thirteen

twenty-five" or "thirteen hundred and twenty five". A TTS system can often infer how to expand a number based on surrounding words, numbers, and punctuation, and sometimes the system provides a way to specify the context if it is ambiguous. Similarly, abbreviations can be ambiguous [9][15][21].

In general, the first task faced by any text-to-speech (TTS) system is the conversion of input text into a linguistic representation. This is a complex task since the written form of any language is at best an imperfect representation of the corresponding spoken forms. Among the problems that one faces in handling ordinary text are the following [5]. First, digit sequences need to be expanded into words. Second, abbreviations must be expanded into full words. Third, ordinary words and names need to pronounce. Fourth, prosodic phrasing is only sporadically marked (by punctuation) in text, and lastly phrasal accentuation is almost never marked. So, those problems are also still in need of addressing in Amharic language.

As Tewodros [2], recommended extensive identification and inclusion of Non-standard words (NSWs) in the language with efficient representation methods is also another area to be addressed in future works. Because of real text contains many non-standard words. Use of non-standard words in any language TTS system is a critical issue to be solved.

It is therefore the aim of this research to explore the possibility of designing TTS synthesizer that can effectively handle unrestricted text contain both standard and non-standard words in Amharic language and generates natural sounding and intelligible speech which is vital for many application areas.

1.3. Objective of the Research

1.3.1. General objective

The general objective of the study is 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.

1.3.2. Specific Objectives

To accomplish the above stated general objectives, the following specific objectives are aimed to achieve:

- Review literatures on related works inwards to identify better ways of implementing Amharic TTS system.
- Identify mapping features of standard and non-standard words to speech/voice.
- Select efficient techniques for handling both standard and non-standard words (NSWs) in Amharic.

- Develop prototype Amharic TTS system that converts standard and non-standard words in to speech.
- Evaluate and report the performance of Amharic speech synthesis using test datasets.
- Forward concluding remarks and recommendation by identifying potential research areas as a further work.

1.4. Scope and Limitation of the study

This research prototype considers the TTS system for Amharic language. The system developed using diphone-based text-to-speech synthesis. In particular, with the extensive identification and inclusion of Amharic standard and non-standard words (NSWs). Moreover, the study attempt to select efficient representation method for integrating them with Amharic TTS synthesis system. The necessary models (language model, model algorithm, data model and model synthesizer) considered in this research. Representative data sets collected for evaluating the synthesis.

However, due to time constraint rule-based mapping process is done for limited number of non-standard word categories (such as cardinal numbers, date, years, time, ration, special characters and abbreviation) that are identified and considered. Besides unavailability of standardized corpus data, limits to handle

ambiguities found in Amharic non-standard words and also pronunciation of all Amharic words.

1.5. Area of Application

The problem and the reason that lead for the requirement of TTS is universal need for the technology. There are many applications of text to speech synthesizers which lead to the necessity of developing synthesizer for different language. Among those several applications some common once are mentioned below and described briefly as stated in [1][2][3][9][16]:

- *Aid for visually impaired:* Which is probably the most important and useful application field in speech synthesis is the reading and communication aids for the visually impaired. Instead of using special bliss symbol keyboard, which is an interface for reading the Braille characters, speech synthesis make their life easier to get information from computer with speech [9].
- *Aid for the hearing-impaired and vocally handicapped*: People who are born-deaf cannot learn to speak properly and people with hearing difficulties have usually speaking difficulties. Synthesized speech gives the deafened and vocally handicapped an opportunity to communicate with people who do not understand the sign language. With a talking head it is

possible to improve the quality of the communication situation even more because the visual information is the most important with the deaf and dumb.

- *Application for language education:* High Quality TTS synthesis can be coupled with a Computer Aided Learning system, and provide a helpful tool to learn a new language [16]. A computer with speech synthesizer can teach 24 hours a day and 365 days a year. It can be programmed for special tasks like spelling and pronunciation teaching for different languages [9].
- *Application for Telecommunication and multimedia:* The newest applications in speech synthesis are in the area of multimedia, which stimulating smart user friendly interfaces. TTS is applicable for language learning, entertainment, talking characters, proof reading, and productivity tools, online talking assistance. Synthesized speech has been used for decades in all kind of telephone enquiry systems and may also be used to speak out short text messages (SMS) in mobile phones [9][16].
- *Speaking alarm system, announcement system:* In principle, speech synthesis may be used in all kind of human-machine interactions. Synthesized speech applicable to give more accurate information of the current situation [9].

1.6. Research Methodology

A research methodology defines what the activity of research is, how to proceed, how to measure progress, and what constitutes success.

1.6.1. Review of related Literature

Before starting the actual work, literature review is made that are written on TTS system area to have a clear picture about the work. Books, journal articles, conference, papers and Internet are the main sources. Literatures written on Amharic language are reviewed so as to understand the nature of the language and how to design new system. Specifically, different research publications and the Internet are reviewed for extensive identification of non-standard words in Amharic language. The different speech synthesis techniques are also studied to identify their differences and select suitable one for the research.

1.6.2. Prototype design and development tool

In view of the fact that connecting prerecorded natural utterances is the easiest way to produce intelligible and natural sounding synthetic speech [9]. This work develops a prototype Amharic language Text-to-Speech (TTS) system that converts both standard and non-standard words (NSWs) in to speech. The

system is implemented using the Festival speech development toolkit framework.

The Festival Speech Synthesis System is an open-source and a complete TTS synthesis system, with components supporting front-end processing of the input text, language modeling, and speech synthesis using its signal processing module [11][12]. In addition, Festival is also a concatenative TTS system using diphone or other unit selection speech.

The possible sets of diphone database are recorded and manually collected using Praat[1]. Praat is freely available speech analysis tool and it incorporates sound recording, spectral analysis, sound segmentation, pitch analysis, and segment concatenation, etc functionalities. In line with the utility, and also because it is easy to use, Praat is preferred as speech analysis tool in this work.

Construction of a diphone database and implementation of the natural language processing modules are done. Residual Excited Linear Predictive Coding (RELP Coding), which is method of concatenation currently distribute with Festival framework is used as the method for synthesis. In addition, as required by this method, pitch marks, Linear Predictive Coding (LPC) parameters and LPC residual values are extracted for each diphone in the diphone database.

[1] *Available from: http://www.praat.org.*

1.6.3. Testing Procedure

The performance of the Amharic synthesizer is evaluated using representative test datasets. The precision of the system is examined for the given standard and non-standard words. In addition, the overall quality (that is, intelligibility and naturalness) of synthetic speech created by the system is also measured by the user.

On the other hand, to test the performance of TTS system a number of testing techniques are suggested in literatures [2][8][9], Modified Rhyme Test (MRT), and Diagnostic Rhyme Test (DRT), Mean Opinion Score (MOS), Pair Comparison (PC) and Semantically Unpredictable Sentences (SUS) are some of the performance testing methods. In this thesis work, Mean Opinion Score (MOS) is used, which is the most widely and simplest method to evaluate speech quality. It is also suitable for overall evaluation (intelligibility and naturalness) of synthetic speech. MOS is a five[2] level scale from bad (1) to excellent (5) and it is also known as ACR (Absolute Category Rating) [9]. Respectively, for overall evaluation five native speaker of Amharic language are invited. And, then the evaluators give their option based on MOS scale for both intelligibility and naturalness criteria of synthesized speech.

[2] *See table 5.2 for detail value of each MOS scale.*

1.6.4. Organization of the Thesis

This thesis is organized into five chapters. The first chapter presents an introduction to the subject matter under study, description of the statement of the problem along with justifications and objectives of the research. It also states an overview on speech synthesis, methodologies adopted on the research, applications of speech synthesis systems and recipient of the new system.

Chapter two provides literature review on human speech production system, overview of speech synthesis. It also presents different techniques of speech synthesis used by researchers. Works done to implement a TTS system for Amharic are also discussed to show newness of present research.

Chapter three discusses the Amharic language phonologies in general. Consonant and vowel sounds with their characteristics are also described. At the last identified Amharic non-standard word (NSWs) are presented.

In chapter four, algorithms used in this study are explained. Some of the techniques are Residual Exited Linear Predictive (RELP) Coding algorithm.

Chapter five present the prototype designed for Amharic language. And also discusses experimentation and testing of the Amharic speech synthesizer in

detail. The last chapter provides concluding remarks and recommendations for future research in the area.

CHAPTER TWO

REVIEW OF LITERATURE

Speech is the primary means of communication between people. The human speech production system is composed of organs ranging from the diaphragm and lung to the vocal and nasal cavity [8]. Text-to-speech synthesis system, on the other hand takes the input text and gives speech utterance.

2.1. Historical Background

The historical root of speech and speech science goes many years back, since it is a vital means of communication. As Henock [8] noted, speech science could be believed to start about 2,500 years ago. However, speech synthesis mainly has under gone through two stage of development: mechanical and electronic stage of development, ever since late 18th century [1][9].

Under the mechanical stage of development, efforts were exerted to produce synthetic speech over two hundred years ago. In 1779, the Danish scientist Christian Kratzenstein, working at the Russian Academy of Sciences, built the first talking machine [8]. He constructed models of the human vocal tract that could produce the five long vowel sounds ([aː], [eː], [iː], [oː] and [uː]) and

activated the resonators with vibrating reeds like music instrument [1][9][13]. The machine was composed of tube like acoustic resonator – each with its own shape – and each of which produced specific vowel.

A few years later, the first attempt to synthesize speech was built by Wolfgang von Kempelen of Vienna, Austria in 1791 [14]. His talking machine was powered by bellows and this bellows-operated synthetic speech machine called "acoustic-mechanical speech machine". Kempelen's machine added models of the tongue and lips, enabling it to produce consonants as well as vowels. As Lemmetty [9] noted, the essential parts of the machine were a pressure chamber for the lungs, a vibrating reed to act as vocal cords, and a leather tube for the vocal tract action. By manipulating the shape of the leather tube he could produce different vowel sounds. Consonants were simulated by four separate constricted passages and controlled by the fingers that had been used to produce consonant sounds [1][8][9].

Tewodros [2], quoting Holmes and Holmes (2003), discusses that in about mid 1800's Charles Wheatstone constructed his famous version of von Kempelen's speaking machine. It was a bit more complicated and was capable to produce vowels and most of the consonant sounds [9].

In late 1800's [8][9] Alexander Graham Bell with his father, inspired by Wheatstone's speaking machine, constructed same kind of speaking machine. Bell made also some questionable experiments with his terrier. He put his dog between his legs and made it growl, then he modified vocal tract by hands to produce speech-like sounds [9].

The second stage of development, electronic stage, and the first full electrical synthesis device was introduced by Stewart in 1922 [1]. The synthesizer had a buzzer as excitation and two resonant circuits to model the acoustic resonances of the vocal tract. The machine was able to generate single static vowel sounds with two lowest formants, but not any consonants or connected utterances [9].

First device to be considered as a speech synthesizer was VODER (Voice Operating Demonstrator) introduced by Homer Dudley in New York World's Fair 1939 [14]. The VODER was inspired by VOCODER (Voice Coder) developed at Bell Laboratories in the mid-thirties. The original VOCODER was a device for analyzing speech into slowly varying acoustic parameters that could then drive a synthesizer to reconstruct the approximation of the original speech signal. The VODER consisted of wrist bar for selecting a voicing or noise source and a foot pedal to control the fundamental frequency. The source signal was routed through ten band pass filters whose output levels were controlled by fingers. It

took considerable skill to play a sentence on the device. As a result, the VODER need an operator to manipulate and one problem with it was not easy to manipulate and it needed a trained operator [8][9]. The following figure 2.1, show the architecture of VODER speech synthesizer [9].

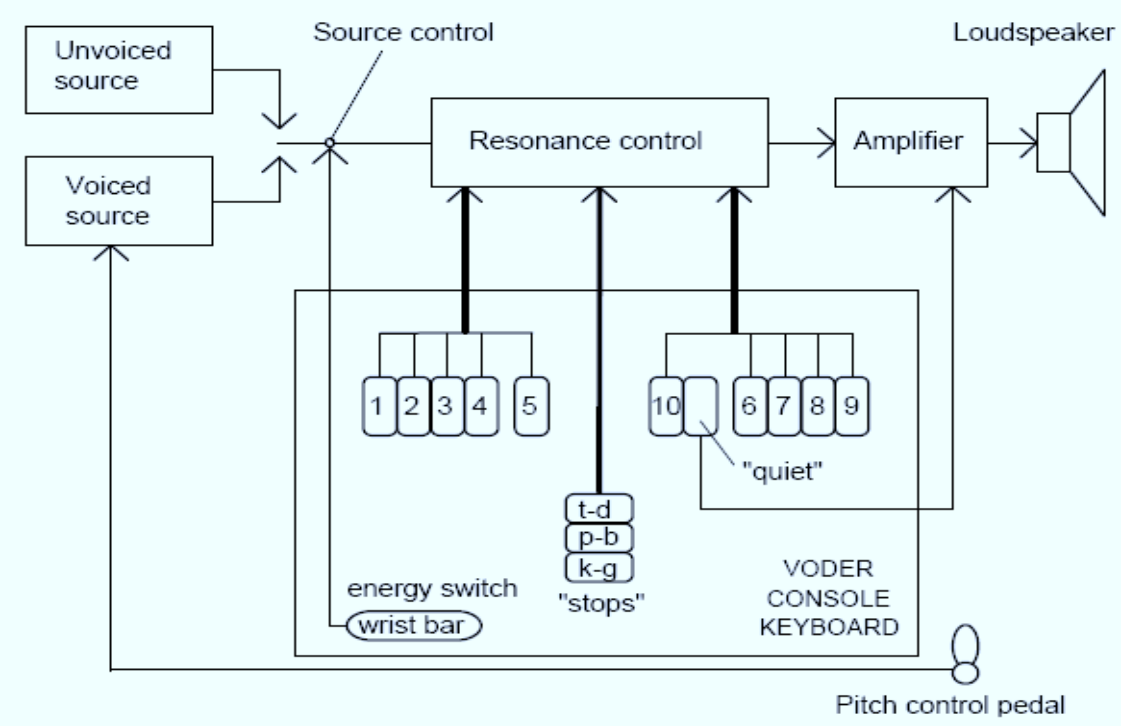

Figure 2.1: The VODER speech synthesizer.

After demonstration of VODER the scientific world became more and more interested in speech synthesis. Later on, a number of developments were achieved. The PAT (Parametric Artificial Talker), which is the first formant synthesizer and the first articulatory synthesizer, was introduced by Walter Lawrence in 1953 and by George Rosen at the Massachusetts Institute of

Technology (M.I.T) in 1958, respectively [9]. The first full text-to-speech system for English was developed in the Electrotehnical Laboratory, Japan 1968 by Noriko Umeda and his companions [9].

Each attempts in past was put their own mark and play a vital role for the current available systems. The present systems are adequate enough some for specific application areas. Even though the achievement is good, the existences of single challenge on the most important qualities of speech synthesis system (i.e. naturalness and intelligibility) become impossible to conclude generally high.

2.2. Human Speech Production System

In human speech production system, sound is generated in several ways and at several locations in the vocal trace. It is composed of organs ranging from the diaphragm and the lungs to the vocal and nasal cavities. The simplified version of speech organs are presented in figure 2.2 [14]. In between there are a number of organs like the trachea (windpipe), larynx, pharyngeal cavity (throat), oral cavity (mouse), and nasal cavity (nose). The pharyngeal and oral cavities are typically referred to as the vocal tract, and the nasal cavity as the nasal tract [3][8].

Lung and diaphragm are the main energy source of the human speech production system. When speaking, the air flow is forced through the glottis between the vocal cords and the larynx to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities [9]. This air flow becomes the cause of vocal cords, in one side, to constrict and vibrate, and on the other side, to relax and let the air to pass effortlessly. As a result, voiced and unvoiced sounds are produced respectively. As Nadew [7] discuses, if the speech sound made the vocal folds (vocal cords) close together and oscillates against one another, the sound is said to be *voiced*. When the folds are too slack or tense to vibrate periodically, the sound is said to be *unvoiced*.

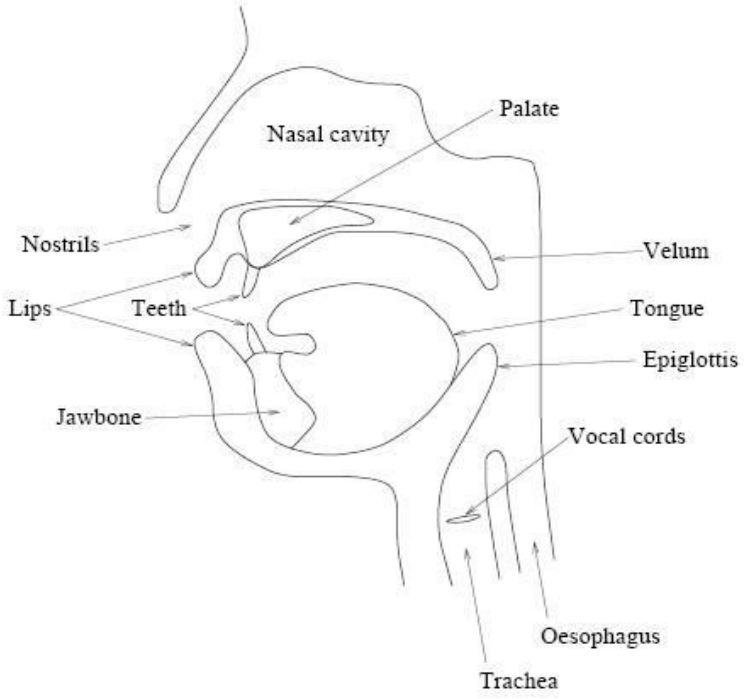

Figure 2.2: Human speech production system.

The larynx is the structure that holds and manipulates the vocal cords. The place where the vocal cords come together is called the glottis [7]. It modulates the air flow by rapidly opening and closing, which results to produce different buzzing sounds and each with different frequency [8]. Moreover, depending on the positions of the various articulators; namely jaw, tongue, velum, lips, and mouth can produce different sounds [3].

As Tesfaye [3], quoting Eker (2002) discusses that, the vocal trace is bound by hard and soft tissue structure. The structure are either essentially immobile or are movable. The movable structures associated with speech production are also referred to articulators (i.e. jaw, tongue, velum, lips, and mouth). Movement of these articulators appeared to account for most of the variation in vocal trace shape associated with speaking. Despite the above, the air passing through the nasal cavity will not encounter any further obstacles.

The tongue plays critical role in speech production via moving around mouth. The dorsum is the main part of the tongue, lying below the hard and soft palate. It shaped away from the palate for vowels, placed close to or on the palate or other hard surfaces for consonant articulation. The teeth and lip are also another place of articulation used to, brace the tongue for certain consonants and move against the teeth as necessary, respectively [7][9].

2.3. Fundamentals of Speech Synthesis

The communication between people mainly conducted through speech, which is a giant and most common way of communication. However, the speech communication happens only between human. As a result, technological struggles have been passing through many challenges and improvements as briefly described above to create and have speech communication between human and machine. Moreover, producing intelligible and natural synthetic speech is requirement in the course of developing a process of Text-to-Speech (TTS) synthesis system.

TTS is defined as the production of speech by machines, by way of the automatic phonetization of the sentences to utter [16]. Speech synthesis is a process of building the system that can generate human-like speech from any text input to mimic human speakers [7]. The ultimate objective of Text-to-Speech (TTS) synthesis systems is to create applications which listeners, and users in general, cannot easily determine whether the speech he or she is hearing comes from a human or a synthesizer [17]. This could possibly assert that, the ideal speech synthesizer should possess both high intelligibility and high naturalness of synthesized speech, to achieve its crucial objective.

Those two parameters - naturalness and intelligibility of speech- are applied to the description of quality of speech synthesis' system. The quality of a speech synthesizer is judged by its similarity to the human voice, and by its ability to be understood. In need of addressing those, the text-to-speech (TTS) synthesis procedure contains a number of steps and synthesized speech can be produced by different methods[3] [9].

Natural Language Processing (NLP) and Digital Signal Processing (DSP) are generally two main phases of TTS systems in the process of converting written text into speech [16]. The former one is targeted to produce phonetic transcription of the text, together with the desired intonation and rhythm. This phase is also known as high-level synthesis [9]. The later one is transforms the symbolic information it receives from the former phase into speech [16]. This phase is also known as low-level synthesis [9]. Any TTS system contain the above two phases as show in figure 2.3 [16].

[3] See section 2.6 for detail discussion of methods.

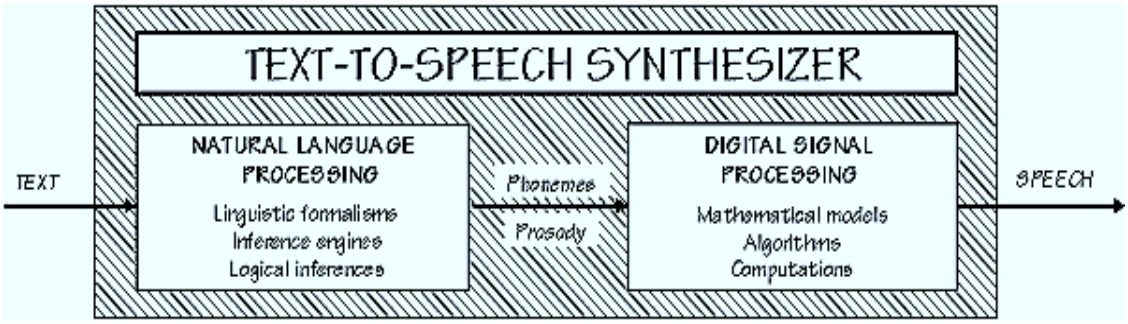

Figure 2.3: TTS system architecture

2.4. Natural Language Processing (NLP)

The aim of Natural language processing (NLP) is to produce phonetic transcription of the text, together with the desired intonation and rhythm. In any written text of a language the presence of symbols like numbers, abbreviations and acronyms are very common. Since the ways we write and speak are different, there is a need to design a mechanism that enable symbols converted to their equivalent speech forms. In addition, phonetic transcription of the text with correct prosody and pronunciation also needed and addressed in this phase even if it is difficult because of written text does not contain explicit emotions. In essence, NLP phase consist of three processing stages: Text Analysis, Phonetization and Prosody generation [9][16].

2.4.1. Text Analysis

Text analysis is one the most influential and complex task of NLP module. This step is responsible for analysis of raw text into pronounceable words [9]. In order to achieve these responsibility four main modules [16]: pre-processing, morphological analysis, contextual analysis and syntactic-prosodic parser module are needed. Initially pre-processing module [16], organizes the input sentences into manageable lists of words. It identifies numbers, abbreviations, and acronyms and transforms them into full text.

Second, morphological analysis module, propose all possible part of speech categories for each word taken individually, on the basis of their spelling. Inflected, derived, and compound words are decomposed into their elementary graphemic units (their *morphs*) by simple regular grammars exploiting lexicons of stems and affixes [16].

The contextual analysis module considers words in their context, which allows it to reduce the list of their possible part of speech categories to a very restricted number of highly probable hypotheses, given the corresponding possible parts of speech of neighbouring words. This can be achieved either with *n-grams*, which describe local syntactic dependences in the form of probabilistic finite state automata (i.e. as a Markov model), to a lesser extent with *mutli-layer perceptrons*

(i.e., neural networks) trained to uncover contextual rewrite rules, or with *local, non-stochastic grammars* provided by expert linguists or automatically inferred from a training data set with *classification and regression tree* (CART) techniques. At the last, a syntactic-prosodic parser examines the remaining search space and finds the text structure [16].

However, converting raw text that contains non-standard words (NSW) like numbers, abbreviations and acronyms into the equivalent of written-out words is challengeable. Moreover, finding the correct full written word with the given context of the sentence made the process more difficult because missing the give parameter representation of non-standard words (NSW) will lead confusion on listeners while the system speak.

As Lemmety [9] describes, digits and numerals must be expanded into full words. For example in English, numeral 243 would be expanded as *two hundred and forty-three* and 1750 as *seventeen-fifty* (if year) or *one-thousand seven-hundred and fifty* (if measure). Related cases include the distinction between *the 747 pilot* and *747 people*. Fractions and dates are also problematic. 5/16 can be expanded as *five-sixteenths* (if fraction) or *May sixteenth* (if date). Expansion ordinal numbers have been found also problematic. The first three ordinals must be expanded differently than the others, 1st as *first*, 2nd as *second*, and 3rd as *third*. Same kind

of contextual problems are faced with roman numerals. Chapter III should be expanded as *Chapter three* and Henry III as *Henry the third* and *I* may be either a pronoun or number [9].

In written Amharic language also, the existence of non-standard words (NSW) integrated with standard words are common. For instance, number 1889 can be considered and expand as like the normal measurement number "አንድ ሺ ስምንት መቶ ሰማንያ ዘጠኝ" and the year "አስራ ስምንት ሰማንያ ዘጠኝ". As a result, in Amharic language TTS system should consider those variations for system effectiveness.

Abbreviations may be expanded into full words, pronounced as written or pronounced letter by letter [9]. There are also some contextual problems. For example kg can be either *kilogram* or *kilograms* depending on preceding number, St. can be *saint* or *street*, Dr. *doctor* or *drive* and ft. *Fort*, *foot* or *feet* [10]. The same issues are also arising in Amharic even if the extent is different. "ወ/ሮ", "ዓ.ም", "ቅ.ገብርኤል", "አ.አ", "የተ.መ.ድ" are some examples of Amharic abbreviations that in need of expanding.

Special characters and symbols, such as '$', '%', '&', '/', '-', '+', cause also special kind of problems. In some situations the word order must be changed. For example, *$71.50* must be expanded as *seventy-one dollars and fifty cents* and *$100*

million as *one hundred million dollars*, not as *one hundred dollars million* [9]. Here in Amharic to some extent also face the same problem, in written word like 50% the symbol must be expand in to full Amharic word for the speech consistency and effectiveness.

2.4.2. Phonetic Analysis

The second task in NLP is to find correct pronunciation for different contexts in the text after normalized word strings are taken from text analysis. In addition, as Nadew [7] describe this stage is to produce a pronunciation for each word along with possible diacritic information. However, some words, called *homographs*, cause maybe the most difficult problems in TTS systems [9]. Phonetic analysis is thus often referred to grapheme-to-phoneme conversion.

The task of the phonetic analysis module can be organized in many ways, often roughly classified into *dictionary-based* and *rule-based* strategies [16]. *Dictionary-based* solutions consist of storing a maximum of phonological knowledge into a lexicon. In order to keep its size reasonably small, entries are generally restricted to morphemes, and the pronunciation of surface forms is accounted for by inflectional, derivational, and compounding morphophonemic rules which describe how the phonetic transcriptions of their morphemic constituents are modified when they are combined into words. After a first phonemic

transcription of each word has been obtained, some phonetic post-processing is generally applied, so as to account for coarticulatory smoothing phenomena [16].

On the other hand, rule-*based* transcription systems transfer most of the phonological competence of dictionaries into a set of letter-to-sound (or *grapheme-to-phoneme*) rules [16]. This time, only those words that are pronounced in such a particular way that they constitute a rule on their own are stored in an exceptions dictionary. Since many exceptions are found in the most frequent words, a reasonably small exceptions dictionary can account for a large fraction of the words in a running text. In English, for instance, 2000 words typically suffice to cover 70% of the words in text [16].

Amharic is nearly phonetic language, meaning that a simple grapheme-to-phoneme conversion is possible for most of the words due to close relationship with Amharic orthography and phonology [7]. Nevertheless, in Amharic, depending on the context, the same word may be pronounced differently because of homograph influence even though the pronunciation does not change completely.

2.4.3. Prosody

The final stage of NLP is prosody generation. The term *prosody* refers to certain properties of the speech signal which are related to audible changes in pitch, loudness, and syllable length [16]. As Lemmetty [9] describe, this stage focus on finding correct intonation, stress, and duration from written text and these features together are called prosodic or suprasegmental features and may be considered as the melody, rhythm, and emphasis of the speech at the perceptual level. The intonation tries to estimate how the pitch pattern or fundamental frequency changes during speech. Here is a fact that, in usual situation while everybody speaks there are variations of pitches, but usually in written text these features does not explicitly included and/or contain very little information. More clearly as Henock [8], quoting Rodman (1999) discusses, in prosodic property may also convey emotional or other connotation in addition to changing the meaning of a sentence or word.

Specifically, in Amharic prosodic features have crucial role in the meaning of the sentence. For example, the Amharic sentence "አንተ አትመጣም" may possibly elaborate the case. A sentence spoken by varying the level of intonations gives two meanings; one, it becomes command and the other is as question. In addition, the sentence may also express emotional connotations such as anger

and/or sadness. As a result, suprasegmental features can even change the meaning of the sentence.

Prosodic features create segmentation of the speech chain into groups of syllables. This gives rise to the grouping of syllables and words in larger chunks [1]. Prosody can also be understood at different levels. At linguistic level, we know the tone, intonation or stress of a speech. That is what prosody exhibits in linguistic level. Prosody is also perceived by human as pitch, loudness, length and strength. This is what we get at perceptual level. Prosody, if expressed at acoustic level, is actually fundamental frequency, duration, amplitude etc, and that is what we operate in the synthesis process. The proper combination of these acoustic factors makes speech natural, expressive and active. Generally, figure 2.4 show some basic factors contributing to prosodic features [9].

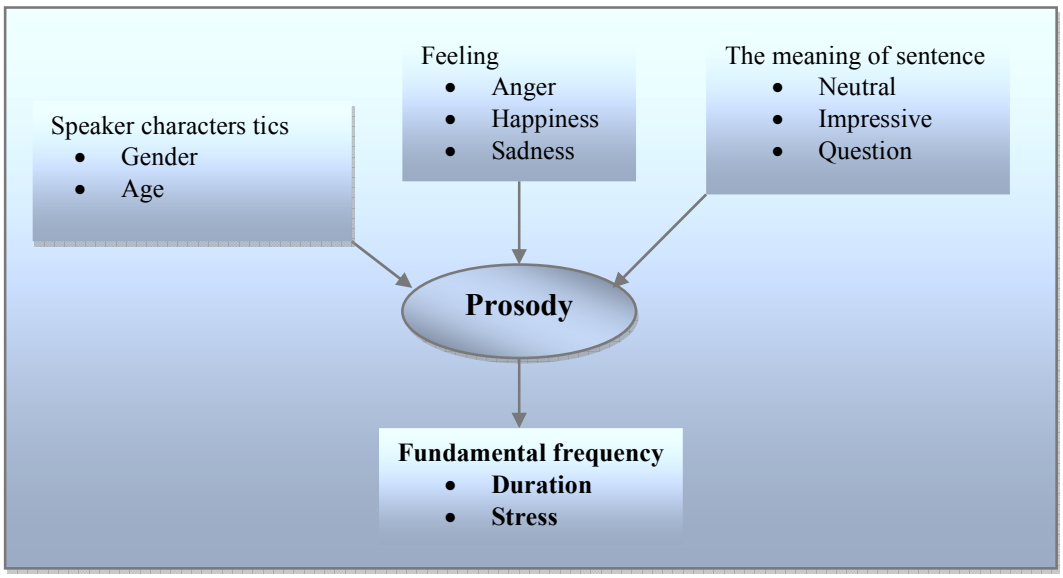

Figure 2.4: Factors contributing to prosodic feature.

2.5. Digital Signal Processing (DSP)

Digital signal processing (DSP) is the second module of TTS system. It transforms the symbolic information (such as, phonetic transcription and prosodic information) it receives from the NLP phase into speech [9][16]. The input of traditional speech synthesis systems is a phonetic transcription with its associated prosody. The input can also include the original text with tags; this may help in producing higher-quality speech [7].

There are many methods to produce speech sounds after text and prosodic analysis. Usually the methods are classified in to three groups: articulatory, format and concatinative methods [16]. Articulatory synthesis attempts to model

the human speech production system directly. Formant synthesis, which models the pole frequencies of speech signal transfers function of vocal tract into based on source-filter-model. Concatenative synthesis uses different length of prerecorded samples derived from natural speech [14].

2.6. Speech Synthesis Techniques

As Donovan [14] discuss, speech synthesis techniques can be broadly divided into two categories: system model approach and signal model approach. System model approach tries to model directly the human vocal system and this technique is also called articulatory synthesis. On the other side, signal model approach attempts to focus model speech signal only and with two varieties namely; rule-based formant synthesis and concatenative synthesis.

In line with the above, speech synthesis can also be classified according to the degree of manual intervention in the system design as *synthesis by rule* and *data-driven synthesis* [8]. In the former, a set of rules are used to drive a synthesizer, and in the latter, the synthesizer's parameters are obtained from real speech data. While concatenative system is, thus, data driven, formant synthesis is rule-based. Currently, there is a new approach named as data-driven-formant synthesis, which uses the combination of the two, to take the advantages of each and to get the best out of each [7].

2.6.1. Articulatory Synthesis

Articulatory synthesis refers to computational techniques for synthesizing speech based on models of the human vocal tract and the articulation processes occurring there [14]. Articulatory synthesis tries to model the human vocal organs as perfectly as possible. So it is potentially the most satisfying method to produce high-quality synthetic speech [9]. Since it models the human system directly, it is also one of the most difficult methods to implement and the computational load is also considerably higher than with other common methods. Thus, it has received less attention than other synthesis methods and has not yet achieved the same level of success [14].

As discussed by Henock [8], the articulatory synthesis models the articulators and vocal cords. The articulators are modeled with a set of area functions between the glottis and the mouth. Attributes like lip aperture, lip protrusion, tongue tip position, and velic aperture serve as articulatory control parameters. The first articulatory model was based on a table of vocal tract area functions from larynx to lips for each phonetic segment.

The rule-based articulatory synthesis is very difficult to optimize [9]. This is due to the fact that, while speaking articulators move and change shape of the vocal tract which caused by the vocal tract muscles and become the basis of different

sounds. X-ray analysis of natural speech derived data is usually used for articulatory model. However, this data is usually only 2-D when the real vocal tract is naturally 3-D, so the unavailability of sufficient data of the motions of the articulators during speech triggers it to face difficulty. Other deficiency with articulatory synthesis is that X-ray data do not characterize the masses or degrees of freedom of the articulators. Also, the movements of tongue are so complicated that it is almost impossible to model them precisely [9].

2.6.2. Formant Synthesis

Formant synthesis is the other method used to produce synthesized speech. Probably the most widely used synthesis method during last decades has been formant synthesis which is based on the source-filter-model of speech [9]. It does not use human speech samples at runtime. Instead, a set of rules is used to determine the parameters necessary to synthesis a desired utterance using a formant synthesizer. Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. This method is sometimes called *rules-based synthesis*; however, many concatenative systems also have rules-based components [8][9][14][16].

Formant synthesis models the speech spectrum and its changes in time as we speak, rather than model the production mechanisms themselves [18].

Parameters such as pitch, voicing, and noise levels are varied over time to synthesize speech waveforms [7][9][14][16]. This method is also called parametric based synthesis. Formant synthesis uses a relatively simple system to select from a small number of parameters, with which to control a mathematical model of the speech sounds. A set of parameters is picked for each speech sound and they are then joined up to make the speech. This stream of parameters is so turned into synthetic speech using the model [18].

In formant synthesis there are two basic structures [7][9][14][18]: parallel and cascade. But for better performance some kind of combination of these is usually used. A cascade formant synthesizer consists of band-pass resonators connected in series and the output of each formant resonator is applied to the input of the following one. The cascade structure needs only formant frequencies as control information. The main advantage of the cascade structure is that the relative formant amplitudes for vowels do not need individual controls [9].

A parallel formant synthesizer consists of resonators connected in parallel. Sometimes extra resonators for nasals are used. The excitation signal is applied to all formants simultaneously and their outputs are summed. Adjacent outputs of formant resonators must be summed in opposite phase to avoid unwanted zeros or antiresonances in the frequency response. The parallel structure enables

controlling of bandwidth and gain for each formant individually and thus needs also more control information [9].

The strength of formant synthesis is its relative simplicity and the small memory footprint needed for the engine and its voice data [18]. It also provides infinite number of sounds which makes it more flexible than for example concatenation methods [9]. This can be important for embedded and mobile computing applications. Formant synthesis generates highly intelligible, but not completely natural sounding speech.

2.6.3. Concatenative Synthesis

Concatenative synthesis is another method that is recently gaining much attention. It operates based on the concatenation of segments of recorded speech [14]. Connecting prerecorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech [9]. In concatenative synthesis, signal processing techniques should be applied to alter parameters like pitch and duration, and to smooth the discontinuity created by concatination points [8].

This method often criticized due to, the differences between natural variations in speech and the nature of the automated techniques for segmenting the

waveforms sometimes result in audible glitches in the output. In addition, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods. But despite its weak side, concatenative synthesis produces the most natural-sounding synthesized speech [9].

One of the most important aspects in concatenative synthesis is to find correct unit length. The selection is usually a trade-off between longer and shorter units [16]. With longer units high naturalness, less concatenation points and good control of coarticulation are achieved, but the amount of required units and memory is increased. With shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. In present systems units used are usually words, syllables, demisyllables, phonemes, diphones, and sometimes even triphones [3][9][16].

Word is perhaps the most natural unit for written text and some messaging systems with very limited vocabulary. Concatenation of words is relatively easy to perform and coarticulation effects within a word are captured in the stored units [9]. The great difference between words which spoken in isolation and in continuous sentence, makes the continuous speech to sound very unnatural. If the words are recorded separately, intonation will be lost. Moreover, the system

will be limited with the prerecorded words and this makes the word concatenation unsuitable unit for unrestricted TTS systems [3].

The number of different syllables in each language is considerably smaller than the number of words, but the size of unit database is usually still too large for TTS systems. For example, there are about 10,000 syllables in English [16]. Unlike words, the coarticulation effect is not included in stored units. So, using syllables as a basic unit is not very reasonable. There is also no way to control prosodic contours over the sentence. At the moment, no word or syllable based full TTS system exists. The current synthesis systems are mostly based on using phonemes, diphones, demisyllables or some kind of combinations of these. On the other hand, demisyllables represents the initial and final parts of syllables [9][16].

Phonemes are probably the most commonly used units in speech synthesis because they are the normal linguistic presentation of speech. Using phonemes give maximum flexibility with the rule-based systems. However, some phones (such as, plosives) that do not have a steady-state target position are difficult to synthesize. The articulation must also be formulated as rules. Phonemes are sometimes used as an input for speech synthesizer to drive for example diphone based synthesizer.

Diphones are defined as a stretch from the least varying (most stable or steady-state) part of a phone to a similar point in the next phone [16]. The idea of introducing diphones was to capture the transition between phones within the acoustic model in order to reduce mismatches between phones [17]. That means that the concatenation point will be in the most steady state region of the signal, which reduces the distortion from concatenation points [9].

One advantage of diphones is that, since the boundaries between diphones during synthesis thus occur at the middle of phones, the coarticulation effects need no more to be formulated as rules [14]. In principle, the number of diphones is the square of the number of phonemes (plus allophones), but not all combinations of phonemes are needed [9]. Nevertheless, the number of data is still tolerable and with other advantages, diphone is a very suitable unit for sample-based text-to-speech synthesis. The number of diphones may be reduced by inverting symmetric transitions, like for example /as/ from /sa/. Another reason to use diphones is, despite the price of storing a large number of permutations of units, this choice is much more convenient than using syllables because they would need a considerably larger number of permutations [17].

In general, as Lemmetty [9] describes, there are several problems in concatenative synthesis compared to other methods.

- Distortion from discontinuities in concatenation points, which can be reduced using diphones or some special methods for smoothing signal.
- Memory requirements are usually very high, especially when long concatenation units are used, such as syllables or words.
- Data collecting and labeling of speech samples is usually time-consuming. In theory, all possible allophones should be included in the material, but trade-offs between the quality and the number of samples must be made.

2.7. Speech Synthesis Engines

The development of speech science were introduce many speech synthesizer system tools, such as, Osprey, ATR's CHATR system, Festival, Festvox, AT&T Bell labs TTS system and java-based speech synthesis system from Sun Microsystems called FreeTTS. The two most widely used tools to develop a speech synthesizer system currently are Festival and Festvox [2][10]. Specifically, these tools are recently made possible for many localization projects to being undergoing in support of different languages[4], mainly through customization of Festvox [6].

[4] *such as see reference, [11], [12] ,[17] and [26].*

2.7.1. Festival TTS System

The Festival TTS system was developed in CSTR at the University of Edinburgh by Alan Black and Paul Taylor and in co-operation with CHATR, Japan [10][26]. The current system is available for American and British English, Spanish, Welsh and recently Festival also speaks Italian language. The system is written in C++ and supports residual excited LPC and PSOLA methods and MBROLA database. With LPC method, the residuals and LPC coefficients are used as control parameters [10]. With PSOLA or MBROLA the input may be for example standard PCM files. As a University program the system is available free for educational, research, and individual use.

The system is developed for three different users aspect [10][26]. The first group is those who want simply use the system from arbitrary text-to-speech. The other group is people who are developing language systems and wish to include synthesis output, such as different voices, specific phrasing, dialog types and so on. Lastly, for those who are developing and testing new synthesis methods.

The Festival Speech Synthesis System is an open-source and a complete TTS synthesis system, with components supporting front-end processing of the input text, language modeling, and speech synthesis using its signal processing

module [11][12]. In addition, Festival is also a concatenative TTS system using diphone or other unit selection speech.

Festvox is also a tool which facilitate for creating voices in different languages with appropriate documentation that use in Festival Speech Synthesis system. The Festvox package is developed in 2003 by Alan Black and Kevin Lenzo at the Language Technologies Institute at Carneige Mellon University, Pittsburgh [10]. The purpose with this project was to build synthesized voices in a more systematic way and to make the documentation process better. The goal with this package is to make it easier for anyone with little knowledge of the subject to be able to build an own synthetic voice. Basic skeleton files are included in the Festvox distribution that will facilitate the building of a new voice [2][10].

2.8. Related Researches in Local languages

As to the knowledge of the researcher, so far some remarkable works have been done on speech synthesis for local languages. Among those works done using concatinative synthesis are discussed below.

2.8.1. TTS for Afaan Oromo

Morka [19] develops a prototype TTS for Afaan Oromo. The technique that he uses in his research was concatenative speech synthesis and diphones were used

as the basic concatenation units to synthesize sample Afaan Oromo texts. It was indicated in the research paper success on recognizing the utterance of the transcribed phonetic unit was 43.33% for native speakers. He recommended incorporating spectral smoothening technique to smooth the transition points of the diphones.

2.8.2. TTS for Tigrigna

Tesfay [3] develops a prototype for Tigrigna Language TTS System. He deals with diphones as the basic concatenation units to synthesize sample Tigrigna texts and also Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) used as technique to generate the synthetic speech. He adopted Mean Opinion Score testing method and evaluation was done for both Tigrigna and Amharic language. The average result obtained was 3.05, which is closer to the scale level good. He finally recommended the need incorporate number converter for word normalization.

2.8.3. TTS for Wolaytta

Tewodros [2] develop a prototype of Wolaytta language, using speech synthesis architecture of Festival. The system was developed based on concatenative synthesis and diphones were used as the basic concatenation units to synthesize

Wolaytta sample text. As has been indicated in the research paper, Residual LPC technique was used to generate the synthetic speech. The overall performance of the system is found to be 78%. He also adopts Mean Opinion Score (MOS) testing method for intelligibility and naturalness of the synthesized speech, in which intelligibility resulted 3.17 and 2.77 for naturalness. As a final point, he recommend extensive identification and inclusion of non-standard words (NSWs) in the language with efficient representation method.

2.8.4. Concatenative Amharic TTS System

A research work by Henok [8], applies concatenative speech synthesis. Diphones and syllables were used as the basic concatenation units to synthesize sample and with Time-Domain Pitch Synchronous Overlap and Add (TD-PSOLA) technique to generate the synthetic speech. In addition to this, he has also considered prosodic effects, like anger, happiness and emotions, into account. For system evaluation he adopted two methods: Open Rhyme Test (ORT) and Mean Opinion Score (MOS). The ORT test result was obtained 88% and 75% for diphone based and syllable based synthesized words, respectively. Based on the MOS test result he concluded that syllable based synthesis gives better result than diphone based.

Another Amharic language TTS system was done by Sebsibe [20] with other abroad researchers, it titled as "Unit Selection Voice for Amharic Using Festvox". They developed a unit selection concatenative speech synthesizer by using transliteration scheme to work with Amharic scripts and incorporated Amharic phone set, syllabification rules, letter to sound rules into Festvox. Festvox, which is a voice building framework used for building unit selection voices in a new language Festival speech synthesis system were used. The perceptual evaluation indicated in the research used six levels ranging from Excellent (5) to Very Poor (0) and resulted is 2.9. Finally, they recommend the proper selection of unit and optimal selection of corpus will give better quality.

Even if these works have a great contribution in the area, speech synthesis on Amharic language has not yet thoroughly explored like other abroad languages. 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 language with the integration of non-standard words (NSWs), such as, numbers, abbreviations, acronyms, currency and dates representations found in orthographic text of a language. Hence in this study an attempt is made to develop a generalized TTS system that handles both standard words and non-standard words using Residual Exited Linear Predication (RELP) coding technique in Festival tool.

CHAPTER THREE

AMHARIC PHONOLOGY

Amharic is the official language of Ethiopia, and working language of most regional states (such as Amhara, Southern Nation and Nationality People (SNNP), Addis Ababa). Amharic has been a written language for roughly 600 years and has as rich legacy of both typeset and calligraphic literature [23]. It has its own non Latin based syllabic script called "Fidel" or "Abugida" [6][24]. Written Amharic uses a unique script originating from the Ge'ez alphabet, which is the liturgical language of the Ethiopian Orthodox Church [24]. Ge'ez is the ancient language of Ethiopia that is analogous in the role that Latin played for the Romance language of Europe [23]. In addition, written Ge'ez can be traced back to at least the 4th century AD [24].

3.1. Amharic Phonology

Amharic script has 33 core characters and out of each 32 of them are consonants having seven orders to show the seven vowels [20][24]. Basic alphabets are organized in the form of 33 rows by 7 columns table, as shown in *Appendix 'A'* [7]. Characters in the first column are called First order, the remaining columns are also labeled as: Second, Third, Fourth, Fifth, Sixth and Seventh orders

according to their position in the table. Out of the seven derivatives six of them are CV (Consonant vowel) combinations while the sixth is the consonant itself [20]. Other symbols representing labialization, numerals, and punctuation marks are also available [7].

3.1.1. Consonant Phonemes

The phonetic alphabet is usually divided in two main categories: namely, vowels and consonants [9]. Consonants are characterized by significant constriction or obstruction. This is mainly due to articulators come close to each other and even sometimes touches with each other, while producing consonants. Consonant sounds may be either voiced or unvoiced [9]. In addition, as opposed to vowels, consonants involve a very rapid change. They are more difficult to synthesize properly.

In Amharic there are about 32 consonants. But there are many cases where a single phoneme produced under numerous symbols that have extremely different orthographic form. For instance, (‘ሰ’ and ‘ሠ’), (‘ጸ’ and ‘ፀ’), (‘አ’ and ‘ዐ’) and (‘ሀ’, ‘ሐ’, ‘ኀ’ and ‘ኸ’) are redundant in Amharic [23]. Those have different orthographic forms but still having the same meaning and representing the same phoneme. Therefore, there are only 27 phonemes [6][7][8][22], which represent

different sound. The remaining are redundant sounds that represent the same sound.

The sound of a given consonant can be determined by three major factors. These are, the place of articulation, manner of articulation and voicing [7][22]. The place and manner of articulation refers the place where the narrowing occurs – which articulators gets close to each other and the relative position and activity of articulators while forming the obstruction, respectively. The third, voicing, refers the state of vocal cords.

Generally, Amharic consonants are classified as stops, fricatives, nasals, liquids, and semi-vowels [7][20][22]. Table 3.1 summarizes the phonetic representation of the consonants of Amharic and their manner of articulation, voicing, and place of articulation.

Manner of Articulation	Voicing	Place of Articulation											
		Labials		Alveolar		Palatals		Velars		Labiovelar		Glottal	
Stops	Voiceless	p	ፕ	t	ት			k	ከ	kx	ኩ	ax	ዕ
	Voice	b	ብ	d	ድ			g	ግ	gx	ጉ		
	Glottalized	px	ጵ	tx	ጥ			q	ቀ	qx	ቁ		
Fricative	Voiceless	f	ፍ	s	ስ	sx	ሽ					h	ሀ
	Voice	v	ቭ	z	ዝ	zx	ዥ						
	Glottalized			xx	ጽ							hx	ኋ
Affricative	Voiceless					c	ች						
	Voice					j	ጅ						
	Glottalized					cx	ጭ						
Nasals	Voiced	m	ም	n	ን	nx	ኝ						
Liquids	Voiced			l	ል								
				r	ር								
Glides		w	ው			y	ይ						

Table 3.1: The Amharic consonants and their phonetic representation

3.1.2. Vowel Phonemes

Vowels are always voiced sounds and they are produced with the vocal cords in vibration [9]. The tongue shape and positioning in the oral cavity do not form a major constriction of air flow during vowel articulation [7]. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically [9].

Vowels can be classified by the position of the tongue and the lips. The tongue and the lips produce different vowels by altering the shape of the vocal trace and enable the vibrating air produce sound in which different frequencies are emphasized [8].

As Getahune [22] discussed, vowel sounds in Amharic language are seven in number, these are (ኧ, ኡ, ኢ, ኣ, ኤ, እ and ኦ). However, another Amharic character (ዐ) and its variation also represent as vowel out of the 33 basic forms [7], this work only considers those seven vowels listed above. Amharic vowels are found with a combination of each consonant in CV (Consonant Vowel) manner. Each consonant should be pronounced as one with a combination of its vowel, otherwise consonants by itself cannot give sound independently [22]. Table 3.2 shows how Amharic vowels amalgamate with each consonant, specifically with character 'በ' as present in [22] and CV transcription adapted from [20].

1st	2nd	3rd	4th	5th	6th	7th
በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
ብኧ	ብኡ	ብኢ	ብኣ	ብኤ	ብእ	ብኦ
B/e/	B/u/	B/i/	B/a/	B/ie/	B	B/o/

Table 3.2: Amharic CV and transcriptions

Amharic vowels can be classified in to different categories depending on the position of tongue and lip. Under the shape of lip, vowels in Table 3.2 can be categorized as rounded (ኡ/u/ and ኦ/o/) and unrounded (ኧ/e/, ኢ/i/, ኣ/a/, ኤ/ie/, and እ/ee/) [22]. For instance let us consider two Amharic words "metu" /መጡ/ which means *they came* and "lbie" /ልቤ/ which means *my heart*. The first one "metu" /መጡ/ can be expanded as /ምጣጥኡ/, while saying the last vowel /ኡ/ lips end with rounded shape. On the other hand, the last vowel /ኤ/ in the second word "lbie" /ልቤ/ can be also expanded as /ልእብኤ/ force lips to be round out.

According to the positions of tongue linguistically important dimensions are generally in the ranges of [front to back], [high to low] and reverse [7][22]. Moreover, Figure 3.1 shows Amharic vowels dimensional ranges [22].

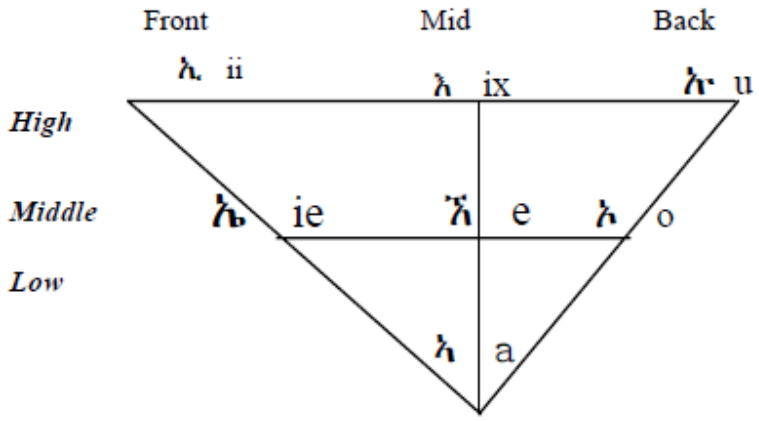

Figure 3.1: Amharic vowels map.

Though Amharic vowels are found with a CV manner, in table 3.2 CV transcription of the orthographic form the sixth order is the consonant itself only. Despite that the symbol may possibly associate with the vowel **/ix/** in its spoken form [7] and in another literature [8] associate with **/ɨ/**, this work preferably follow the CV transcription present in [20] and which shows in the above table.

3.2. Amharic Non standard words (NSWs)

Written documents of Amharic language contain unrestricted texts which include both standard words (common words and proper names) and non-standard words (NSWs). Standard words (SWs) have a specific pronunciation that can be phonetically described either in a lexicon, using a disambiguation processing to some extent, or by letter-to-sound rules [15][27]. By definition, NSWs comprise numerical patterns and alphabetic strings that do not have a regular entry in a lexicon and their pronunciation needs to be generated by a more complicated natural language process [15][27].

Since the core reason of this thesis is to identify and integrate Non-standard words in Amharic language and producing speech for unrestricted text of the language TTS system become fundamental part. As described above, NSWs in Amharic also cannot give sense through an application of letter-to-sound rule. As a result the text analysis part more worked on normalizing NSWs to convert

into pronounceable words. Some examples of them in Amharic language are given in table 3.3.

NSW category	Written format	Pronunciation
Cardinal number	12, ½, 0.6	አስራሁለት, ግማሽ, ዜሮ ነጥብ ስድስት
Ordinal number	ለ1, 3ኛ, በ10	ለአንድ ፣ ሶስተኛ, በአስር
Date	12/7/92	አስራሁለትኛው ቀን ሰባትኛው ወር ዘጠናሁለት
Telephone number	0912358690	ዜሮ ዘጠኝ አስራሁለት ሰላሳአምስት ሰማንያ ስድስት ዘጠና
Year(s)	1997	አስራዘጠኝ ዘጠና ሰባት
Time	2:30	ሁለት ሰዓት ተኩል
Ratio	3:4	ሶስት ለ አራት
Range	10-15	አስር እስከ አስራአምስት
Special character	%	በመቶ
Acronym	አአዩ ሜ	አዲስ አበባ ዩኒቨርሲቲ ሜጭር
Abbreviation	አ.ም (ዓ.ም)	አመተ ምህረት

Table 3.3: Category of NSWs in Amharic language[5]

However, there are many ambiguities in Amharic NSWs, which is difficult to determine in one category as like above listed. For instance, the given NSW

[5] *These are collected from different common Amharic news papers like Addis Admass, Addis neger, Awramba times and Addis Zemen.*

“3:10” can be consider as time and also ration. The same way NSWs like “1997/8” and “1990-1995” can also be pronounces as range of years or cardinal number range. In addition, it is common in Amharic text to write date as “ሰኔ 05/2001” format, and others like “ቀበሌ 01/02”, “5.50 ብር”, “ቢሮ ቁጥር 405/2”, “1.70 ሜትር” and “አንቀፅ 8.1.1” require to be pronounced differently.

3.3. Syllable Structure

Amharic words are characterized by weak, indeterminate stress; presence of glottal, palatal and labialized consonants; frequent geminate consonants; high frequency of the central vowels, and use of an automatic helping vowel /ix/ [20]. Though strict definition of syllable is difficult, a word in Amharic could be monosyllabic like “na” /ና/ (meaning come) or polysyllabic like “al.me.ta.ciim” /አልመጣቸም/ (meaning she didn’t come), which consists of four syllables. All syllables have a vowel nucleus. Several researchers studied the syllable structure of Amharic language and came up with different syllable template. For example, [20][22] states the six possible syllabic structures in Amharic as V, VC, VCC, CV, CVC, and CVCC. Moreover rarely initial cluster could exist when the second consonant in the cluster is liquid (and form CCV and CCVC) [20]. As Sebsibe et al [20], quoting Mulugeta (2001) discusses that the syllable structure of Amharic as CV and CVC only and also different in [24].

In this chapter a number of language related issues are assessed. The detail discussion of Amharic language, helps to understand Amharic phonology, standard and non-standard words and syllable structures in order to select techniques/algorithms that are appropriate to design a speech synthesis to Amharic language. Next chapter, discuses algorithms used in this research work.

CHAPTER FOUR

TTS ALGORITHM

From the speech synthesis methods, concatenative synthesis is one of frequently used to join short segments of speech usually taken from a pre-recorded database. But, in the course of converting text to speech distortion of synthetic speech can be introduced mainly due to inappropriate of selection the boundaries between speech segments or insufficient merging of segments and by the prosodic modification process, due to an insufficiently robust speech modification model [28].

Applying signal processing techniques to the synthesis speech units which change their pitch and duration, and to smooth away spectral concatenation discontinuities between units, can solve the above problems [28].

The algorithms described here are based around the LPC analysis/synthesis framework, and achieve prosodic modification by time-domain processing of the LPC residual.

4.1. Residual Excited Linear Prediction (RELP) Coding

The Residual Excited Linear Prediction (RELP) coding is one of the Linear Predictive Coding (LPC)-based codes. It is one of the standard methods for resynthesis, which is also used in Festival [10][29]. Its compression rate is moderate because the RELP needs to encode a sequence of residual signals for exciting the vocal tract model synthesized from speech signals. Moreover, the quality of synthesized speech is superior to other kinds of LPC. The system is robust since there is no need to analyze whether the sound is voiced or unvoiced nor to analyze the pitch period. The RELP consists of five functional blocks: an LPC analyzer, a residual encoder, a residual decoder, a spectral flattener, and an LPC synthesizer. As required by this method, pitch marks, Linear Predictive Coding (LPC) parameters and LPC residual values had to be extracted for each diphone in the diphone database [12].

The RELP method key concepts are discussed as follows [29]. In this method, first LPC analysis has to be carried out on the original speech to obtain LPC parameters. In particular, this is a technique for modifying the original residual so as to produce a synthetic residual of desired fundamental frequency (F0) and duration [28]. Then, inverse filtering is performed to get the residual signal. For instance, consider original speech sample x[n] which can be predicted as a linear

combination of the previous p (linear prediction order) samples, as given below [29]:

$$\tilde{x}[n] = \sum_{i=1}^{p} - a_i \ \chi[\mathrm{n} - \mathrm{i}]$$

……..……..Equation 4.1

where ai are prediction coefficients and x[n-i] are past speech samples. The prediction error due to this approximation is [29]:

$$e[n] = x[n] - \tilde{x}[n] = x[n] + \sum_{i=1}^{p} - a_i \ \chi[\mathrm{n} - \mathrm{i}]$$

…… Equation 4.2

This error is known as the *residual signal*, which can be used as the excitation to the LPC filter to get a perfect reconstruction of the speech signal [29]. The LPC coefficients were used to create an inverse filter and the speech frame was filtered using the inverse filter also the output of the inverse filter is the LPC residual [30].

During LPC analysis the LPC parameters computed using asymmetric Hanning-windowed pitch-synchronous frames of the original speech as shown in figure 4.1 [29].

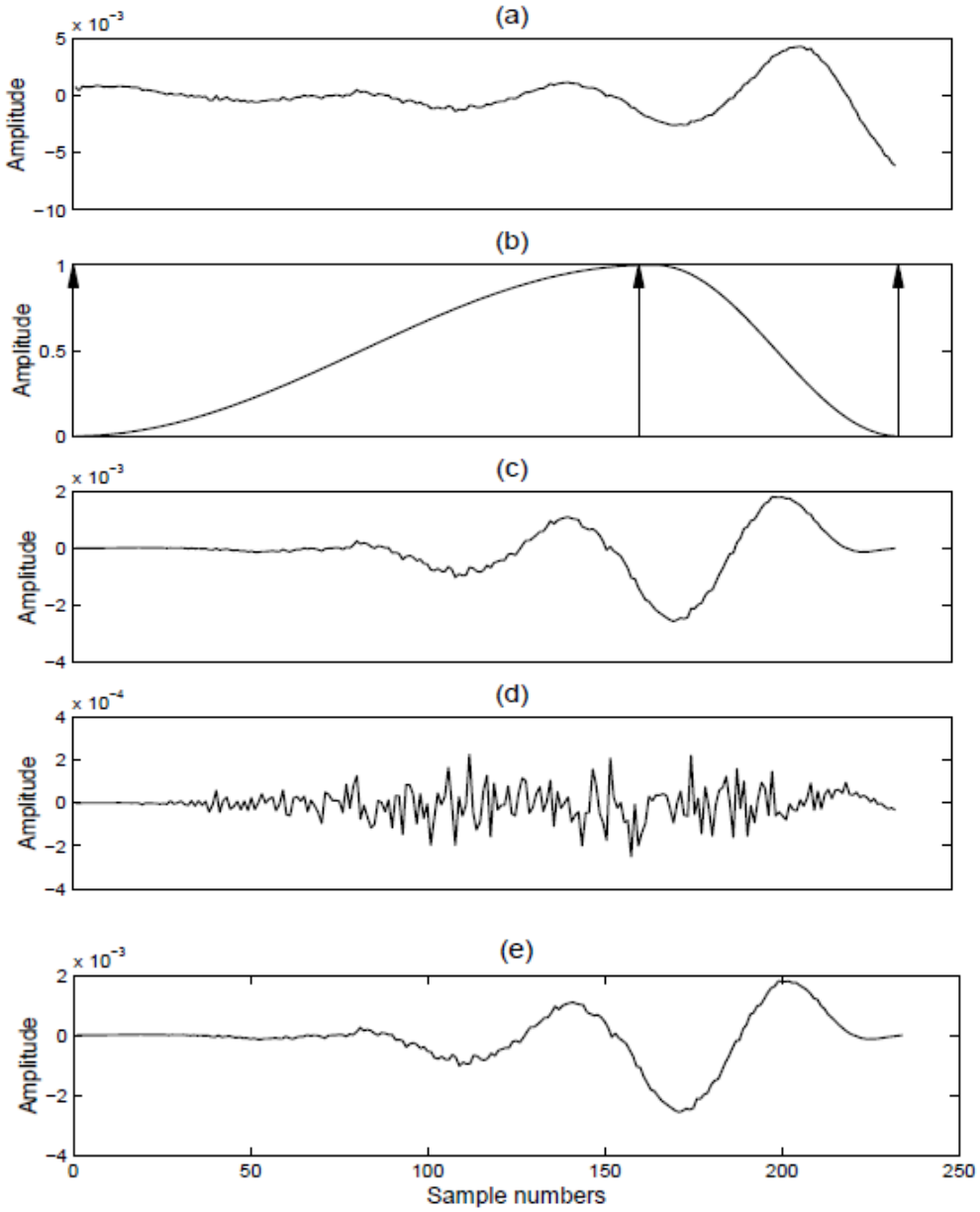

Figure 4.1: RELP synthesis using an asymmetric window

In the above figure: the first one (a) show the original waveform, (b) asymmetric window with pitch marks shown as arrows, (c) windowed original waveform, and figure (d) and (e) shows the residual signal and reconstructed waveform, respectively.

The advantage of using the asymmetric window can be observed in the figure, where successive pitch periods are very different in size and the window is not centered. The sample plots shown in the figure are two pitch periods in length. The residual is computed by passing the windowed original speech (plot (c)) through the inverse LPC filter. A sample residual signal is depicted in plot (d) of the figure 4.1. The residual is not modified by the smoothing operation. Then, the LPC filter is excited using the residual to reconstruct the output speech waveform. In figure 4.1, the output waveform is depicted in the last plot, which is a reconstruction of the original signal.

Finally, Festival divided each diphone into overlapping pitch-synchronous windows and the respective LPC coefficients and the residual signal are stored in the diphone database. And, then to get the full synthetic waveform for an utterance these pitch-periods output waveforms are overlap and add. For instance, while the sample texts given, windowing function applied to the edge of each diphones so that the samples at the juncture have low or zero amplitude, and if both diphones are voiced diphones are joined pitch-synchronously. Figure 4.2 shows the signal zero crossing position that Festival breaks and when the system synthesizes the voice its try to match [11].

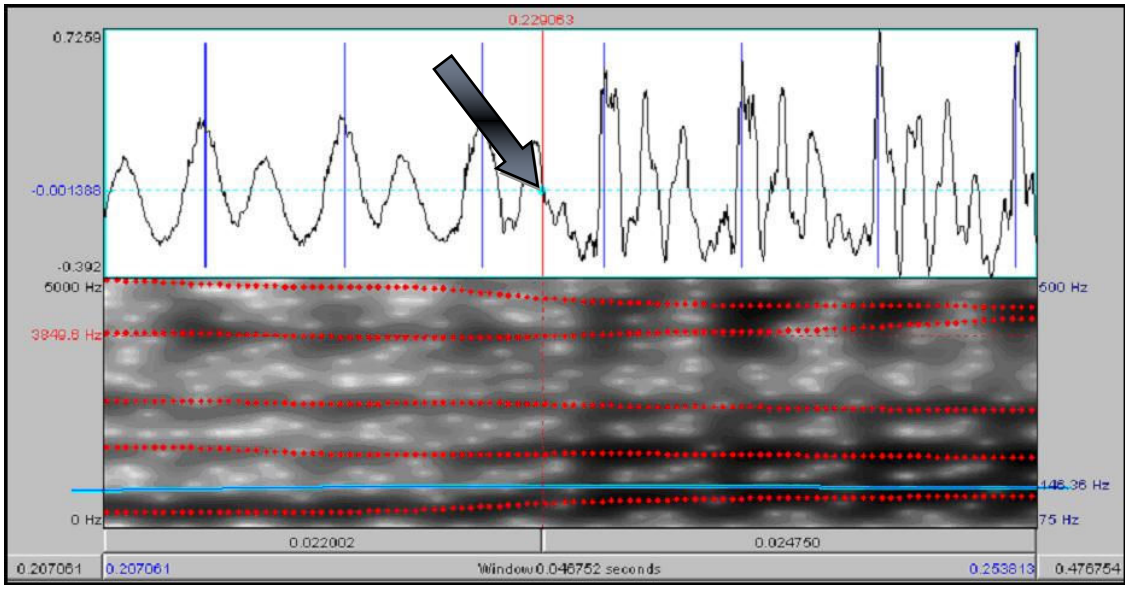

Figure 4.2: Zero crossing position that signals split and match

4.2. Linear Prediction theory

Linear Prediction (LP) is a good tool for analysis of speech signals. From a frame of natural speech and using computationally efficient algorithm the digital filter is estimated automatically. LP synthesis has been used extensively in concatenation system, since it enables the rapid coding of concatenation unit and the relationship between the coefficients used to define the LP filter [14].

Basically linear prediction theory assume that the current speech sample *y(n)* can be predicted as a linear combination of the previous *P* sample of speech, plus a small error term *e(n)*, and for which $1 \leq n \leq N$ [14]

$$e(n)^n = \sum_{i=0}^{p} a(i)y(n-i)$$ ……... Equation 4.3

where *a(0)=1* and the *a(i)* are termed the linear prediction coefficients, and *P* the linear prediction order. The coefficients, *a(i)*, are found minimizing the sum of the squared errors over the frame of speech under analysis. Two methods of performing this calculation are commonly used, termed the *covariance* method and the *autocorrelation* method, which differ in the range of *n* over the error is minimized. But this thesis work only consider autocorrelation function $r_y(i)$ and is defined as [14].

$$\mathbf{r}_{\mathbf{y}(i)} = \sum_{n=1}^{N-i} y(n)y(n+i)$$ …… Equation 4.4

The algorithm discussed here is used to synthesize Amharic TTS. The system is developed using residual excited LPC technique to synthesize words in Festival framework. The next chapter deals with the detailed description of the TTS developed.

CHAPTER FIVE

EXPERMENTATION AND EVALUATION

In this research, Text to speech synthesizer for Amharic language is implemented. The system is enabled to convert both standard words and non-standard words in the text. The performance of the system is also evaluated using test datasets and its level of intelligibility and naturalness is checked based on users' acceptance test.

5.1. Generalized Architecture of Amharic TTS System

The integration of speech and language technologies is the next step in the evaluation of Human Computer Interaction (HCI) which enabling users to carry out spoken dialogue with computer. As discussed in chapter two, concatinative synthesis techniques give the most natural sound in speech synthesis. However, the unit chosen and algorithm used to concatenate the units smoothly play a crucial role to produce high quality speech, even if a trade-off exists between longer and shorter unites.

The TTS system for Amharic are performed mainly through two processing modules of synthesizer and within different phases as shown in figure 5.1.

Besides, in order to keep doing with those modules initially appropriate diphone database is developed. The text analysis phase, within natural language processing module, converts non standard words to standard ones. The phonemic analysis phase is a grapheme-to-phoneme converter, converting the written text into a sequence of phonemic symbols. The prosodic analysis module then takes the phoneme sequence, and assigns to each phoneme the required pitch and duration. Both the phonemic and prosodic analyses are typically language dependent. Then, finally digital signal processing module accepts individual phonemes associated with their prosodic information and then produce synthesized speech, which is the outcome of the process.

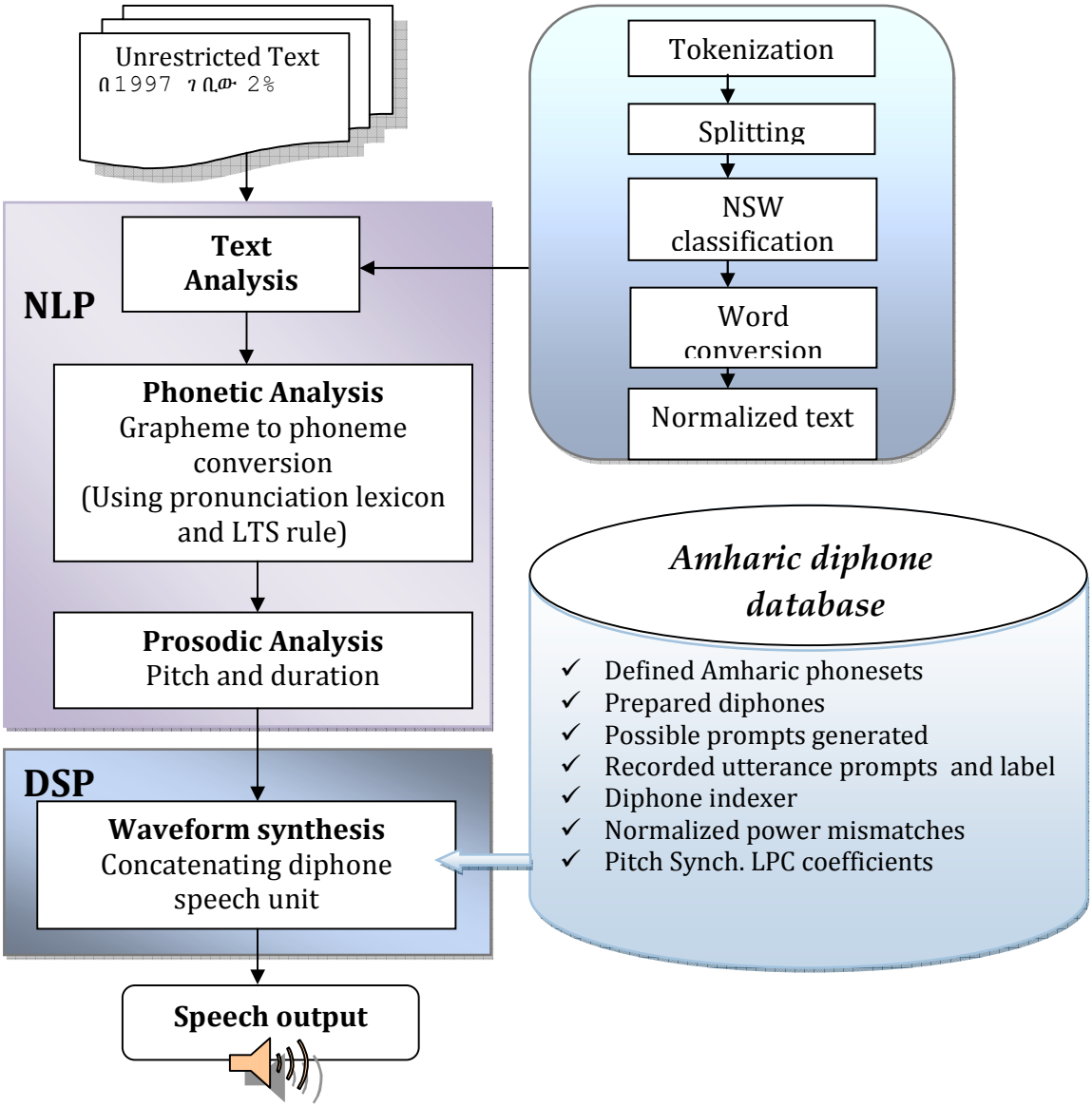

Figure 5.1: The architecture of generalized Amharic TTS system

5.2. Diphone Database Construction

One of the most vital tasks in the development of a text to speech system for Amharic language speech synthesizer is a construction of diphone database. Prior to constructing the diphone database, answers to the diphone-pairs exist in

the language and carrier words. Carrier words should be used to compile set of words each containing an encoded diphone, and, then it afford diphones to be concatenated by the synthesizer.

As initial point for construction of Amharic diphone database the required set of skeleton files and directories are generated using Festvox speech tools command[6]. In addition, as a standard adopted in Festival speech tools, three arguments 'aau', 'am' and 'alu' are added at the end of the command, those represent the institution, language and speaker name, respectively.

Amharic phone set definition is the first and cornerstone task for proceeding further in database construction. The identified phone-set for Amharic in section 3.2[7] is implemented in the file '*festvox/aau_am_alu_phoneset.scm*'. The proposed set of transcription symbol scheme is found to be a representation scheme for Amharic phoneme. Along with the consonant and vowel transcription symbols, features such as length, height, place of articulation, manner of articulation and voicing are defined. Apart from the identified set of phonemes, new phoneme that is useful in describing silence is also defined.

---

[6] *$FESTVOXDIR/src/diphones/setup_diphone aau am alu*

[7] *Specifically see table 3.1 for Amharic consonants and their phonetic representation.*

The number of diphones in a language is roughly the square of the number of phoneme. In Amharic language there are about 39 phonemes [8]. Therefore, roughly 1521(39 x 39) diphones exist. Using the phone-set defined above, the first phase involved the preparation of matrices mapping all possible combinations of consonants and vowels; i.e. CV, VC, VV, CC, #V, #C, C# and, V#. Here '#' denotes a short period of silence. Then all possible combinations of diphones are generated automatically using a command[8] available in Festvox speech tools. This command produce list of diphones as shown in *Appendix 'B'* and with a file name 'amdiph.list'.

However, more to the point of those generated diphones, due to various phonotactic constraints, not all phone-phone pairs occur physically. Such non-existent diphones were identified, but to the knowledge of the researcher this is the challengeable task to go through all diphones generated and checking the existence or not in the language. Those identified diphones were filtered manually from automatically generated diphone list. Finally, 1443 diphones were determined.

[8] *Festvox/diphlist.scm Festvox/am_schema.scm '(diphone_gen_schema "am" "etc/amdiph.list")'*

These diphones were embedded in carrier sentences including four other nonsensical context words. A care was taken when coining these nonsensical words, so that these words act in accordance with phonotactics of the Amharic language. After the diphones determined, using the voice intended for the database the prompts and their corresponding label files need to be prepared so that the recordings are done.

The next phase involved recording the diphones utterance on speech analysis software tool called Praat. Due to the shortage of time, for this work only writer's speech was recorded.

Literatures advice sounds preferable to be recorded in appropriate sound studio it in order to reduce noise level. As a result, recording was done with an optimum noise free environment time (at night time and quite room). These diphones were embedded in carrier phrases, which were generated with diphones for keeping consistency while recording. The diphone is put in the middle syllable of the middle phrase, minimizing the articulatory effects at the start and end of the diphone. Also, the use of carrier phrases helped the speaker to maintain a neutral prosodic context. Then the speech signal recorded with stereo microphone at 16 kHz sampling rate. Recorded speech signals were split into individual *wav* form files, pitch marks and diphone boundaries hand-labeled

using the speech analysis software tool Praat. Here again because of the challenges to ensure the presence of diphones in Amharic language and specifically with non-standard words (NSWs), those manually determined diphones are collected for Amharic speech synthesizer corpus.

5.3. Natural Language Processing (NLP)

When building a new voice using Festvox, the NLP modules should be constructed according to the language requirements. Hence, the language specific scripts (phone, lexicon, tokenization and text normalization) and speaker specific scripts (duration and intonation) should be configured. These language specific scripts are done under text and phonetic analysis phases of NLP module. On the other hand, speaker specific scripts are handled by prosody phase.

5.3.1. Text analysis

Along with language specific scripts and the whole Text-to-Speech system design the first is text analysis, which identifies pronounceable words from raw text. In the current text analysis field, text normalization is considered as a crucial component of text analysis in TTS [21]. It involves the work on the real text, where many Non-Standard Word (NSW) representations appear, for example, numbers (year, time, ordinal, cardinal, floating point), abbreviations, acronyms,

currency, dates, URLs. However, NSW cannot be detected by an application of "letter-to-sound" and may be recognized as different standard words depending on both the local text and the text genre. As a result, all of these non-standard representations should normalize, or in other words convert to standard words. As shown in figure 5.1, text analysis is designed to capable normalizing those words.

Initially text tokenization methodologies are implemented under the file template[9] produced for the language before normalization is done. White-space is most commonly used delimiter between words and is extensively used for tokenization [11]. In this work, as a delimiter for the tokenization process, white-spaces, tabs, new lines and carriage return characters are used to tokenize Amharic text. For example, the Amharic sentence "ሰለሞን ዛሬ 7 ዓመት ይሞላዋል", is tokenized as "ሰለሞን", "ዛሬ", '7', "ዓመት" and "ይሞላዋል" tokens. Then Festival converts the give sentence into ordered list of tokens for further process. Once the text has been tokenized, text normalization is carried out.

5.3.1.1. Amharic Non-Standard Words (NSWs) Normalization

This step converts non-standard words (NSWs) like digits, numerals, abbreviations, and non-alphabetic characters into word sequence depending on

[9] *Festvox/aau_am_alu_tokenizer*

the context. Using those identified Amharic non-standard words listed in chapter three, so each NSWs are identified as separate token by token identifier rules.

Those tokenized words above ("ሰለሞን", "ዛሬ", '7', "ዓመት" and "ይሞላዋል") given to *token to word* function which is responsible to return list of word for the token string match with the token identifier rule. For this work token identifier rule is constructed based on the regular expression schema of Festival and adapted to normalize Amharic non-standard words.

For instance, to identify the first ten (0-9) numeric digit token normalization rule set as, *(string-matches name "[0-9]")* and then those tokens matches with this condition rule its appropriate string word returned from the list set for it.

In line with tokens above, for example token word '7' match with token identifier rule set for its numeric value, then the appropriate string "ሰባት" is returned and replace it as ""ሰለሞን ዛሬ ሰባት ዓመት ይሞላዋል". Besides for numeric tokens these values greater than [0-9] rules are designed as like above.

For the rest of Amharic non-standard word rules set as, (*string-matches name "[0-9][0-9]*/[0-9][0-9]")* and *(string-matches name "[0-9][0-9]*.[0-9] [0-9]")* for cardinal number category. Ordinal number categories like *ክ70, ሰ9* and $3^{ተኛ}$ handled by the rule set as *(string-matches name "[YyBbKk]?[Ee]?[1-9]?[0-9]")* and *(string-matches*

name "[1-9]tenxa"). Date and year values are also handle by the rule, (*string-matches name "[0-9]?[0-9]/[0-9]?[0-9]/[0-9][0-9]")* and *(string-matches name "[0-9][0-9][0-9][0-9]a.m")* then for instance, the give Amharic text may contain *12/07/02* or *1979ዓ.ም* then these are normalized through above rules. Rules (*string-matches name "[0-9]?[0-9]:[0-9][0-9]"),* (*string-matches name "[1-9]:[1-9]") and (string-matches name "[1-9]?[0-9]*-[1-9]?[0-9]")* are set for NSWs category of time, ration and range, respectively. Lastly, for both acronym and abbreviation category of non-standard words, rules are set based on each type of the character that it contains and then their pronunciation string words are listed accordingly.

Nevertheless, still there are difficulties that need address for further work. The similarity between token strings, for instance token "2:10" only feat and handled by the rule set for the time, and can be pronounce as "ሁለት ሰዓት ከአስር", but it also can be ration and needed pronounce as "ሁለት አስረኛ". The value which contain token string "1870" match with the *token to word* rule set for cardinal number. However, it can be also year but the system only consider the token strings as year if it only contain "a.m" at the end as "1870a.m". Moreover, token "5.50" may represent as money (ብር), cardinal number or distance measure value (ሜትር). In general, the existence of ambiguities in NSWs become challenging and those are not consider in this work. Further work on Amharic language can

benefit from further finding on handling ambiguity in NSWs based on their contextual analysis.

5.3.2. Phonetic Analysis

Ones the given text is tokenized and normalized into pronounceable word, the second phase of natural language processing module of TTS system is phonetic analysis. This phase converts the pronounceable text word to its pronunciation form. The common means[10] for finding pronunciation of a word need large list of lexicon and Letter-to-Sound (LTS) rule [16]. Nevertheless, this is not just merely providing list of phonemic representation of each entry, but also the syllabic structure, part of speech tags and stress markers of Amharic language.

Phonetic analysis within Festival expects to build large amount of lexicon under its framework for the language. As a result, selected 936 Amharic language lexicon training datasets[11] are compiled by hand and prepared for pronunciation lookup process using the embedded scheme interpreter. In addition, for those words that their pronunciation are not explicitly sets in the lexicon, Festival provide as an optional way of finding the pronunciation words through Letter-to-Sound (LTS) rules.

---

[10] *Those can be also classified dictionary-based and rule-based, see section 2.4.2 for more discussion and reference [16].*

[11] *Collected from different (see footnote 5) Amharic news papers.*

The lexicon structure that is basically available in Festival takes both a word and a part of speech (POS) to find the given pronunciation. For instance, the Amharic word "ይመጣል" represent in the lexicon entry structured and compiled as *("ymetxal" nil ((y) 0) ((m e) 0) ((tx a) 1) ((l) 0)))*. Here the word is categorized in three parts. "ymetxal" stands for the root word entry to be pronounced. The second one is "nil"[12] which represent the part of speech, and identified syllable structure pronunciations and stress markings. Each entry in the Amharic lexicon is compiled using phoneme sets and these three parts. Moreover, since lexicon look-up process use binary search while new instances given, each lexicon entries are arranged based on their alphabetical order.

On the other hand, Festival also supports a smaller list called an *addenda.* Addenda, primarily provided to allow special applications and add entries that are not in the existing compiled lexicon. Nevertheless, when new instance is run into searching is done linearly; due to this reason when it compared with lexicon, it is less efficient. For instance, to add entry of Amharic special character as; *(lex.add.entry '("%" nil (((b e) 0) ((m e) 0) ((t o) 1))))*.

The second common means of finding the pronunciation of a word is Letter-to-Sound (LTS) rule; it is also called the Grapheme-to-phoneme (G2P). However,

[12] *"nil" value indicate part of speech tag is not consider in this work.*

because of the difficulty in building LTS for Amharic language, and the associated computational complexity in interpreting these rules, it is not considered in this work.

5.3.3. Prosodic Analysis

After the language specific scripts set along with speaker specific scripts the duration and intonation information should be configured here in prosodic analysis phase. As discussed in section 2.5.3, this stage focuses on finding correct intonation, stress, and duration from written text. Basic factors contributing to prosodic features are speaker characteristics, feeling and meaning of the sentence. During database preparation (i.e. duration) and in phonetic analysis phase (i.e. stress) the prosodic information set. In addition, the required diphone with their intonation and duration information is fetched and given to prosody.

For instance, for prosody phase an Amharic sentence "አንተ አትመጣም" is given. In this case the required diphone *"pau a an nt te e pau // pau a at tm me etx txa am m pau"* with their intonation and duration information is submitted to prosody. However, when a sentence is spoken by varying the level of intonations it gives two meanings. One, it become command "አንተ አትመጣም!" and the other is as question "አንተ አትመጣም?". In addition, the sentence may also express emotional

connotations such as anger and/or sadness. As a result, amalgamating those prosody features information is challenging.

A pitch-mark (pitch period), the location of the short-time energy peak of each pitch pulse in a speech signal, in other words, the beginning of a pitch period. Since pitch synchronous speech synthesis algorithms require the beginning location of the pitch period for every voiced segment prior to speech synthesis, residual excited Linear-Predictive Coding (LPC) also requires it. RELP coding, which is currently the only signal processing method publicly distributed in Festival framework.

The pitch mark identification for Amharic diphones speech signal database adapted from the Festvox speech tool script[13] and modified in this work. Mainly the modification involves the minimum allowed pitch period (min) and the maximum allowed pitch period (max) values to fit with the range of speaker is the critical. As discussed in literatures [10][25], a good starting point for a male speaker values in the range 0.005 and 0.012 (200 to 80 Hz), and then those values are set. Irrespective to extracting the pitch marks, one of the worthwhile post-processing stages that move the pitch marks to the nearest peak are done.

[13] *bin/make_pm_wave*

5.4. Digital Signal Processing (DSP)

Digital signal processing (DSP) transforms the symbolic information it receives from the NLP module into speech. Those symbolic information consists of phonetic transcription and prosodic information consisting of a list of phones associated with duration and a set of fundamental frequency (F0) targets.

In this work, using the residual excited LPC method of concatenation in Festival framework diphone concatenative synthesis is used for creating waveforms in the system. A diphone is a phone-like unit going from roughly the middle of one phone to the middle of the following phone. From a prerecorded database of diphones, diphone synthesizer model generates a waveform via identifying sequence of phones and then concatenate diphones.

Specifically, using LPC coefficients and the residual signal values of each respective diphones, those stored while LPC analysis process of diphone database construction. And, then each diphone is divided into overlapping pitch-synchronous windows. Consequently, making waveforms for each sample words encounter done, through a windowing function[14] applied to the edge of each diphones that the given word contains. Besides this is done either the samples at the juncture have low or zero amplitude, or diphones are joined pitch-

[14] *See section 4.1 for detail.*

synchronously if both diphones are voiced. That is mainly due to the pitch periods of the each diphones must line up with each other.

5.5. Performance Evaluation

The performance of the Amharic synthesizer is evaluated for both standard and newly integrated non-standard words (NSWs) of a language. As discussed before, standard words (SWs) have a specific pronunciation that can be phonetically generated either in a lexicon or by letter-to-sound rules. On the other hand, NSWs comprise numerical patterns and alphabetic strings that do not have a regular entry in a lexicon and their pronunciation needs to be generated by a more complicated natural language processing rule.

Performance of the system is evaluated using selected 80 Amharic words[15]. Those testing words are containing 50 standard words (common words and proper names)) and 30 non-standard words (NSWs). Table 5.1 presents performance evaluation results of Amharic synthesizer developed in this study.

[15] *See Appendix 'C' words used for evaluation.*

Performance Measure						
	Correctly Pronounced		Partially Pronounced		Not Correctly Pronounced	
	In number	In %	In number	In %	In number	In %
SWs	37	76.7	6	12	7	14
NSWs	21	70	3	10	6	17.3
Average %	73.35		11		15.65	

Table 5.1: Performance measure evaluation of Amharic synthesizer

Comparatively standard words (SWs) pronounce correctly than non-standard word (NSWs) based on the words given for both category evaluation 76.7% and 70%, respectively. In conclusion the overall performance of the system is found to be 73.35%.

The overall performance of the system is inclined by the performance of non-standard words. As observed from analysis the pronunciation of non-standard words did not map to its word string that could possibly found in the lexicon look up process. This is mainly due to the non-standard words need more complex rule to normalize. Moreover, existence of ambiguities in non-standard words while normalization is done and difficulties to find their appropriate string word in the specified rule conditions.

For instance, while non-standard word “01/02” is given the system generates error, due to leading zero of the token. But, when “1/2” is given it is pronounces as “ግማሽ”.

5.6. Evaluation of Sound Quality

Intelligibility and naturalness of synthesized speech is assessed in the second phase of experiment. Even if other technique of evaluation are available, for this thesis work Mean Opinion Score (MOS) technique is preferable chosen. It is a five level scale as shown in table 5.2. Based on the scale evaluators give their evaluation and the average of the opinion taken as the performance of the system [9].

Value	MOS
5	Excellent
4	Very Good
3	Good
2	Fair
1	Bad

Table 5.2: Scales used in MOS

Accordingly to evaluate the intelligibility and naturalness of synthesized speech six Amharic sentences are prepared as test dataset[16]. Each sentence contains a mixture of standard and non-standard words. Before the actual evaluation, questioner which filled by evaluators while they assess arranged by way of MOS technique require.

Then after six native speaker of Amharic language are selected and invite to assessment the system. The result of intelligibility and naturalness of synthesized speech evaluation present in table 5.3 and 5.4 respectively.

Evaluators	**Sentence and Ranks given by evaluators for intelligibility**					
	Sentence ***1***	*Sentence* ***2***	*Sentence* ***3***	*Sentence* ***4***	*Sentence* ***5***	*Sentence* ***6***
1	3	3	4	3	2	4
2	3	4	3	2	2	3
3	3	3	3	3	3	3
4	4	3	2	3	3	3
5	3	3	3	3	3	3
Average	***3.2***	***3.2***	***3***	***2.8***	***2.6***	***3.2***
Mean average	***3***					

Table 5.3: Amharic Speech Synthesizer Intelligibility (MOS) Scores

[16] *see Appendix 'D' the sentences used for evaluation*

Evaluators	**Sentence and Ranks given by evaluators for Naturalness**					
	Sentence ***1***	*Sentence* ***2***	*Sentence* ***3***	*Sentence* ***4***	*Sentence* ***5***	*Sentence* ***6***
1	2	3	3	3	3	3
2	3	3	3	2	2	3
3	3	4	3	3	3	3
4	2	2	3	3	2	3
5	3	3	3	3	3	3
Average	***2.6***	***3***	***3***	***2.8***	***2.6***	***3***
Mean average	***2.83***					

Table 5.4: Amharic Speech Synthesizer Naturalness (MOS) Scores

The quality of synthetic speech produced by the system is evaluated using five Amharic sentences. Based on the suggestion of six native speakers, the system produces synthetic speech with score of 3 and 2.83 intelligibility and naturalness, respectively. Accordingly, as per MOS score test the intelligibility of synthesized speech by the synthesizer is "good" and naturalness is also near "good".

As observed form listener's, even if utterances are recorded preferably in quite room, still the noise that originated from recording reduce the naturalness of the synthesized sound. This is due to unavailability of sound laboratory. Also the output quality gets hampered due to the problems occurred while transferring from one word to another word and concatenation of word with large diphones.

CHAPTER SIX

CONCLUTION AND RECOMMENDATION

5.1. Conclusions

This thesis work describes the development and evaluation of a generalized TTS system for Amharic language texts, with a consideration of non-standard words (NSWs) possibly found in Amharic text. The system is developed based on the framework of Festival using Residual Excited Linear Predictive (RELP) coding synthesizer. The system designed and described in this work has three major parts; namely, diphone database construction, natural language processing and digital signal processing parts. Diphone database constructed to be used as a base store for Amharic data, such as phone-sets, text utterances and recorded diphone sounds. NLP used to process language specific scripts like phone, lexicon, tokenization and text normalization, and also speaker specific scripts such as, duration and intonation. Consequently, DSP is responsible to perform speech production.

The performance of the system shows on the average an accuracy level of 73.75% for Amharic text containing both NSWs and SWs. In addition, the system achieves 3 and 2.8 MOS score for intelligibility and naturalness. The result looks

encouraging and shows the possibility of integrating non-standard words (NSWs) during text to speech conversion. Normalization of non-standard words to appropriate string word is challengeable because of inconsistency in the use of abbreviations and pronunciation.

Improvement of intelligibility and naturalness also depends in particular on proper lexical stress assignment, duration and a more sophisticated generation of prosodic features.

The attempts at developing Amharic TTS system integrated with non-standard words (NSWs) and that can synthesize Amharic non-standard words are promising.

5.2. Recommendation

This research is an attempt to see into the possibility of integrating Amharic non-standard words on TTS system. Based on the findings, the following recommendations are forwarded for further work to improve the level of non-standard word integration and total quality of the system.

Text normalization is not a trivial task; in this work rule-based mapping process is followed to convert non-standard words to their equivalent standard words. But, the inconsistency of usage of non-standard words it is challenging to

generate rule-based mapping schema. Hence, to consider all NSWs, there is a need to use statistical techniques such as, n-grams, Markov model, neural networks or classification and regression tree (CART).

In addition, the existence of ambiguities in non-standard words also another challenge that in need addressing in further work. Using more advanced rule such as Speech Synthesis Markup Languages (SSML) and Java Speech API Markup Language (JSML) for raw data converted to simple XML based markup format. For splitting NSWs, ambiguities detection and also to incorporate all type of non-standard word.

The lexicon look up process find the pronunciation of new arrival words, first to the compiled lexicon, if it is not found in the lexicon, then it find to LTS rule of the system. Developing large number of pronunciation lexicon, automatic lexicon entries instead of adding manually, find out LTS or Grapheme-to-Phoneme (G2P) rule to handle unknown words also important.

Speech synthesizer has to consider prosody, which take in to account speaker specific intonations and speaker specific duration. The major challenge for building prosody for Amharic is the lack of a part of speech POS tag set, POS tagger and tagged text corpus. To have better system quality with prosody analysis building part of speech is crucial.

Still the quality of the system is affected by unavailability of sound laboratory. Recording in appropriate sound studio for having better quality output is also contributing significant role.

REFERENCE

[1]. **Habtamu Haye,** (2007) "Amharic concatenative Text- To-Speech (TTS) synthesis system using syllabic unit", MSc project, Addis Ababa University, Addis Ababa, Ethiopia.

[2]. **Tewodros Abebe,** (2009)**. "**Text-to-Speech Synthesizer for Wolaytta Language", MSc thesis, Addis Ababa University, Addis Ababa, Ethiopia.

[3]. **Tesfay Yihdego,** (2004) "Diaphone Based Text-To-Speech Synthesis System for Tigrigna Language", MSc thesis, Addis Ababa University, Addis Ababa, Ethiopia.

[4]. **Nebiyou Tsegaye,** (2005). "Speech to text conversion using Amharic characters", MSc thesis, Addis Ababa University, Addis Ababa, Ethiopia.

[5]. **Richard Sproat**, (1996). "Multilingual text analysis for text-to-speech synthesis", Speech Synthesis Research Department, Bell Laboratories, Murray Hill, USA.

[6]. **Tadesse Anberbir and Tomio Takara**, (2009). **"**Development of an Amharic Text-to-Speech System Using Cepstral Method", Association for Computational Linguistics, Athens, Greece.

[7]. **Nadew Tademe,** (2008). "Formant based speech synthesis for Amharic vowels", MSc thesis, Addis Ababa University, Addis Ababa, Ethiopia.

[8]. **Henock Lulsegede,** (2003). "Concatenative Text-to-Speech (TTS) synthesis for Amharic language", MSc thesis, Addis Ababa University, Addis Ababa, Ethiopia.

[9]. **Lemmetty S.**, (1999). "Review of Speech Synthesis Technology". MSc. Thesis. Laboratory of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland.

[10]. **Black, Alan W., and Lenzo, A.,** (2003). "Building Synthetic Voices". Available at URL: http://Festvox.org/ [Accessed on: December, 2009].

[11]. **Alem F., Kanti Nath P. and Khan M.,** (2007). "Text To Speech for Bangla Language using Festival" BRSC University, Bangladesh.

[12]. **Wasala, A., Weerasinghe, R., Gamage, K,** (2007). "A Sinhala Text-to-Speech System", Language Technology Research Laboratory, University of Colombo School of Computing, Sri Lanka.

[13]. "**Speech synthesis"** – Wikipedia the free encyclopedia, available at URL: http://en.wikipedia.org, [Accessed on: January,2010]

[14]. **Donovan, E.,** (1996). "Trainable Speech Synthesis" PhD Dissertation, University of Cambridge, UK.

[15]. **G. Xydas, G. Karberies and G. Kouroupertroglou** (2004). "Text Normalization for the Pronoucation of Non-Standard Words in an Inflected Language" 3rd Hellenic Conference on Artificial Intelligence, Samos, Greece.

[16]. **Dutoit, T.,** (1997). "A Short Introduction to Text-to-Speech" Kluwer Academic Publishers, Dordrecht, Boston, London.

[17]. **Medina A., Herrera J. and Alvarado M.,** (2009). "Towards the Speech Synthesis of Raramuri: A Unit Selection Approach based on Unsupervised Extraction of Suffix Sequences" A. Gelbukh (Ed.) Advances in Computational Linguistics, Research in Computing Science 41, 2009, pp. 243-256

[18]. **Yibeltal Tafere,** (2008). "Formant-Based Speech Synthesis: a Case of Amharic Words" MSc project, Addis Ababa University, Addis Ababa, Ethiopia.

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

[20]. **Sebsibe H/Mariam, Kishore, S., Black, Alan W., Kumar, R., and Sangal, R.,** (2004). "Unit Selection Voice for Amharic Using Festvox", ISCA Speech Synthesis Workshop, Pittsburgh, PP. 103-107.

[21]. **Tao Zhou, Yuan Dong, Dezhi Huang, Wu Liu, Haila Wang,** (2008). "A Three-Stage Text Normalization Strategy for Mandarin Text-To-Speech System", Beijing University of Posts and Telecommunication, and France Telecom R&D, Beijing. Available at URL: http://www.ieee.com/ [Accessed on: February, 2010].

[22]. **ጌታሁን አሚረ (ረ/ፕሮፌሰር)** ፣ (2001 ዓ.ም) "*ዘመናዊ የአማርኛ ሰዋሰው በቀላል አቀራረብ*" ፣ ቁጥር 9 ፣ አልፋ አሳታሚዎች ታተመ ፣ አዲስ አበባ፣ ኢትዮጵያ።

[23]. **Daniel Yacob,** (2005). "Developments towards an Electronic Amharic Corpus", Ge'ez Frontier Foundation, USA, Available at URL: ftp://ftp.geez.org/pub. [Accessed on: February, 2010].

[24]. **BjÖrn G., Fredrik O., Atelach Alemu, Lars Asker,** (2009) "Methods for Amharic Part-of-Speech Tagging" Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT,, pp- 104–111, Athens, Greece.

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

[26]. **Nigel Rochford,** (2003). "Developing a New Voice for Hiberno-English in the Festival Speech Synthesis System" Computer Science, Linguistics And French, Trinity College Dublin

[27]. **Sproat, R., Black, A., Chen, S., Kumar, S., Ostendorf, M., and Richards, C., (1999)** "Normalization of Non-Standard Words", WS '99 Final Report, *Computer Speech and Language*.

[28]. **Edgington, M., and Lowery, A**., (2000). ''Residual –Based Speech Modification Algorithms for Text-To-Speech synthesis'' BT Laboratories, Martlesham Health, UK. Available at URL: http://www.asel.udel.edu [Accessed on: February, 2009]

[29]. **J. Vepa and S. King,** (2005). "Subjective Evaluation of Join Cost and Smoothing Methods for Unit Selection Speech Synthesis" IEEE Trans, On Speech And Audio Processing: Submission Draft For Review.

[30]. **Ananth N. Iyer, Melinda Gleiter, Brett Y. Smolenski and Robert E. Yantorno,** (2003). "Structural Usable Speech Measure Using LPC Residual" International Symposium on Intelligent Signal Processing and Communication Systems **(**ISPACS), pp 236-240, Awaji Island, Japan,

Appendix A: Basic 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	ዶ	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: Sample Amharic Diphone List

```
( am_0017 "pau t aa z a z aa pau" ("z-a" "a-z") )

( am_0018 "pau t aa h a h aa pau" ("h-a" "a-h") )

( am_0019 "pau t aa l a l aa pau" ("l-a" "a-l") )

( am_0020 "pau t aa c a c aa pau" ("c-a" "a-c") )

( am_0021 "pau t aa nx a nx aa pau" ("nx-a" "a-nx") )

( am_0022 "pau t aa r a r aa pau" ("r-a" "a-r") )

( am_0023 "pau t aa w a w aa pau" ("w-a" "a-w") )

( am_0024 "pau t aa y a y aa pau" ("y-a" "a-y") )

( am_0029 "pau t aa b u b aa pau" ("b-u" "u-b") )

( am_0030 "pau t aa p u p aa pau" ("p-u" "u-p") )

( am_0031 "pau t aa d u d aa pau" ("d-u" "u-d") )

( am_0032 "pau t aa t u t aa pau" ("t-u" "u-t") )

( am_0033 "pau t aa g u g aa pau" ("g-u" "u-g") )

( am_0034 "pau t aa k u k aa pau" ("k-u" "u-k") )

( am_0035 "pau t aa q u q aa pau" ("q-u" "u-q") )

( am_0036 "pau t aa cx u cx aa pau" ("cx-u" "u-cx") )

( am_0414 "pau t aa t a t aa pau" ("t-a") )

( am_0415 "pau t aa t u t aa pau" ("t-u") )

( am_0416 "pau t aa t o t aa pau" ("t-o") )

( am_0417 "pau t aa t e t aa pau" ("t-e") )

( am_0418 "pau t aa t ie t aa pau" ("t-ie") )

( am_0419 "pau t aa t i t aa pau" ("t-i") )

( am_0981 "pau t aa s - b aa t aa pau" ("s-b") )

( am_0982 "pau t aa s - p aa t aa pau" ("s-p") )
```

```
( am_0983 "pau t aa s - d aa t aa pau" ("s-d") )

( am_0984 "pau t aa s - t aa t aa pau" ("s-t") )

( am_0985 "pau t aa s - g aa t aa pau" ("s-g") )

( am_0986 "pau t aa s - k aa t aa pau" ("s-k") )

( am_0987 "pau t aa s - q aa t aa pau" ("s-q") )

( am_1374 "pau u t aa pau" ("pau-u") )

( am_1375 "pau o t aa pau" ("pau-o") )

( am_1376 "pau e t aa pau" ("pau-e") )

( am_1377 "pau ie t aa pau" ("pau-ie") )

( am_1378 "pau i t aa pau" ("pau-i") )

( am_1380 "pau b aa t aa pau" ("pau-b") )

( am_1381 "pau p aa t aa pau" ("pau-p") )

( am_1382 "pau d aa t aa pau" ("pau-d") )

( am_1383 "pau t aa t aa pau" ("pau-t") )

( am_1384 "pau g aa t aa pau" ("pau-g") )

( am_1385 "pau k aa t aa pau" ("pau-k") )

( am_1386 "pau q aa t aa pau" ("pau-q") )

( am_1387 "pau cx aa t aa pau" ("pau-cx") )

( am_1388 "pau j aa t aa pau" ("pau-j") )

( am_1389 "pau hx aa t aa pau" ("pau-hx") )

( am_1390 "pau tx aa t aa pau" ("pau-tx") )

( am_1391 "pau kx aa t aa pau" ("pau-kx") )

( am_1392 "pau f aa t aa pau" ("pau-f") )
```

Appendix C: Words Used to Test Amharic TTS

Standard words		Non-standard words	
አበበ	ኢትዮጵያ	200	120
መቸ	ወረደ	1910	1/3
ነበር	ብዛት	2፡30	3ኛ
ይጀምራል	የደን	2	¼
ትምህርት	የአገራችን	8-9	2.5
ዘወትር	ና	%	
አጠቃላይ	ይመጣል	ዶ/ር	
ዓመት	ገቢው	510	
ምርት	ሰለሞን	7	
የቡና	ደረሰ	ዓ.ም	
በየአመቱ	ዛሬ	አ.አ	
ቀነ ሰ	ግን	በ18	
ሸፋን	ብር	½	
ላይ	ህዝብ	አአዩ	
ቤቱ	ተባለ	10/3/68	
አልመጣም	ምግብ	ወ/ሮ	
ሰላም	መንገድ	የተ.መ.ድ	
በላ	ማጡ	0.6	
እናት	ልቤ	5፡6	
አደጋ	ሲያድግ	ት/ቤት	
ፍለጋ	ህይወት	ቅ/ገ ብርኤል	
ያድጋል	ይሞላዋል	01/02	
ኑሮ	አባየ	$	
ተስፋ	ዘመናዊ	70	
ለጋስ	ምርጫ	አአዱ	

Appendix D: Sentences Used to Test Amharic TTS

1. አበበ 200 ብር ሰጠኝ፡፡

2. የአገራችን የደን ሽፋን 2% ደረሰ፡፡

3. ዶ/ር ስለሽ መቐ ይመጣል?

4. በ 1910ዓ.ም የኢትዮጵያ ህዝብ ብዛት 8-9 ሚሊዮን ነበር ተባለ፡፡

5. ትምህርት ቤቱ ዘወትር 2፡30 ላይ ትምህርት ይጀምራል፡፡

6. በአጠቃላይ የቡና ምርት በ 0.6 ሲያድግ ገቢው ግን በ ½ ቀነሰ፡፡

Declaration

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: _____

Signature: _____

Date: _____

Confirmed by advisor:

Name: _____

Signature: _____

Date: _____

Place and date of submission: Addis Ababa, July 2010.