ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

SYLLABLE-BASED TEXT-TO- SPEECH SYNTHESIS (TTS) FOR AMHARIC

BY

MULAT SHIFERAW SIYOUM

A THESIS SUBMITED TO

THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION SCIENCE

AAU, June, 2012

ADDIS ABABA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

FACULTY OF INFORMATICS

DEPARTMENT OF INFORMATION SCIENCE

SYLLABLE-BASED TEXT-TO- SPEECH SYNTHESIS (TTS) FOR AMHARIC

BY

MULAT SHIFERAW SIYOUM

Signature of the Board of Examiners for Approval

Name	Signature
1. Dereje Teferi (PhD) , Advisor	_____
2. Solomon Teferra (PhD), Examiner	_____
3. _____	_____

AAU, June, 2012

DEDICATION

*To my mom, who has always been there for me, I love my mom so I can't thank her enough for everything***.**

DECLARATION OF ORIGINALITY

I hereby declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute. Any help that I have received and used in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated and cited in the thesis work.

Declared by:

Name: _____

Signature: _____

Date: _____

Confirmed By:

Dereje Teferi (PhD)

Signature: _____

Date: _____

ACKNOWLEDGEMENT

I am humbly grateful to my Lord for guiding me and helping me all the way through. I owe a special debt of gratitude Dereje Teferi (PhD), whose guidance and critical advice, as my thesis advisor, greatly contributed to the making, and letting me undertake this thesis and showing me lots of kindness whenever I visited his office.

I gratefully would like to acknowledge my deepest appreciation to my dear sister Meseret and Anchinalu Ejigu for those invaluable help through my work.

My special gratitude goes to my best friend Nirayo H. and Solomon G., who had been helping me on commenting and providing direction to my work.

Lastly I would like to express my special thanks, but certainly not least to my family, especially my mother who helped me immeasurably with her encouragement, and kept me going at the most desperate moments, and to brothers and sisters for their continuous support and encouragements throughout my studies.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Appendices

List of Abbreviations

ASCII…….American Standard Code for Information Interchange

ASR…….Automatic Speech Recognition

CELP…….Code Excited Linear Prediction

CTTS…….Concatenative Text-to-Speech

CV…….Consonant-Vowel

DSP…….Digital Signal Processing

ECSA…….Ethiopian central statistical authority

F0 …….Fundamental Frequency (pitch)

FD-PSOLA…….Frequency Domain Pitch Synchronous OverLab-Add

G2P…….Grapheme-to Phoneme

GCI…….Glottal Closure Instant

H/S…….Hybrid Harmonic/Stochastic Model

HMM…….Hidden Markov Model

HNM…….Harmonic Noise Pulse Model

IPA…….International Phonetic Association

IVR…….Interactive Voice Response

L2P…….Letter-to-Phoneme

LP…….Predictive Coding

LPC…….Linear Predictive Coding

MBR-PSOLA…….Multi-Band Re-synthesis Pitch-Synchronous OverLap-Add

MOS……Mean Opinion Score

NLP…….Natural Language Processing

OLA…….OverLap -Add

ONC…….Onset-Nucleus-Coda

OOV…….Out of Vocabulary

OR…….Onset-Rhyme

ORT……Open Rhyme Test

PCM ……Pulse code Modulation

PSOLA …….Pitch-Synchronous OverLap-Add

RMS…….Root Mean Square

SNNPR…….South Nation Nationality People Representative

SOLA…….Synchronous OverLap-Add

TD-PSOLA…….Time Domain Pitch Synchronous OverLab-Add

T2P …….Text-to-Phoneme

TTS …….Text-to-Speech

VODER…….Voice Operating Demonstrator

WSOLA…….Waveform Similarity Synchronous OverLap-Add

ZCR…….Zero Crossing Rate

ABSTRACT

The goal of Text-to-Speech synthesis is to convert arbitrary input text to intelligible and natural sounding speech so as to transmit information from a machine to a person. In speech synthesis, the capability of information extraction is crucial in producing high quality synthesized speech. This paper describes the design of a syllable based concatenative speech waveform synthesizer for Amharic language using TD-PSOLA algorithm for the prosodic modification and speech waveform analysis/synthesis purpose. This approach is based on the decomposition of the signal into overlapping frames synchronized with the pitch period.

In concatenative corpus-based TTS systems, the acoustic units of varying sizes are selected from a large speech corpus and then concatenated to produce speech waveforms. The speech corpus contains more than one instance of each unit to capture prosodic and spectral variability found in natural speech; hence the signal modifications needed on the selected units are minimized if an appropriate unit is found in the unit inventory. A syllable unit is chosen primarily because Amharic language is syllable centred; Consonant-Vowel (CV) assimilated language. The unique syllable units are then added to a syllable repository. Further, concatenation at syllable boundaries can lead to smaller error owing to the spectrum being similar across different syllable boundaries. Syllable based approach to speech processing is an interesting alternative to the diphone (triphone) - based approach, especially for the syllable-timed languages, Amharic.

The system was implemented and tested using selected Amharic texts found in the language Amharic. The result gives 97.8% of word accuracy rate for automatic syllabification, which leads to improve prosody and synthesis models as well as speech waveform generation and an average score of 89.58% and 3.45 for ORT and MOS respectively based on the subjective assessment of users' for intelligibility and naturalness of the synthesized speech respectively. Subjective listening tests performed on the synthesized speech there is an improvement of in the quality of synthesised speech.

Keywords: *Text-to-speech, concatenative synthesis, syllable, TD-PSOLA, CV-assimilated, prosodic modification, unit selection*

CHAPTER ONE

1. INTRODUCTION

1.1. Background

Language is a fundamental part of everyday life. Whether we are using speech, sign language, emotion or a coding system that conveys meaning through touch, we use language to express our thoughts, intentions, reactions, and experiences (James , 1965).

Speech provides an international forum for communication among researchers in the discipline that contribute to our understanding of the production, perception, processing, learning and using it. It is one of the most important tools for communication between human and their environment and plays a great role in man-machine interaction. The advancement of technology has provided us many tools in the area that helps in our day to day activities to be come interactive with us or machines. Speech enabled interfaces are desirable because they promise hands-free, natural, and ubiquitous access to the interacting device (Lewis, *et al.* , 2000, Lewis, 2001) . It is one of the oldest and most widely used means of communication between people. Nowadays human beings have been struggling for developing intelligent machines which handles communication automatically.

Modern speech processing techniques and research community on speech science and technology concerned about the naturalness and intelligibility of speech produced by synthesizer. The ultimate goal of speech research is to build systems that mimic human capabilities in understanding, generating and coding speech for a range of human-to-machine interactions (Jilei , 2006) .

As Jielei (2006) discusses, speech synthesis is a process of making artificial speech, understanding the language speaking rules, styles, features and produce synthetic speech. Text-to-Speech (TTS) synthesis is a process that can be applied for various applications such as services over telephone, e-document reading, and speaking system for visually impaired people.

Amharic speech synthesis can be designed by accepting normalized Amharic texts and generate acoustic and prosodic features, and then produce speech output. Normalization of Amharic texts includes transliteration, epenthesis, and gemination, and syllabification. The speech result produced by Amharic speech synthesizer needs to be natural, intelligible and pleasant same as human beings to overcome the communication using synthetic natural speech.

There are three main approaches to speech synthesis: articulatory, formant, and concatenative (Kenny, 1998). Articulatory synthesis tries to model the human articulatory system, the vocal cords, the vocal tract, but requires high computational power in practical systems. Formant synthesis employs some set of rules to synthesize speech using the formants that are the resonance frequencies of the vocal tract. Since the formants constitute the main frequencies that make sounds distinct, speech is synthesized using these estimated frequencies which make the automatic prediction of the parameters harder. On the other hand, concatenative speech synthesis is based on the idea of concatenating pre-recorded speech units to construct the utterance (i.e. the use of uttered sounds for auditory communication). Concatenative systems tend to be more natural than the other two since original speech recordings are used during synthesis of the speech waveform, such that both commercial and research systems are predominately corpus-based (Möbius, 2000 and Hasim, 2004). The first two approaches are rule based methods whereas the last is data-driven method. In the search for more natural synthesis there has been a move to use recorded human speech rather than techniques to construct it from its fundamental parts. Concatenative synthesis techniques not only give the most natural sounds speech synthesis, it also is the most accessible to the general user in that it is quite easy for us to record speech and the technique used here to analyse it are, to the most part, automatic (Shah, *et al.*, 2004). The most commonly used techniques in present systems are based on formant and concatenative synthesis. The concatenative one is becoming more and more popular since it minimizes the problems with the discontinuity effects in concatenation points (Hasim, 2004 and Lyons, 2004). In concatenative systems, speech units can be variable length units such as syllables and phones. The latter approach is known as unit selection, since a large speech corpus containing more than one instance of a unit is recorded, and variable length units are selected based on some estimated objective measure to optimize the synthetic speech quality.

Currently the state-of-the-art of speech synthesis consists of the phonetic, prosodic and synthesis module. The phonetic module converts the input texts into speech units including diacritic information such as stress marks, and stress indicators, whereas the prosodic module indicates the syllable duration, pitch and intensity. Then the synthesis units are extracted from a phonetic inventory and the concatenation rule is applied to produce final pitch and energy contour on the speech output using synthesis module.

In speech synthesis, the most commonly used techniques in present systems are based on formant and concatenative synthesis in present time. The latter one is becoming more and more popular since it minimizes the problems of discontinuity effects in the concatenation

points. The concatenative synthesis is a data-driven method which snippets of recorded speech features and store transition on database and produce high natural sounds speech.

Recent progress in speech synthesis has produced synthesizers with very high intelligibility, but the sound quality and naturalness still remain to be a major problem. In addition to this, most synthesizers currently manipulate a small number of parameters in a highly constrained manner to produce speech and thus lack flexibility. However, the quality of present products has reached to an adequate level for several applications, such as multimedia, telecommunications, and other assistive technology but speech waveforms generated is far from natural sounds (Lewis, *et al.* , 2000). Researches on text-to-speech synthesizers for Amharic languages have used synthesis techniques that require prosodic models for good quality synthetic speech but they have not reached acceptable standards.

Although there are works done by different researchers (Nadew, 2008, Tadesse, *et al.*, 2010 and Tadesse & Yoon, 2011) using different techniques, and different speech units, there is still no research work which output speech waveforms as intelligible and natural sounds except the research done by Habamu (2006), Diphone based speech synthesis system for Amharic and tries to introduced a significant improvement in quality of natural sounds when diphone speech units are used. Diphone based synthesis also requires a significant amount of prosody modelling for duration, intonation and energy which in turn requires analysis of a voluminous amount of data and deduction of proper rules from the data. Naturalness of synthetic speech produced by state-of-the-art of speech synthesis systems is mainly attributed to the use of concatenative speech synthesis (Lewis & Mark, 2001) that uses syllables as speech units. Naturalness describes how close the output sounds is to human speech, while intelligibility is how the output can be understood by end users.

The main challenges here in speech synthesis are producing natural, intelligible and pleasant sounds similar with human beings, since it requires versatile aspect of speech unit selection, appropriate techniques used, and effective pre-processing tasks. The pre-processing tasks include gemination, epenthesis vowel insertion, syllabification in the text analysis, and prosodic features like stress assignment, duration and pitch modification, labeling should handle properly. When we develop Amharic speech synthesizer main focuses should be: the speech unit selection strategies, techniques used for linguistic and prosodic analysis, and synthesis part to produce intelligible and natural speech sounds.

Therefore, our target is to building speech synthesizer for Amharic syllabified words that maximizes the naturalness and intelligibility similar with human beings by accepting normalized Amharic texts. In other words, developing a workable syllable based Amharic

TTS as communication-aid to express users' intentions and emotions for using syllables as speech unit.

1.2. Statement of the problem

In the last decades, the performance of speech processing system like speech synthesizer have improved dramatically, resulting in an increasingly widespread use of speech science and technology in real world scenario. Even though, the TTS technology is growing from time to time to make things easy, people in developing countries like Ethiopia are unable to utilize it. This is because no availability of Amharic speech synthesizer, limited access to information technology, lack of knowledge about foreign languages, and other economical and political issues to afford and use it.

It is not difficult to imagine that if an individual loses both their ability to speak and their means of expressing their feeling and emotions, which makes their isolated and depressing. Several systems have reached the test stage, but no system emotions, either vocally or even physically, due to paralysis or other factors, his or her life would is yet available on a commercial basis that aids such hardship applications for Amharic language users.

Amharic is the official language of Ethiopia; it belongs to Semitic language family that has the largest number of speakers next to Arabic language. The language uses a unique script, which has originated from ancient language, the Ge'ez alphabet and this script is phonetic in nature (Yibeltal, 2008). Therefore, this language requires linguistic and prosodic analysis so as to come up with a more natural, intelligible and pleasant sounds and to allow an easy access for different TTS applications described on section 1.6. By using TTS systems we can avail an assistive technology which handles the difficulties that arise on speech related areas especially for visually impaired users.

Many researches have been done on modelling speech synthesis for different language such as English, Dutch, Spanish, Finnish and Germany, and when we see attempts in case of Amharic language there are few published work done on modelling speech synthesizer using formant based speech synthesis technique using vowels and words as speech unit. On those works the system is evaluated using (Nadew, 2008) mean opinion score (MOS) method and score 88.85% of the vowels are correctly recognized by the listeners, and (Yibeltal, 2008) on word level but did not specify the accuracy achieved. There is project works done by (Habamu, 2006) and (Laine, 1998) on Diphone based speech synthesis system for Amharic language and achieve remarkable level. Even though, there are technological attempts on speech synthesizers for Amharic language, they are not alleviating problems related to speech naturalness and intelligibility aspects and this work tries to address issues related to

naturalness and intelligibility of synthetic speech. This is because producing natural sounds depends on the strategy of speech segments selection, the techniques used and pre-processing tasks and prosodic features modification of Amharic language on their work. Since to find speech output which is natural, intelligible and pleasant sound from synthesizer we should take attention on the pre-processing Amharic texts carefully. Pre-processing includes text analysis, linguistic analysis, and prosodic feature analysis.

Identification of syllables structures of words play an important role in speech synthesis and recognition apart from their purely linguistic significance. The pronunciation of a given phoneme tends to vary depending on its location within a syllable. While actual implementations vary, TTS systems must have, at minimum, three components: letter-to-phoneme (L2P), prosody, and synthesis module. The recent works tries to describe the syllable structure of Amharic words, for instance, (Mulugeta, 2001) and (Aster, 1981) are among those works. Syllable segments as speech unit contains basic prosodic properties, such as pitch, duration and loudness (i.e. stress, and accent).

Speech units generated by the concatenative method are syllables (i.e. composed of more than one phoneme) in linguistic sense, having a vowel nucleus with optional preceding and/or following consonants (Aster, 1981). A syllable can be described by a series of grammars. The simplest grammar is the phoneme grammar, where a syllable is tagged with the corresponding phoneme sequence. The consonant-vowel grammar describes a syllable as a CVC sequence. The syllable structure grammar divides a syllable into onset, ONC (Mulugeta, 2001, Sebsibe, *et al.* , 2004, and Nirayo, 2011). For this study the improved rule for syllabification algorithm mainly adopted from (Nirayo, 2011) is used.

The concatenative synthesizer provides high quality natural sounds; due to that concatenative synthesis itself puts together (concatenate) units selected from the voice database (i.e. stores the units features only, since the number of possible words for Amharic is virtually unlimited) and after decoding (optional), output the resulting speech signal.

The primary aim of this thesis work is to improve naturalness of speech produced by Amharic speech synthesizer by handling gemination, epenthesis (i.e. insertion of a vowel or consonant into a word to make its pronunciation easier), syllabification and modifying prosodic features based on rules and characteristics of the language, Amharic.

Whenever we develop a speech synthesizer for Amharic language words, after segment the units into equivalent syllables, we have to answer the following questions:

- How to improve the intelligibility and naturalness of speech synthesis system?

- How to select speech units suitable for speech synthesis system?
- How to identify appropriate techniques for Amharic speech synthesizer?

The research tries to answer and address the above questions and remark as a challenge to the next generation as research work, for instance stress assignment.

1.3. Objectives

1.3.1. General objectives

The general objective of this research is to develop and to investigate the possibility of having appropriate text-to-speech model for Amharic language by accepting normalized texts as an input.

1.3.2. Specific objectives

To accomplish the general objective, the specific objectives are:

- To understand related works done so far in the area of speech synthesis.
- To study the acoustic and phonetic characteristics of Amharic phonology.
- To design & model text-to-speech system for Amharic.
- To develop a prototype text-to-speech for Amharic
- To test and evaluate the performance of Amharic speech synthesizer

1.4. Research Methodology

In this study, first we study the syllable structure, linguistic rules and prosodic features for speech synthesis of Amharic words in parallel with a survey of relevant literature, a set of rules are identified, modelled and implemented.

The main issue here in speech synthesis system is that getting intelligible, natural and pleasant speech waveforms. To tackle the above problem we consider the methods which take into account the phonetic and prosodic features during representation, generation of waveforms by accepting normalized Amharic texts by focusing on the pre-processing task to find quality datasets which leads to quality output from the synthesizer.

1.4.1. Literature review

In this section various related literature in the area of TTS system had been reviewed to understand the state-of the- art of the speech synthesis system especially in the area of natural language processing and digital signal processing with different speech synthesis techniques and approaches have been reviewed. Moreover, related works will be reviewed so far on

Amharic speech synthesis system, and others languages in speech communication and technology from different resources like books, journals, and articles.

1.4.2. Data collection

To thoroughly study the speech synthesis rule of the language an interview and consultation will be made with the language experts from linguistic department. We also collect parallel corpus (speech & text), which is normalized (syllabified, epenthesis, geminated) Amharic words as representative dataset.

By measuring the acoustic intensities of sounds (waveforms), the sonority of a sound can be estimated to ideal. Since waveform synthesisers have inputs for prosody features (i.e. intensity, duration and fundamental frequency (f0) (Black and Lenzo, 2003) in relation with the classes of vowel and consonant sounds that are usually distinguished along this dimension.

In order to build a robust speech synthesizer, it is crucial that all acoustic (sub-word) models receive representative training samples, taking into account many variations can occur in speech sounds. It follows that a training corpus should be representative, which consisting of speech samples spoken by men and women from different age groups and if possible from different linguistic background to come up quality speech output from the synthesizer.

Recording sample datasets will be the major undertaking sub-tasks like selection of phonetically rich and phonetically balanced texts and sentences which is geminated and syllabified correctly. Selecting appropriate participants for evaluation purpose of the system, transcribing words into equivalent transcription (syllables in text) and recording data are become time consuming. For example, different types of speech units may be stored in the inventory of a concatenative TTS system. Storing whole speech units is impractical for general TTS because of the tremendous demands on a voice talent that would have to read a few hundreds of thousands of words in a consistent voice and manner. when we recorded sample utterances we should select the appropriate room, speech units and recording devices due to different factors which affects speech waveforms such as background noise and speakers emotions (sad, fear, happy). Even if recorded successfully in multiple sessions spread over several days and weeks, a lack of coarticulation and phonetic recoding at syllable boundaries may result in unnatural sounding speech. The recording takes place in normal emotion of speakers and listeners, normal pc speaker and at normal room.

1.4.3. Modeling Tools and techniques

Different tools and techniques are thoroughly inculcated to accomplish the proposed research. Amongst from different tools for acoustic and speech analysis software tool called PRAAT is used. PRAAT have multiple functionalities such as speech analysis/synthesis, manipulating and labeling, extracting features and segmenting, recording and listening for experiments and modifying prosodic features. The main strength of PRAAT is its powerful graphic interface which is used to label extract, and segment speech units easily and flexibly. According to the system requirements of the thesis we use Matlab program for speech waveform analysis/synthesis and prosodic modification, Microsoft visual studio (C#) software for Amharic syllabification and text to speech.

In this paper, we propose the data-driven concatenative speech synthesis technique, which tries to find high quality speech waveforms having naturalness and intelligibility as major criteria. It is an approach which uses different length pre-recorded samples sound data derived from natural speech (Engin, 2006). The data-driven concatenative synthesis is used due to its less computational complexity and need not prior knowledge about acoustic units or samples speech corpus.

Before recording the sample datasets first we will geminate, syllabify the selected Amharic texts, to come up with natural sounds, by accepting the already geminated, syllabified texts in to the system and further pre-processing is done such as: text analysis- that handles transcription of the input texts in to syllables and extraction of the speech parameters and, then engineer speech synthesis system in terms of a simple linear string of units (syllables). During recording sample datasets, we did not record and store a different version of every speech sound in every possible context; rather we recorded each speech sound in variety of different contexts including each transition.

Finally the synthesis task is done by generating the artificial speech by voice rendering from database, concatenating each segment of speech units, synthesizing it and generating natural and synthetic speech waveforms.

1.4.4. Analysis and evaluation

The most important and the commonest testing parameter used in evaluating speech synthesis systems is performance of synthesizer. During evaluation the most important qualities of a speech synthesis system are assessed with respect to naturalness and intelligibility aspects of the speech sound produced. The quality of the speech synthesizer is to maximize both characteristics.

The ultimate goal for all synthesis research with few exceptions is to produce as high speech quality as possible. The quality and the intelligibility of speech is usually a very difficult task to measure. No single measure is able to pinpoint where the problems are. To evaluate the performance of the proposed model, results will be computed using word and juncture accuracy of syllabified texts and recorded data. One of the complex problems in speech synthesis is evaluating the result due to that there is no standard method to test the performance system.

Nevertheless, the performance of the pre-processing of system is evaluated using selected speech corpus from different Amharic books and literatures, which is representative enough. In the syllabification word accuracy is simply the number of words syllabified by the method in exactly the same way as is given by the lexicon used. In addition, juncture accuracy compares speech synthesis at the sub-word level (i.e. syllables). Each position between letters is assessed to determine whether it will be classified correctly or not in to expected syllabification and other rules. The juncture between two neighbouring rhythm units is realized by intonation, stress, and pause (Aster, 1981).

Furthermore, the naturalness and intelligibility of the speech output is evaluated using MOS and ORT respectively. MOS is well known subjective scoring methods, which measure performance by providing opportunity to the listeners to rate speech quality with same set of texts as dataset and performing cumulative average of the result from the subjective assessment. It is also possible to use ORT to measure the intelligibility of the speech by assigning the listener on the selected sample data by marking there perception whether they understated the recorded and labeled sampled data correctly or not.

Finally result using two methods having different scale and criteria the selected evaluator invited to provide their perception and then taking the mean average of the users' acceptance test to evaluate the performance of the system.

1.5. Scope and limitation of the study

Syllable based approach to speech processing is an interesting alternative to the diphone or triphone based approach, especially for the syllable-timed languages like Amharic. Syllables; as the basic prosodic element, carry prosodic attributes such as pitch, duration, stress (accent) and intensity (energy) of speech.

Our main focus is to develop a speech synthesizer for Amharic texts which produce natural speech using appropriate speech selection units and techniques in addition to handling gemination, epenthesis, syllabification and modifying prosodic features. The stress, duration

and intonation modeling is beyond the scope of thesis work. But it is possible as future work, since those issues are related with syllable with variable length as speech unit.

1.6. Application of Results

The potential applications of high quality TTS Systems are indeed numerous. TTS used for enhancing the quality of speech produced by synthesizers considering the acoustic and prosodic features (Yibeltal, 2008), in addition to gemination and epenthesis. Detecting the syllable will help to model phone durations and behaves certain acoustic traits like intensity and duration to improve the synthesized speech intonation (Lewis & Mark, 2001). Advances in this area also provide benefit work in the fields of speech analysis, speech recognition and speech synthesis when dealing with natural variability. After implementing the speech synthesizer end users will be benefit from Amharic TTS system for different applications, for instance, an assistive technology for visually impaired people. By implementing such synthesizer the end user gets several potential applications; few of benefits are:

- Aid to handicapped (visually impaired persons): Voice handicaps originate in mental or motor/sensation disorders.
- Telecommunications services: TTS systems make it possible to access textual information over the telephone.
- Language education: Quality TTS synthesis can be coupled with a Computer Aided Learning system, and provide a helpful tool to learn a new language.
- Talking books and toys: The toy market has already been touched by speech synthesis.
- Vocal Monitoring: In some cases, oral information is more efficient than written messages. The appeal is stronger, while the attention may still focus on other visual sources of information. Hence the idea of incorporating speech synthesizers in measurement or control systems.
- Multimedia, man-machine communication: In the long run, the development of high quality TTS systems is a necessary step (as is the enhancement of speech recognizers) towards more complete means of communication between human and computers.
- Fundamental and applied research: TTS synthesizers possess a very peculiar feature which makes them wonderful laboratory tools for linguists: they are completely under control, so that repeated experiences provide identical and detailed results.

After modeling and implementing the TTS system for Amharic it is possible to apply any of the above applications especially for those who are visually impaired and reading disabilities people as an assistive technology.

1.7. Organization of the Thesis

This thesis is organized as the follows: Chapter 2 presents literature review and related works on human speech production system, speech synthesis. In this chapter related works on speech synthesis system on different language using different speech units, techniques tried by researchers are reviewed. In Chapter 3, we present phonology aspect of Amharic language, syllable structure, linguistic and prosodic feature of speech synthesis in Amharic and novelty of the research is discussed. Design prototype of automatic speech synthesis algorithm for Amharic is presented in Chapter 4. The experimental results and evaluation is discussed in Chapter 5. In Chapter 6, conclusion and future work directions are pointed out.

CHAPTER TWO

2. REVIEW OF LITERATURE

2.1. Fundamental of Speech Synthesis

In speech science and technology field, the major efforts have been done in language technology areas especially in speech synthesis and speech recognition. Speech is the primary means of communication among people and it has long been considered as the most natural form of human communications. The human speech production system is composed of organs ranging from the diaphragm and lung to vocal and nasal cavity (Henock, 2003). Speech synthesis can be described in simple words as a machine speaking to people. TTS technology is based on state-of-the-art algorithms and techniques that allow the creation of synthetic voices from natural recordings and maintaining the voice characteristics of the original speaker (Alessandro, 2009). As described in the first chapter, the principal objective of this study is to design and implement prototype of a generalized speech synthesizer for Amharic syllabified texts to come up with a more natural sounds.

Speech synthesis is the process of converting a written text into speech and this technology have the ability to convert arbitrary text into audible speech, with the goal of being able to provide textual information to people via voice messages (Yibeltal, 2008). In addition, it is an artificial production of human speech from raw text and used extensively to aid people with disabilities. Tadesse (2009) sited lemmetty (1999) and discussed, TTS synthesis is a process that artificially produces synthetic speech for various applications such as human-machine interaction, hands and eyes free access of information, interactive voice response systems (IVR), and screen reader software for the visually impaired.

It is also called speech prosthesis is computer-generated speech for people with physical disabilities that make it difficult to speak intelligibly and/or multimodal speech synthesis which incorporate an animated face synchronized to complement the synthesized speech (Alessandro, 2009). Researchers in different backgrounds collaborate to put together their knowledge in computational linguistics, phonetics, prosody, physiology, vocal tract modeling, signal processing, image synthesis, experimental psychology. At the turn of our century, the natural language processing and digital signal processing allows synthetic speech to be widely used.

Speech synthesis consists of text (i.e. normalization), phonetic (i.e. G2P) and prosodic (i.e. stress and intonation) analysis, and speech waveform synthesis steps are shown in Figure 2.1 below. Mainly classified as text analysis, where the input text is transcribed into a phonetic or

some other linguistic representation; this is the most sophisticated part of the job and generation of speech waveforms, where the acoustic output is produced from this phonetic and prosodic information.

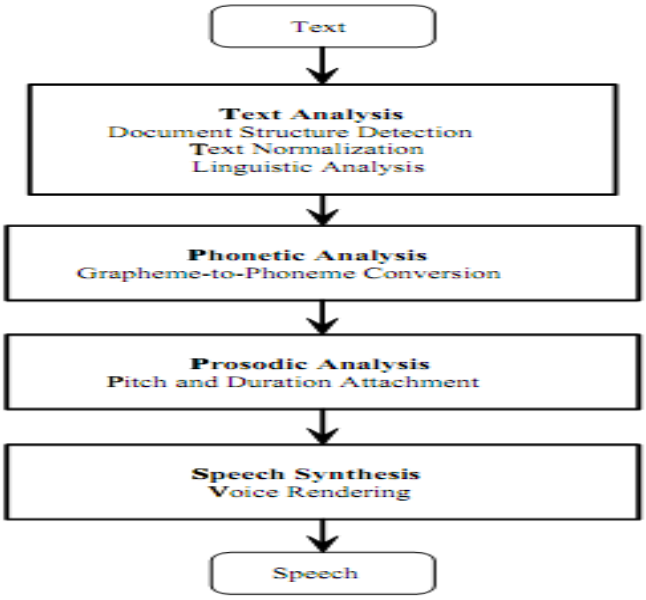


Figure2. 1: Steps of TTS System, source:(Morais & Violaro, 2005)

The ultimate goal of TTS system is to create applications which listeners, and users in general, can't easily determine whether the speech comes as sources from a human or a synthesizer (Zen, *et al.* , 2009), to convert ordinary orthographic text into an acoustic signal that is indistinguishable from human speech. In general, the speech synthesizer depends on the TTS system architecture inculcated to produce intelligible and natural sounds from the synthesizer.

2.2. Historical Background

Although speech synthesis has had a long history (Rashad & Mastorakis, 2003), progress is still being made and recent attention in the field has been primarily focused on concatenating real human speech selected from a large corpus and produce natural sounds. Producing artificial speech using speech synthesizer has been a dream of the humankind for centuries. The earliest efforts to produce synthetic speech were made over two hundred years ago (Lemmetty, 1999), as mechanical and electrical stage of development. Under the mechanical state of development in 1779, Russian Professor Christian Kratzenstein explained physiological differences between five long vowels *(/a/, /e/, /i/, /o/, and /u/)* and made apparatus to produce them artificially. A few years later, in Vienna in 1791, Wolfgang von Kempelen introduced his "Acoustic- Mechanical Speech Machine"(see Figure 2.2), which was able to produce single sounds and some sound combinations.

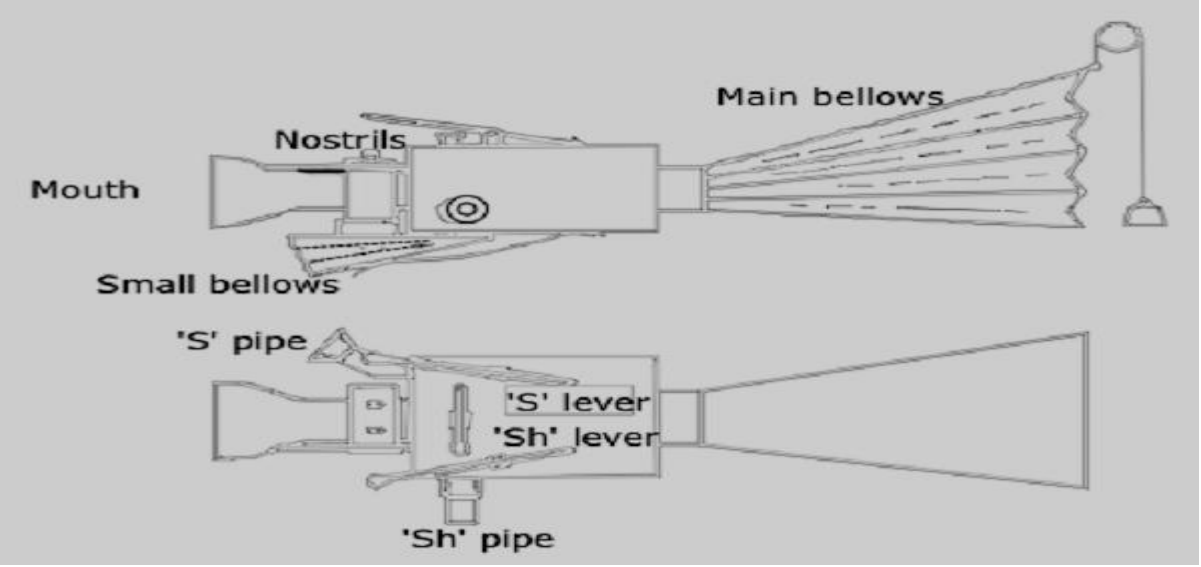


Figure2. 2:von Kempelen‘s talking machine , Source:(Mons and Fitzpatrick, 1997)

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 able to produce vowels and most of the consonant sounds. The Stewart’s machine was able to generate single static vowel sounds with two lowest formants, but not any consonants or connected utterances. In late 1800’s, Alexander Graham Bell with his father, inspired by Whetstone’s speaking machine, constructed same kind of speaking machine. He put his dog between his legs and made it growl- to utter low dull rumbling sounds, then he modified vocal tract by hands to produce speech like –sounds.

In the electronic stage of development of speech technology, the first full electrical synthesis device introduced by Stewart in 1922, which consists of buzzers and two resonant circuits for the first 2 formants and in 1923, he added the third formant into the system. First device to be considered as a speech synthesizer was called VODER (Voice Operating Demonstrator), introduced by Homer Dudley in New York World's Fair 1939, works by operators using keyboard and pedal shown in Figure 2.3 below.

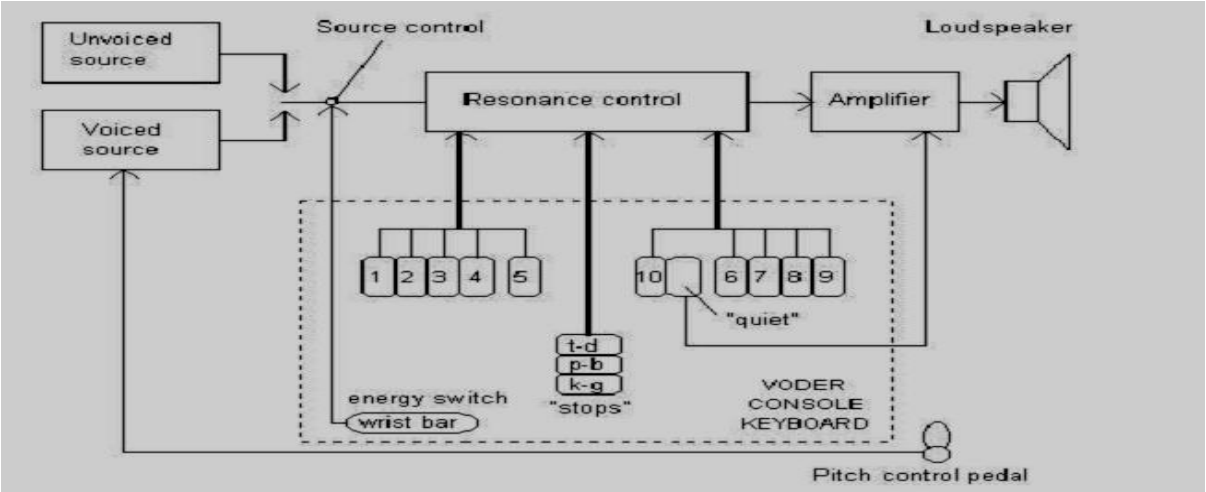


Figure2. 3: Block scheme for VODER, source: (Kuhn, 2009)

As different researchers discusses about the development of speech synthesis system, some of the milestones are shown in Figure 2.4 below. The integration of speech science and technology on speech processing are mainly classified into mechanical and electrical stage of development for scientific research in the world became more and more interested and providing remarkable results throughout the world.

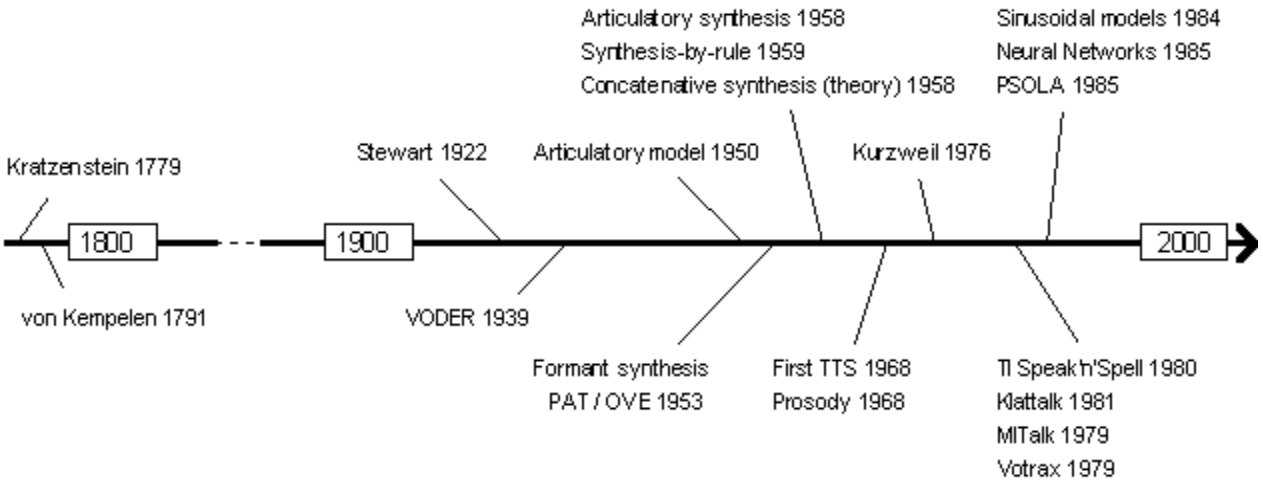


Figure2. 4: Milestones of Speech synthesis development, source:(Lemmetty, 1999)

Since the mid-1990s, significant advances have been made in the area of TTS synthesis which enabled a broad class of new applications and services in telecommunications, entertainment, language education, and other areas. At the same time, research efforts have shifted from providing more intelligible and natural synthetic sounds (Pijper & Krahmer, 2000). The enablers for these advances are, among others, data-driven methods and larger speech segment units (i.e. syllables and words) that are now being applied in all aspects of TTS system and technology.

Speech synthesis, automatic generation of speech waveforms, has been under development for several decades (Raj, *et al.* , 2007). Recent progress in speech synthesis has produced synthesizer with high intelligibility but the sound quality and naturalness perspective still remain a major problem. However, the quality of present products has reached an adequate level for several applications, such as multimedia and telecommunications. Each attempts in the past put a great role to the development of speech science and technology, especially in speech synthesis as commercial purpose for different real applications.

Modern speech synthesis technologies involve quite complicated and sophisticated methods and algorithms to synthesize and find natural human like sounds from the synthesizer. The critical issues for current speech synthesizers concern trade-offs among the conflicting demands of maximizing speech quality, while minimizing memory space, algorithmic complexity, and computational speed and cost (O'Shaughnessy , 2001).

2.3. Articulatory Phonetics

2.3.1. Human Speech Production System

Speech is produced by air-pressure waves emanating from the mouth and the nostrils of a speaker. The gross components of the speech production apparatus are the lungs, trachea, larynx (organ of voice production), pharyngeal cavity (throat), oral and nasal cavity. The pharyngeal and oral cavities are typically referred to as the vocal tract, and the nasal cavity as the nasal tract (Huang, 2001). Figure 2.5 shows the important articulatory motions and models of the human speech production system.

Rong-Wei (2003), quoting Huang (2001), and mentioned that the lung is the source of the air during speech production. 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. The larynx is the structure that holds and manipulates the vocal cords. The "Adam' s apple" in males is the bump formed by the front part of the larynx. The place where the vocal folds come together called the glottis. The epiglottis is the fold of tissue below the root of the tongue. The epiglottis helps to cover the larynx during swallowing; food goes into the stomach and not the lungs. A few languages use the epiglottis in making sounds. For instance, a short distance behind the upper teeth is a change in the angle of the roof of the mouth. This is the alveolar ridge. Sounds, which involve the area between the upper teeth and this ridge called alveolars (i.e. a consonant articulated with the tip of the tongue near the gum ridge). The hard portion of the roof of the mouth called hard palate that the term "palate" by itself usually refers to the hard palate and the Velum (Soft Palate) operates as a valve and

it is the soft portion of the roof of the mouth, lying behind the hard palate called soft palate that separates the oral and nasal cavities. The velum can also move: if it lowers, it creates an opening that allows air to flow out through the nose; if it stays raised, the opening is blocked, and no air can flow through the nose.

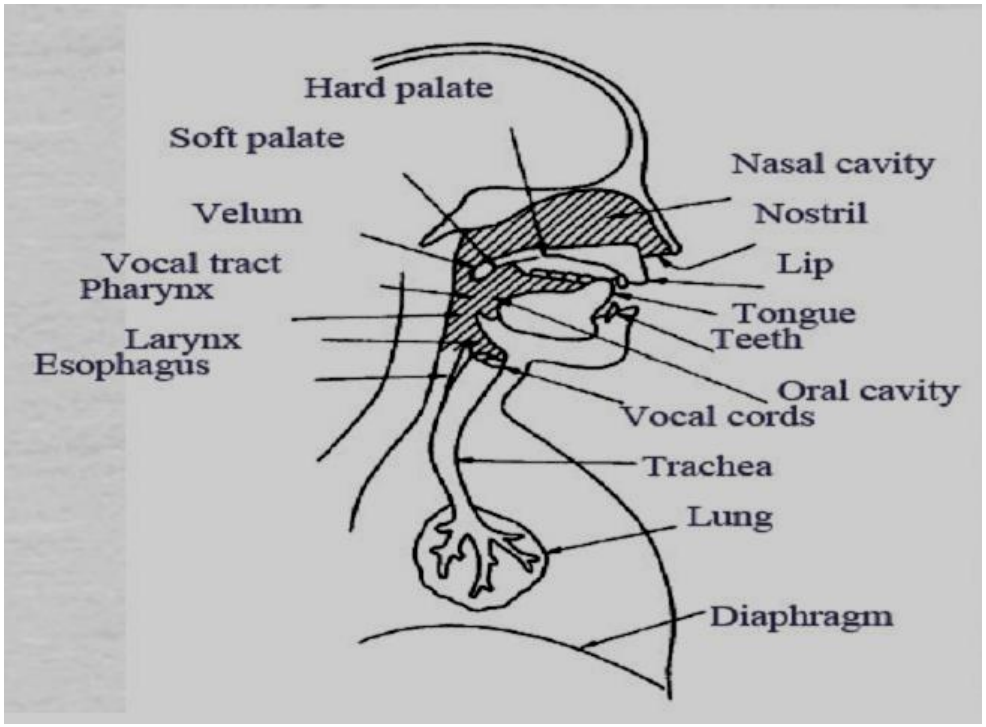


Figure2. 5: Human speech production system, source: (Huang, 2001).

The tongue is among the finer anatomical features critical to speech production. The dorsum is the main part of the tongue, lying below the hard and soft palate. It is found at the back part of the tongue (hence "dorsum", or "back" for Latin) and shaped away from the palate for vowels, placed close to or on the palate or other hard surfaces for consonant articulation. The teeth are also another place of articulation used to brace the tongue consonants (Morais & Violaro, 2005). For example, plosives are considered as the most basic type of consonant, which are produced by stopping the flow of air at some point and suddenly releasing it (see Figure 3.2). They form a complete obstruction to the flow of air out of the mouth and nose, and normally these results in a build-up of compressed air inside the chamber formed by the closure. When the closure is released, there is a small explosion that causes a sharp noise. The basic plosive consonant type can be exploited in many different ways: plosives may have any place of articulation, may be voiced or voiceless and may have an egressive or ingressive airflow. The airflow may be from the lungs (pulmonic), from the larynx (glottalic) or generated in the mouth (velaric).

As mentioned above, the vocal cords play a great role in generating the kind of sound to be produced. The vowel are grouped in voiced sounds and the consonants are unvoiced sounds. Consonants involve constrictions, or gestures that narrow the vocal tract at a particular point. When we classify consonants, one of the most important things to consider is the place where this obstruction is made; this is known as the place of articulation, and in conventional phonetic classification, each place of articulation has an adjective that can be applied to a consonant. For example in Amharic language, as shown in Figure 3.2, there are five place of articulation as: labial, dental, palatal, velar, and glottal (Tadesse, 2009).

The other important thing that we need to know about the human speech production system is what sort of obstruction it makes to the flow of air. A vowel makes very little obstruction, while a plosive consonant makes a total obstruction. The type of obstruction the phonemes make is known as the manner of articulation. In Amharic, the common manners of articulations are stops, fricative, nasals, liquids and affricatives (Daniel, 2006). In general the human speech production, the model and parameter control (see Figure 2.6), which are employed to generate speech waveforms.

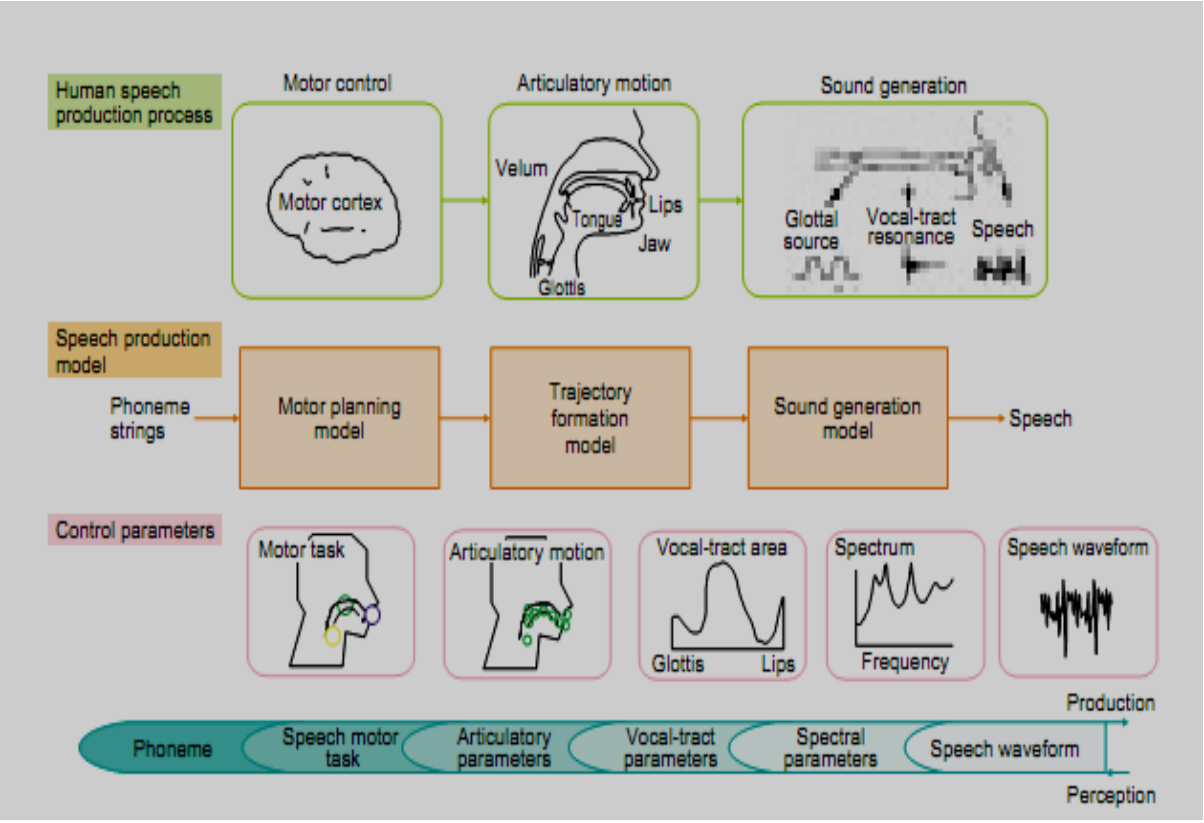


Figure2. 6: Speech production process and the mode, source: (Honda, 2003)

Basically the human speech production process targeted to produce speech waveforms from different articulatory and control parameters, and acoustic units as shown in Figure 2.6 above.

Tesfaye (2004), quoting Elker (2002) and discuses that the vocal tract is bound by hard and soft tissue structure. The structures are either essentially immobile or mobile. The mobile structures associated with high speech production are also referred to as articulators (i.e. jaw, tounge, lips and mouth). Movement of these articulators appeared to account for most of variations in vocal tract shape associated with speaking style. Modelling and controlling the segmental coarticulation and other phonetic factors is an important part of a TTS system.

2.4. The Basics components of Text-To-Speech System

In TTS systems, the process of converting written text into speech contains a number of components and processes are employed. In general, the TTS system architecture is classified into two basic components: the Natural Language Processing (NLP) and the Digital Signal Processing (DSP) as shown in Figure 2.7. The NLP is capable of producing a phonetic transcription of the text read, together with the desired intonation and rhythm (often termed as prosody), whereas the DSP, transforms the symbolic information it receives form NLP to produce speech waveform. Each of the above components contains different subtasks discussed in the next section below.

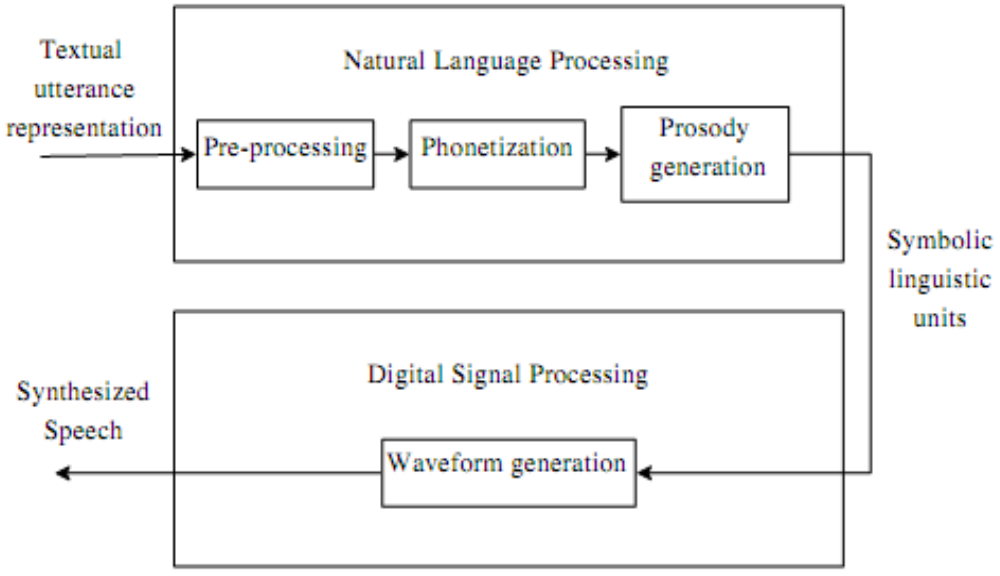


Figure2. 7: Basic components of TTS synthesizer, source:(Shah , 2004)

2.4.1.1. The Natural Language Processing (NLP) Component

Natural Language Processing or text-to-phoneme (T2P) is targeted to produce phonetic transcription of the text, together with the desired prosodic features. It concern how computational methods can aid the understanding of human language and focused on

developing systems that allow computers to communicate with people using everyday language (Dutoit, 2008). The NLP component consists of three processing stages as seen in Figure 2.8. The components are text analysis, automatic phonetization and prosody generation.

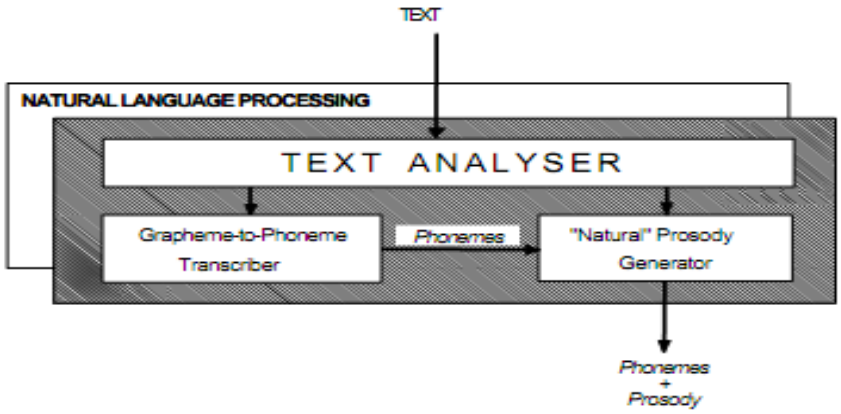


Figure2. 8:The skeleton of the NLP module, source: (NLPA-Phon2, 2007)

2.4.1.1.1. Text Analysis

The first stage of NLP is text analysis, which consists of four modules: pre-processing module (text normalization), morphological analysis module, contextual analysis module and syntactic prosodic parser. Each of these modules will be discussed in the following section.

2.4.1.1.2. Text normalization

The first task of all TTS systems is to pre-process or normalize the input text in a variety of ways. Text normalization includes tokenization, morphological and contextual analysis, and syntactic-prosodic parser.

In a pre-processing module, the input texts are organized into lists of words. The system must first identify these words or tokens in order to find their pronunciations. In other words, the first step in the text analysis is to make chunks out of the input text - tokenizing it. Many tokens in a text appear in a way where their pronunciation has no obvious relationship with their appearance such as abbreviations, acronyms and numbers. Apart from tokenization, normalization is needed where a transformation of these tokens into full text is done.

In a morphological analysis module, all possible part-of-speech categories for each word are proposed on the basis of their spelling. For example inflected, derived and compound words are decomposed into their morphs by simple grammar rules.

A contextual analysis module considers word in their contexts. This is important to be able to reduce possible part-of-speech categories of the word by simple regular grammars by using lexicons of stems and affixes.

Finally, in syntactic-prosodic parser, the remaining search space is examined and the text structure is found. The parser organizes the text into clause and phrase like constituents. After that the parser tries to relate these into their expected prosodic realization (Dutoit, 2008).

2.4.1.1.3. **Phonetization**

The second module is the grapheme to phoneme (G2P), where the words are phonetically transcribed. In this stage, the module also maps sequences of grapheme into sequences of phoneme with possible diacritic information, such as stress and other prosodic features that are important to fluency in naturally sounds speech. It is responsible for the automatic determination of the phonetic transcription of the incoming text into correct pronunciation.

In Dutoit (2008), pronunciation dictionaries refer to word roots only. They do not explicitly account for morphological variations (i.e. plural, feminine, conjugations, especially for highly inflected languages, such as French, Amharic), which have to be dealt with by a specific component of phonology, called morphphonology. Some words actually correspond to several entries in the dictionary, or more generally to several morphological analyses, and pronunciations.

Pronunciation dictionaries merely provide something that is closer to a phonemic transcription than from a phonetic one (i.e. they refer to phonemes rather than to phones). Words embedded into sentences are not pronounced as if they were isolated. Surprisingly enough, the difference does not only originate in variations at word and syllable boundaries but also on alternations based on the organization of the sentence into non-lexical units, that is whether into groups of words or into non-lexical parts of many phonological processes. For instance, Amharic language is sensitive to syllable structure since it is syllable timed languages; the words are assimilated from CV-syllables templates. In addition, not all words can be found in a phonetic dictionary, called out of vocabulary (OOV)-the pronunciation of new loan words and proper names. To minimize the effects for this research syllables with variable unit length are selected.

2.4.1.1.4. Prosody Generator

The last module, the prosody generator, is where certain properties of the speech signal such as pitch, loudness and syllable length are processed. It refers to certain properties of the speech signal, which are related to audible changes in pitch, loudness, and syllable length.

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 (Dutoit, 1997). Prosody is the pitch, speed, and volume that syllables, words, phrases, and sentences are spoken with. The technique that uses to synthesize prosody varies, but there are some general techniques, which are used to identify the correct prosodic features (i.e. intonation, stress, duration). Prosody analysis relies on each level of linguistic competence of the reader: syntax, semantics and pragmatics relations.

Phonological processor takes phonemic stream as input and marks syllable boundaries, stress, and rhythm; ultimately generates phonetic stream after applying phonological rules. Phonological processor constitutes the following modules: syllable marker, stress marker, intonation marker, and phonological rule processor. The prosodic features of speech depends on a number of factors such ass the meaning of the sentences (group of words), the speaker characteristics and emotions. In this work ,the major concern is to see the prosodic feature of Amharic texts clearly to identify, and clarify the texts whether it is correctly pronounced, transcribed, geminated, epenthesized and finally syllabified or not. Even though, generating the appropriative prosody feature depends on such factors are identified in the Figure 2.9 below.

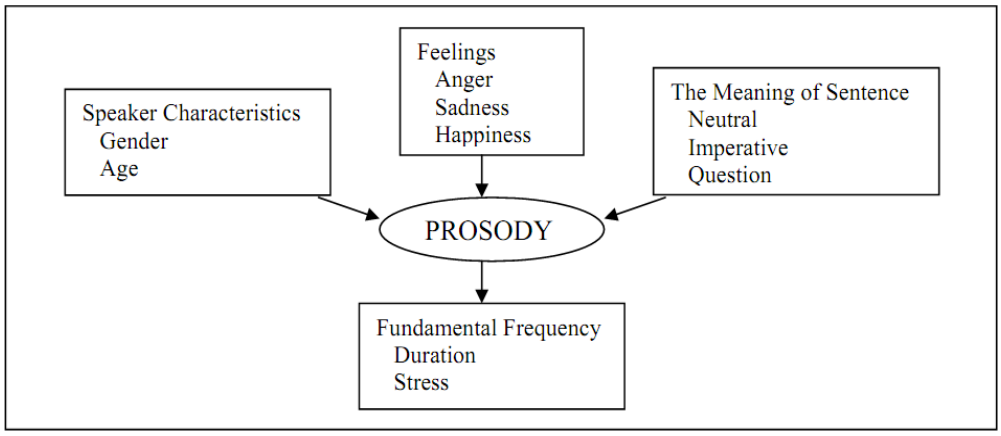


Figure2. 9: Prosodic Dependencies (Lemetty, 1999)

Prosodic features can be divided into several levels such as syllable, word, and phrase level. For example, at word level vowels are more intense than consonants. At phrase level correct prosody is more difficult to produce than at the word level. In case of syllable level the prosodic feature is highly related that it is easy to detect the attributes of the prosody. Since the most obvious prosodic feature in language is the syllable, which also known as suprasegmental features. The main features of the prosody are described below in brief.

Pitch**:** The pitch pattern or fundamental frequency over a sentence (intonation) in natural speech is a combination of many factors. The pitch contour depends on the meaning of the

sentence. For example, in normal speech the pitch slightly decreases towards the end of the sentence and when the sentence is in a question form, the pitch pattern will raise to the end of sentence. In the end of sentence, there may also be a continuous rise which indicates that there is more speech to come. A raise or fall in fundamental frequency can also indicate a stressed syllable (Klatt, 1987). Finally, the pitch contour is also affected by gender, physical and emotional state, and attitude of the speaker.

Duration**:** The duration or time characteristics can also be investigated at several levels from phoneme (segmental) durations to sentence level timing, speaking rate, and rhythm. The segmental duration is determined by a set of rules to determine correct timing. Usually, some inherent duration for phoneme is modified by rules between maximum and minimum durations. In general, the phoneme duration differs due to neighboring phonemes. At sentence level, the speech rate, rhythm, and correct placing of pauses for correct phrase boundaries are important.

Intensity: The intensity pattern is perceived as a loudness of speech over the time. At syllable level, vowels are usually more intense than consonants and at a phrase level, syllables at the end of an utterance can become weaker in intensity. The intensity pattern in speech is highly related with fundamental frequency. The intensity of a voiced sound goes up in proportion to fundamental frequency (Klatt, 1987).

2.4.1.2. The Digital Signal Processing (DSP) Component

The DSP also called phoneme-to-speech (P2S) that transforms the symbolic information it receives from the NLP module into speech signal. Intuitively, the operations involved in the DSP component 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 (Dutoit, 2008). In order to do it properly, the DSP component should obviously, in some way, take articulatory constraints into account, since it has been known for a long time that phonetic transitions are more important than stable states for the understanding of speech (Liberman, 1992).

Computers made it possible to utilize speech synthesis for practical purposes, and several systems with the function of converting text to speech were developed. TTS systems perform a range of processes, from text normalization, pronunciation, and several aspects on symbolic and acoustic prosody, finally generating speech at the last step (Jilei , 2006). Concatenative unit selection method provides the greatest naturalness, because it applies only small amounts of DSP to the recorded speech corpus in a flexible manner. The DSP often makes recorded

speech sound less natural, consequently TTS systems use a small amount of signal processing at the point of concatenation to smooth the waveform and to minimize degradation effects.

According to Jilei (2006), DSP component can be achieved in two ways:

- Explicitly, in the form of a series of rules which formally describe the influence of phonemes on one another;
- Implicitly, by storing examples of phonetic transitions and co-articulations into a speech segment database, and using them just as they are, as ultimate acoustic units i.e. in place of phonemes.

2.5. Speech Synthesis Techniques

Synthesized speech can be produced by employing several different techniques to find natural human like sounds. The major purposes of speech synthesis techniques are to convert a chain of phonetic symbols into artificial speech, to transform a given linguistic representation and to generate speech automatically with information about intonation and stress i.e. prosody (Dutoit, 1997). Birkholz(2007), discussed that the techniques used for speech generation, speech synthesis can be classified into two types: rule-based and data-driven methods.

Rule-based synthesizers are mostly in favour with phoneticians and phonologists, as they constitute a cognitive, generative approach of the phonation mechanism. The broad spreading of the Klatt synthesizer (Klatt, 1987) for instance, is principally due to its invaluable assistance in the study of the characteristics of natural speech and synthesized speech. 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, and syllables. Figure 2.10 shows the comparison of the internal data structure of dictionary and rule-based synthesis methods.

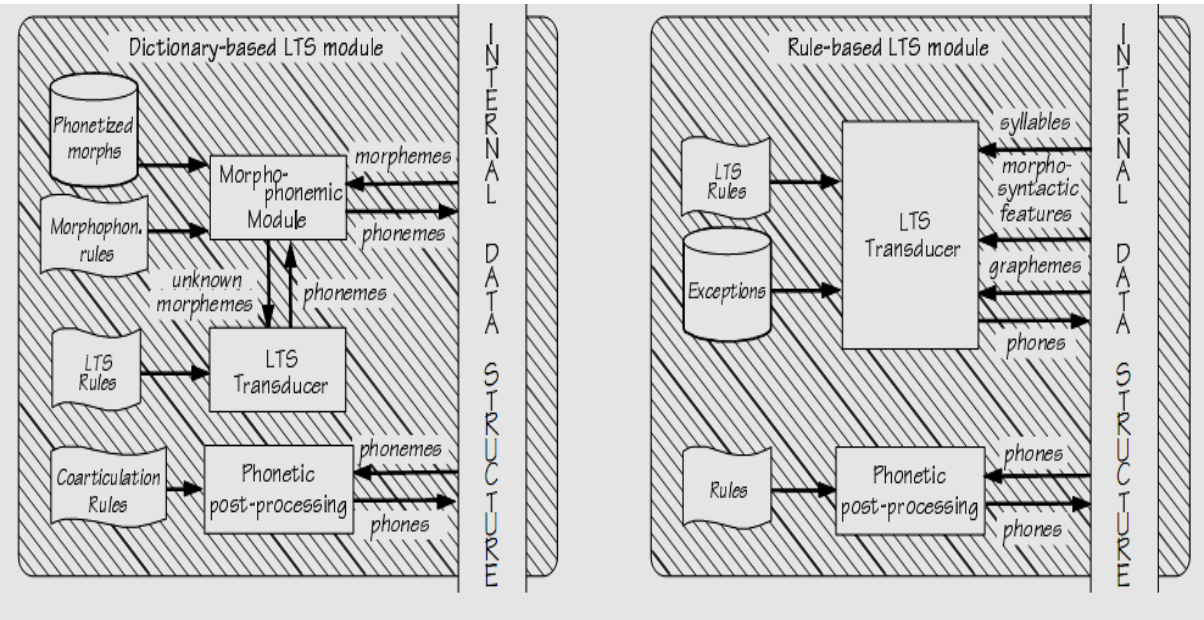


Figure2. 10: Rule and Dictionary-based synthesis method, Source: (Monaghan and Keynes, 2001)

As discussed in the first chapter, speech synthesis techniques are categorized into: articulatory synthesis, attempts to model the human speech production system directly; formant synthesis, models the pole frequencies of speech signal or transfer function of vocal tract based on source-filter-model and concatenative synthesis uses different length pre-recorded samples derived from natural speech. The Articulatory and formant synthesis have traditionally used synthesis by rule-based whereas concatenative synthesis belongs to the data-driven/ dictionary-based category (Minghui, 2000).

2.5.1. Articulatory Synthesis

Articulatory synthesis simulates speech production, which is the most widely accepted method of synthesizing speech for many years, which model the human speech production system directly. It requires a dynamic model of vocal tract that makes it possible to simulate motions of the articulators and glottis model, generates the excitation signal during speech production. The synthesis requires also a method of generating and monitoring air pressure and velocity. Mimicry of speech is achieved by dynamically changing the virtual shape and sizes of these segments according to the corresponding articulatory movement.

In the source-filter model of speech production, the excitation signal is considered as the source and the signal passes through the vocal tract and nasal tract, which acts as a filter. Their location determines the identity of a sound produced. Articulatory synthesis is very attractive for research on speech production and perception, but the quality of speech generated in this way is far from perfect due to its computational and mathematical

complexity and insufficient knowledge concerning each of the articulatory parts and processes involved in the production of speech (Alan, *et al.* , 2007).

2.5.2. Formant Synthesis

Formant synthesis is based on the source-filter-model and knowledge concerning speech acoustics representation. This method uses a set of rules for controlling a highly simplified source-filter model that assumes the source (glottal) is completely independent from the filter (Nadew, 2008). It consists of two components: the generator of an excitation signal as source and formant filters that represent the resonances of the vocal tract as filter. Formant synthesis seeks to mimic human speech by artificially creating the movements of these frequencies and bandwidths are modelled by means of a two-pole resonator.

Researchers, Nadew (2008) and Yibeltal (2008), discussed that diphone based formant synthesis is very useful for the research in the field of speech acoustics, phonetics and speech perception. Speech generated by formant synthesizers is characterized by metallic sounds. However, this sort of speech synthesis can bring satisfying results if control parameters are hand-tuned, which is impossible in a fully automatic system, which needs high intensive knowledge for each of the acoustic representation. For example, few of formant synthesizers used in nowadays are: KlattTalk, MITalk, DECtalk, and Infovox (Klatt, 1987).

2.5.3. Concatenative Synthesis

In concatenative speech synthesis, there is always a unit inventory that stores pre-recorded speech segments. During synthesis, suitable segments are selected and concatenated with or without signal processing. State-of-the-art commercial TTS systems produce speech by concatenating sound units with variable length. These sound units are drawn from a carefully designed database that has units in various prosodic and phonetic contexts. To concatenate those units, it is necessary to select the best synthesis method, i.e. concatenative synthesis which produces natural speech. It is based on the concatenation of speech segments, which gained much attention recently in speech science and technology areas. Connecting pre-recorded natural utterances is probably the easiest way to produce intelligible and most natural sounds synthetic speech. This synthesis method preserve the coarticulation effects and prosody of the language, using digital signal processing techniques to alter parameters like pitch, and duration and to smooth the discontinuity created by concatenation points of speech segments (Dutoit, 1997).

The quality of the synthetic speech produced by a corpus-based synthesizer depends largely upon the suitability and quality of the speech acoustic unit to represent the variability of the

language within the target application domain and that make the unit selection very important.

One of the most important aspects in concatenative synthesis is to find the correct variable unit length, to minimize the diphone and domain specific problems. The selection of unit size is usually a trade-off between longer and shorter units. For instance, in order to obtain natural prosody and smooth concatenation in synthetic speech, for each base unit, rich prosodic and phonetic variations are often expected. This is easy to achieve when smaller base units are used. However, smaller units mean more units per utterance and more instances per unit, and this implies a larger search space for unit selection and thus a longer search time. Besides, smaller units do cause more difficulties in precise unit segmentation. It is found that longer base units are useful as long as enough instances are guaranteed to appear in the database.

In addition, with longer units' high naturalness, less concatenation points and good control of coarticulation are achieved, but the amount of units and memory required is increased. With shorter units, less memory is needed, but the sample collecting and labeling procedures become more difficult and complex. Present concatenative speech synthesis systems may use a wide variety of segments: allophones, diphones, triphones, half-syllables, demi-syllables, syllabic segments and some other types. It depends crucially on the adequacy of the phonological analysis underlying its unit list. In the same way as intelligibility requires concatenative units to be based on a consistent minimal set of allophones, quality requires enriching the allophone inventory.

Henock (2003) and Habtamu (2008) tried their work on concatenative approach, diphone speech unit for Amharic TTS. Researchers addressed that the unit size selection, techniques used, and hardware and software costs required as major problems. During their work they identified that word is the most natural unit for written text and some messaging systems with very limited vocabulary. Concatenation of words is relative easy to perform and less coarticulation effects are captured. However, there is a great difference with words spoken in isolation and in continuous sentence, which makes the continuous speech to sound very unnatural (Raj, *et al.* , 2007). For example, intonation will be lost and because there are hundreds of thousands of different words and proper names in each language, word is not a suitable unit for any kind of unrestricted TTS system. Words need to be divided into syllables, partly because some phonological rules are affected by syllable boundaries, stresses and accents (NLPA-Phon2, 2007), diphones, triphones, and phonemes. Larger units produce quality speech than smaller units (Fék, *et al.*, 2006).

Syllables are groups of phonemes smaller than words in size, which preserve coarticulation effects like words do. 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. Amharic words have an average of three up to four syllables in each word, for example, the word አባላት -"*abalat*" means "member" contains three syllables as /*a*/, /*ba*/and /*lat*/ and ሃብታም- "*habtam*" means "rich" have two syllables, /*hab*/and /*tam*/.

Phonemes are probably the most commonly used units in speech synthesis because they are the normal linguistic representation of speech. Using phonemes give maximum flexibility with rule-based systems. However, as Donovan (1996) discussed synthesizing speech by using phonemes is difficult due to large amount of contextual variations on the phonemes itself and high degree of concatenation discontinuity.

Diphones are extends from the central steady state part of one phone to the central part of the next phone, which contains the transitions between adjacent phones. That means the concatenation point will be in the most steady state region of the signal that reduces the distortion from concatenation points (Dutoit, 2008).

In the context of Amharic languages, the basic units of writing system are characters, which are an orthographic representation of speech sounds. A character in Amharic language scripts is close to the syllable and can be typically of the following form: V, CV, VC, CCV, CCCV, and CCVC, where C is consonant and V is Vowel (Tadesse, 2010, Nirayo, 2011). Amharic syllable structure contains at least one vowel and zero, one or more consonants, that shows that Amharic language depends highly on syllables. Syllable as a speech unit for Amharic is to lesser the concatenations cost and reduces the co-articulation effects during synthesis. TTS synthesis based on syllables seems to be a good possibility to enhance the quality of synthesized speech with comparison to diphone based synthesizers (Hunt & Black, 1996). The syllable-based approach has to face the problem with a relatively large inventory of the syllables and we cannot cover all the syllables of a language in a speech database. In order to address the coverage of the syllables, we have hypothesized that approximate matching of the syllable could be used for TTS synthesis (Libossek & Schiel, 2000).

A syllable is a unit of sound composed of a central peak of sonority (usually a vowel), and the consonants that cluster around this central peak. As we have seen in chapter one, syllabification has importance in a variety of speech applications. For instance, in speech synthesis, syllables are important in predicting prosodic factors like accent and intonation. The realization of a phone is also dependent on its position in the syllable (onset is pronounced differently than coda). In speech recognition syllabification has been used to

build recognizers which represent pronunciations in terms of syllables rather than phonemes (Jurafsky & Martin, 2006), it also helps to detect out of vocabulary words (OOV).

From phonological standpoint syllable is a conventional unit, which is a group of sounds that constitute the smallest unit of the rhythm of a language. These phonological syllables differ from language to language. In English, for example, it is theoretically possible to make a single syllable as CCCVCCCC, where previous studies related to the syllable structure of the standard Amharic have shown that the following syllable templates: V, VC, VCC, CV, CVC and CVCC occur as part of the phonological system of Amharic (Aster, 1981, Mulugeta, 2001, Henock, 2003 and Nirayo, 2011). Technically, the basic elements of the syllable are the onset (zero or more consonants) and the rhyme (similar in sound, especially with respect to the last syllable). The rhyme (sometimes written as 'rime') consists of a vowel, which is treated as the nucleus, following any of consonant(s), described as the coda (final).

According to Weber (2005), when a syllabification procedure is included as a component of a TTS system, a data-driven method is a more appropriate choice than a rule-based approach, even for languages with low syllabic complexity. Taking syllabification during preparation of speech database leads the more natural and intelligible speech sounds.

Different researches have been tried on local language especially for Amharic. These include, Sebsibe (2004), on proper unit selection and optimal corpus preparation, Tadesse (2009), on gemination duration modelling and Nirayo (2011), on syllabification procedures, are issues which must be address to find more natural and intelligible sound, especially for TTS and automatic speech recognition (ASR) implementation. In this thesis works, the researcher tries to find a generalized Amharic speech synthesizer by accepts normalized and syllabified Amharic texts. Normalization includes transliteration from one alphabet to another, gemination-the doubling of consonant(s), epenthesis-the insertion of a vowel into a word to make its pronunciation easier, and syllabification-forming or dividing words into syllables and stress assignment.

2.6. Concatenative syllable based Speech Synthesis Algorithms

Concatenative synthesis techniques work by "gluing" together speech chunks that have been previously recorded in time domain analysis. It concerns the generation of speech from an input text. The speech results from concatenation of acoustic units stored in a database and units are annotated for features referring to linguistic structure and suprasegmental context in which they occur. During synthesis, the database is searched for units that match a target utterance and features of the target utterance are determined on the basis of linguistic and

contextual analyses which are carried out by NLP and DSP module in TTS system (Alan, *et al*., 2007).

The basic idea behind concatenative speech synthesis is that speech can be generated from a limited inventory of acoustic units and that their concatenation should account for coarticulation. Therefore, systems based on phonemes and words are impractical – they do not account for coarticulation and as a result, unnatural synthetic speech is obtained. The selection of units is based on two costs– concatenation and target costs. The former one specifies the amount of discontinuity at the concatenation point between two acoustic units, whereas the latter determines to what extent the unit matches the target utterance specifications. They use inventories of prosodic units that correspond to syllables (Yamagishi, *et al*., 2007 and Tadesse, 2010) and carry out a joint selection of the segmental and prosodic units.

Concatenative syllable based speech synthesis systems render speech by concatenating pre-recorded speech units (syllables) inventory is seen in Figure 2.11. Corpus-based methods use a large inventory to select the units and concatenate (Lewis, *et al*. , 2000) and also called unit selection, has emerged as a promising methodology to solve the problems with the fixed-size unit inventory synthesis, e.g., diphone synthesis (Hunt & Black, 1996). Using a fixed-size unit inventory requires making unit concatenations at each unit join; as a result, the output speech quality is degraded. Moreover, the prosodic modification of the each unit is also expected, since limited number of units exists in the speech unit inventory. These signal modifications further degrade speech quality and result unnatural synthetic speech.

It is characterized by storing, selecting, and smoothly concatenating pre-recorded human utterances (phonemes, syllables, or longer units) (Obin, *et al.*, 2009). In this approach, to prepare speech database, the small pieces are either cut from the recordings or recorded directly and then stored. Then, at the synthesis phase, units selected from the speech database are concatenated and, the resulting speech signal is synthesized as output. Figure 2.11 shows block diagram of the concatenative TTS system. However, differences between natural variations in speech and the automatic segmentation of the waveforms can cause audible glitches in the output and proper union of speech units to affect an intelligible and natural-sounds synthetic speech produced. For that purpose, large speech units need to be processed and concatenated so that discontinuities at concatenation points are minimized.

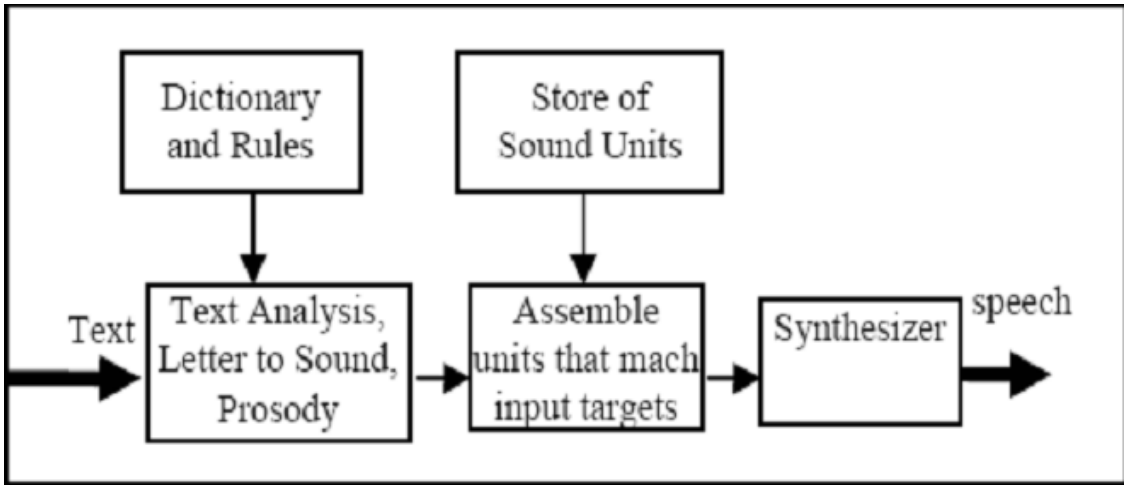


Figure2. 11: Block Diagram of the Concatenative TTS System (Shah, *et al*., 2004)

There are three main sub-types of concatenative synthesis: Unit selection synthesis, Diphone synthesis and Domain-specific synthesis. Form variant types of concatenative synthesis, the unit selection are one of the concatenative approaches which apply only a small amount of digital signal processing for synthesis of the recorded speech providing the greatest naturalness (Zhang, 2004)

Another concatenative approach is called the diphone synthesis that uses a minimal speech database containing all possible diphones (sound-to-sound transitions) in a language. The size of the diphone database may vary depending on the language. These units are combined by DSP techniques in the synthesis process resulting less quality speech.

Domain-specific synthesis concatenates long pre-recorded sample of natural speech, like syllables, words, phrases, and sentences. This method provides high quality synthetic speech, but has a limited vocabulary. So it is used in limited, a particular domain. It is very suitable announcing (transit schedule announcements) and information systems (price list, the weather forecasting report), digit synthesis is concern only pronunciation of number sequence(Utama & Syrdal, 2006). In this approach, the longer the phoneme means the more success of the system. For the units with different lengths, intelligibility and flexibility of system are also different. Flexibility is worse but naturalness of speech is better, if big units are used. Besides big units are for limited domain applications. Conversely, small units are for more flexible systems and wider applications. The Figure 2.12 shows that the variation of the speech unit in relation to flexibility and intelligibility.

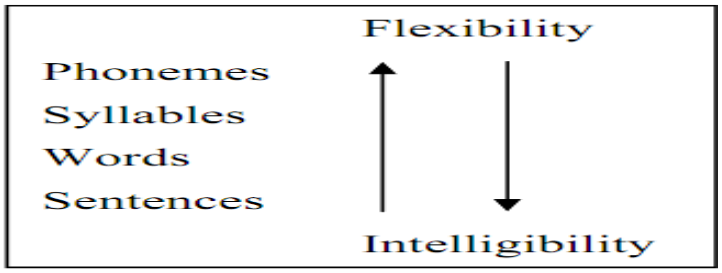


Figure2. 12: Flexibility and Intelligibility variation according to unit length

2.7. Related Works in Local Languages

Speech synthesis is one of the most popular research areas in the field of Speech Science. Past research results indicate that a lot has been done in the area. Currently, researches in speech synthesis are getting more attention by different scholars. Many techniques are available to handle speech synthesis, but hard it is to find a best method that satisfies the naturalness and intelligibility of the synthesizing synthetic speech. Nowadays, investigations are underway in order to get the advantage of the two major speech sound properties for local languages (Tadesse, 2009 & 2011, Sebsibe, *et al.*, 2004, and Yibeltal, 2008). Since Ethiopia is a multi-linguistic society, which needs multi-lingual synthesizer and tried by different scholars. Some of the attempts done using concatenative-based speech synthesis techniques are discuss the next section.

2.7.1. Text-to-Speech synthesis for Tigrigna

Tesfay (2004) developed a prototype for Tigrigna language TTS system using concatenative techniques and diphone based speech segment. The researcher use Time Domain Pitch Synchronous Overlap Add(TD-PSOLA), a method that generate synthetic speech with less computational intensive manner (Tesfay, 2004). The system performance is evaluated by adopting the Mean Score Opinion on the selected speech corpus and obtained an average mean score of 3.05, which is closer to the scale level of good. He recommends as a future work that it is necessary to incorporate the tasks and processes of normalization texts during speech corpus preparation.

2.7.2. Concatenative Amharic Text-to-Speech system

A research work done by Henock (Henock, 2003), applied concatenative as synthesis techniques, speech units of diphone and syllable to synthesize sample and time domain pitch synchronous overlap-add (TD-PSOLA) algorithm to analyze and generate synthetic speech. In addition to this, the researcher also considered prosodic factors. Taking into account such effects in to their work is vital for the effectiveness of speech synthesis to find the natural human like synthetic sounds, which make this work a good attempt in speech science

especially for local Ethiopian language, Amharic (Henock, 2003). For the system performance evaluation, the researcher adopted Open Rhyme Test method to show the system performance and achieved result is 88% and 75% for diphone and syllable respectively. Based on result analysis, the researcher concluded that Diphone based synthesis gives better result than syllable based as speech unit selection. The researcher suggested that syllable as speech units and concatenative technique with large speech database can be applied on different local Ethiopian language for future work.

2.7.3. Unit selection voice for Amharic using Festvox

Amharic language TTS system was done by Sebsibe (2004), called "Unit selection voice for Amharic using Festvox", which 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 is a voice building framework used to build unit selection voices in a new language Festival speech synthesis system were adopted. The performance levels range from Excellent (5) to Very Poor (0) and the system achieved cumulative result of 2.9.

In this work, the researcher suggested as future work that during modelling speech synthesizer for Amharic language the proper selection of unit and optimal selection of corpus will give a better quality of speech waveform from the synthesizer (Sebsibe, *et al.*, 2004). Even though, the above work contributed greatly in the speech synthesis area, speech synthesizer for Amharic language is still not commercialized properly to help in different application as an assistive technology.

Based on the reviewed made so far and knowledge of the researcher, none of the works has tried to design speech synthesiser for Amharic, which synthesize by accepting normalized Amharic texts and generate prosodic features (i.e. intonation, stress) using syllable based approach. The main focus in this work is to find the quality speech corpus, which matters the quality of synthetic speech from synthesizers including linguistic tasks properly.

CHAPTER THREE

3. AMHARIC PHONOLOGY

3.1. Overview of the Amharic Language

Ethiopia is a linguistically diverse country where more than 80 languages are used in day-to-day communication among people. Although many languages are spoken in Ethiopia, Amharic is dominant in that it is spoken as a mother tongue by a substantial segment of the population and it is the most commonly learned second language throughout the country (Bender, *et al*. , 1976). The language is the official working language of the federal government of the country, in different regions. According to the world languages report of the country Ethiopia (Ethnologue, 2006), and the 2007 census of Ethiopian central statistical authority (ECSA, 2007), Amharic is used for more than 32.7% of the population as first and second language.

Amharic *(አማርኛ)* is a Semitic language spoken in North Central Ethiopia by the Amhara. It is the second most spoken Semitic language in the world, after Arabic, and the "official working" language of the Federal Democratic Republic of Ethiopia. It is also the official or working language of several of the states within the federal system, including Amhara region and the multi-ethnic southern nations, nationalities, and people's region (SNNPR), among others, even outside the country. It has been the working language of government, the military, and of the Ethiopian Orthodox Tewahedo Church throughout modern times. According to Bender *et al*.(1976), it is the language of some 2.7 million emigrants (notably in Egypt, Israel and Sweden), and is spoken in Eritrea by some Eritreans of the pre-independence generation and younger deportees from Ethiopia.

Being a Semitic Language of the Afro-Asiatic Language Group, this language is related to Hebrew, Arabic, and Syrian. Unlike Arabic, Hebrew, or Syrian, Amharic language is written from left to right. It is one of the most widely spoken languages in Ethiopia. It has its own script that is borrowed from Ge'ez, another Ethiopian Semitic language (Leslau, 1996). The script is believed to have originated from the South Sabean script. It is a syllabary writing system where each character represents an open CV syllable, i.e., a combination of a consonant followed by a vowel (Daniel, 2006). The Figure 3.1 below shows that the pro-Semitic language family and origin of Amharic language as Ethiopic script.

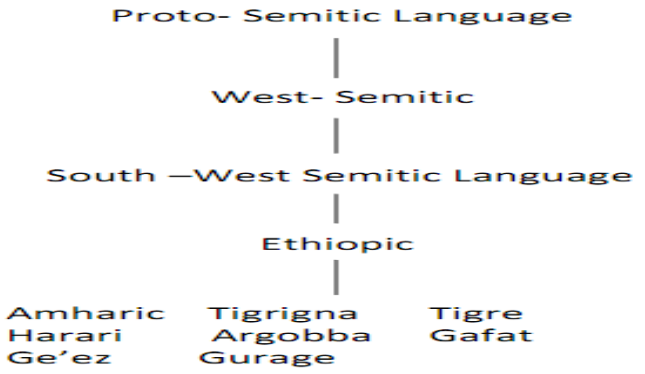


Figure3. 1: Semitic Language Family

3.2. Amharic language script

The Ethiopic (Ge'ez) script was developed as the writing system of the Ge'ez language, a Semitic language spoken in Ethiopia and Eritrea until the 10th to the 12th centuries. The original Ge'ez script was an Abjad - vowels were not written - but the current script is classified as an Abugida. Each symbol represents a CV-syllable, but vowels are not inherent in the consonant. According to Gasser (2009), the original Ethiopic script contained 182 characters, although the basic (unmarked) consonants number only 26. The script has since been extended for other languages and now contains over 500 symbols. Some of the new symbols represent phonological processes such as palatalization, pharyngealization, and labialization. The unmarked set is known as the first order or first form. Each of the first order consonants can be combined with one of six vowels, to produce a syllograph. The resulting sets of syllographs are known as the second, third, fourth, fifth, sixth and seventh orders. Although the language ceased to be used in vernacular speech (it now serves a liturgical function only), the script is still widely used for writing the Ethiopian and Eritrean Semitic languages such as Tigré, Amharic and Tigrigna (Gasser, 2009a). Amharic script, unlike other Semitic scripts, is written from left to right and found in five dialectical variations (Addis Ababa, Gojjam, Gonder, Wollo, and Menz) spoken in different regions of the country.

3.3. Characteristics of Amharic language

As same with other Semitic languages, Amharic has its own characterizing phonetic, phonological, and morphological properties. For example, it has a set of speech sounds that is not found in most of other languages-{P}[ጰ],{S}[ጸ],{x}[ኀ], and {q}[ቀ] (Gasser, 2011). It is morphologically rich but it is extremely complex (Gasser, 2009b, Gasser, 2011), has a complex morphology, with nouns (and adjectives) being inflected for gender, number, definiteness, and case. Definite markers and conjunctions are suffixed to the nouns, while

prepositions are prefixed. Major changes during process of the language in different aspects, for instance, change of meaning, syntax, phonetic, etc (Sisay & Haller, 1987). The case of Amharic is not different; the script underwent changes when it was borrowed from Ge'ez, through the adaptation process and other factors, the Amharic writing system called Amharic syllabary got some problems (Scelta, 2001).

The first problem is the presence of "unnecessary" alphabets (fidels) in the language's writing system. These fidels (alphabets) have the same pronunciation but different symbols, these different fidels can be used interchangeably without meaning change. The fidels are *ሶ* and *ዐ, ጸ* and *ፀ, ሰ* and *ሠ* and *ሀ, ሐ*, and *ኀ*. For example, the word "sun" can be written as, *ጸሀይ, ጸሃይ, ፀኃይ, ፀሃይ*, etc, all mean the same, although they are written differently and produce different orthographic form. The other problem is in the formation of compound words. Compound words are sometimes written as two separate words and sometimes as a single word. For example, the word "school" can be written as "*ትምህርትቤት*" or *"ትምህርት ቤት"*. There are many such compound words, which need some effort to have a standard way of forming them.

Amharic is morphologically rich language where up to 120 words can be conflated to a single stem (Gambäck, *et al.*, 2002). The word units of Amharic are phoneme, morpheme, root, stem, and word. The 34 base characters are a phoneme. A collection of phonemes forms morphemes, which is the smallest meaningful unit in a word (Baye, 1987). An Amharic root is a sequence of base characters. A collection of phonemes or sounds creates a word, which can be as simple as a single morpheme or contain several of them. In addition, in Amharic language, it is common to write some words in shorter form using "/" (forward slash) or "." (dot). The short form of words can be expanded as single or a combination of words. "*አ/አ*", which is expanded as *"አዲስ አበባ"* (means Addis Ababa), is an example for the latter. "*መ/ር*" is a short form of the single word "*መምህር*" (means teacher).

In addition, the language there are different ways of writing a single word due to different reasons such as regional dialects that can influence word formation in the basic level where the words are more likely to be written following their spoken form; "*ሂጂ*" vs. "*ሂጅ*", "*ዓለ*" vs. "*ዓሇ*", etc (Daniel, 2006) and many ways of writing loan words, i.e. words that are taken from foreign languages. For example, the word "computer" can be written as *ኮምፒዩተር, ኮምፒውተር, ኮምፒዉተር*, etc.

3.4. Amharic Alphabet

Amharic is written with a version of the Ge'ez script known as Fidel, present standard writing system of Ethiopic. According to Baye's analysis, there are three writing systems used in Ethiopia, i.e. the Amharic syllabary, the Roman alphabet, and Arabic script (Baye,2006). The Amharic syllabry, which is derived from the writing system of ancient South Arabian inscriptions, is used for Ge'ez, Amharic, with slight modification. The Amharic syllabry is uniquely Ethiopian writing system. The writing system has a similarity with some Semitic languages like Arabic in having vowel marks added to consonant letters. The present writing system of Amharic is taken from Ge'ez script, known as Amharic syllabary. Ge'ez in turn took its script from the ancient Arabian language mainly attested in inscriptions in the Sabean dialect (Baye, 2006 and Gasser, 2009b). The original Sabaean alphabet is said to have had 29 symbols. When Ge'ez became the spoken and written language in common use in northern Ethiopia, it took only 24 of the 29 Sabaean symbols, modify most of them, and add two new symbols to represent sounds of Greek and Latin loan words not found in Ge'ez.

There is no standard way to transliterate Amharic into the Latin alphabet. The Amharic alphabet consists of thirty-three basic characters, each of which has six additional modified characters. The modified characters represent the basic sound of the symbol augmented with a vowel. Thus, the main table of the traditional Amharic syllabary appears as characters set in thirty-four rows and seven columns. Languages using such a scheme have been termed to use an "Abugida" instead of an alphabet (Gasser, 2009b)

However, due to repetition of some characters/fidels like ኦ:ዐ, ጸ:ፀ, ሰ:ሠ and ሀ:ሐ:ኀ and as Baye (2010), discussed that each row is dedicated to the thirty consonants and the columns or orders represent the seven different phonemes resulted on the application of the vowels in a regular fashion. From this, it's apparent that Amharic writing system is partially phonetic: i.e. there is more or less a one-to-one correspondence between the phones and the symbols. The thirty-three core characters by seven orders gives (33*7=231) provide the Amharic distinct symbols. In addition to this there are others that contain special features usually representing labialization like {kwa} [ኵ], {gwa} [ጕ], {qwa} [ቍ]. Regarding the punctuation marks, there are about seventeen, of which only few are commonly used. Thus approximating the writing system can be achieved by taking out the redundant symbols without losing essential understanding (Baye, 2006).

Abugida is a term used for a script whose basic signs denote consonants augmented with a vowel and where consistent modifications of the basic sign indicate augmentation of other vowels (Daniels, 1997). It is derived from the first four characters of one type of ordering of

the Ge`ez script. The Amharic script is an Abugida, although the vowel modifications in Amharic are not entirely systematic. As a result of using an Abugida, each concatenated consonant-vowel (CV) syllable in the Amharic language has its own corresponding orthographic symbol. A few syllables, however, are exceptions to this rule; these syllables are represented by more than one symbol in the character set and these redundant orthographic symbols used to make differences in sound realization as well as meaning in old Amharic. In modern Amharic, these differences do not exist (Ruth, 2008). Another implication of the use of Abugida is that there are a large number of characters in the language, which makes encoding Amharic into computer keyboard is difficult. Using current Amharic text encoding programs and typewriters, it routinely takes 2 or 3 keystrokes to type in a single character. It is not unheard of to find characters that take 4 keystrokes on some of the Amharic text encoding programs. It is instructive to compare this with English that requires only a single keystroke for all its characters except for certain special characters and capital letters in which case only two keystrokes are necessary. The Amharic Abugida is composed of rows and columns instead of a linear list of letters as is the case for the English alphabet. Each row contains seven columns, called vowels/*አናባቢ*/ in Amharic. The rows represent the basic phonemes while the columns represent the vowels appended to the consonants. Thus, each row of the alphabet lists the concatenation of the particular phoneme with the six vowels with one column (the sixth) reserved for the basic phoneme without any concatenation. The order of the vowels in the row is consistent for all phonemes and is given by the following positional assignment. The Amharic Abugida system which is mainly adopted from Habtamu (2006) is shown in Appendix 1.2.

3.5. Consonant phonemes

Consonants, as opposed to vowels, are characterized by significant constriction or obstruction in the pharyngeal and/or oral cavities, that some consonants are voiced; others are not. A set of 34 phones, seven vowels/*አናባቢ*/ and 27 consonants/*ተነባቢ*/ and 7 other repeated consonants ,makes up the complete inventory of sounds for the Amharic language (Baye, 2010). Amharic consonants according to manner of articulation are generally classified as stops/*ዝግ*/, fricatives/*ተሸለቀሏቸ*/- formed when the stricture is very narrow (but without total closure) so that when air flows out, a hissing noise is made, and lateral- is made when air flows out of the sides of the mouth, nasals/*የአፍንጫ*/- air to flow out of the nose, liquids/*የምላስ*/, and semi-vowels/*ግማሽ አናባቢ*/. Figure 3.2 shows the phonetic representation of the consonants of Amharic as to their manner of articulation, voicing, and place of articulation are of voiceless /*የማይነዝር*/, voiced/*ነዛሪ*/, and glottal/*የጉሮሮ*/.

Amharic also has its own inventory of speech sounds. It has thirty consonants (27 simple and 3 complex) (Nadew, 2008) and seven vowels. Amharic consonants are generally classified based on their voicing, manner, and place of articulation. In articulatory phonetics, the place of articulation (also point of articulation) of a consonant is the point of contact where an obstruction occurs in the vocal tract between an articulatory gesture, an active articulator (typically some part of the tongue), and a passive location (typically some part of the roof of the mouth). As same work done by (Habamu,2006), and discussed about the place of articulation are identified as labials, alveolar, palatals, velars and labio-velar and glottal. Figure 3.2 shows the phonetic representation of the consonants of Amharic as to their manner of articulation, voicing, and place of articulation.

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

Figure3. 2: Phonetic representation of Amharic consonants(Sebsibe, et al, 2004)

3.6. Vowel phonemes

Vowels are always voiced sounds and they are produced with the vocal cords in vibration. Most languages have five vowels/*a, e, i, o, u*/, but in case of Amharic, there are seven vowels (see Figure 3.3) below. In addition to the five vowels which is common for different languages, Amharic has two central vowels, /*e*/ and /*ix*/, the latter with a mainly epenthetic function, used for epenthesis vowel insertion/. It is impossible to analyze the effects of vowels when we are speaking, rather we can observe clearly the effects of those when we write, due to that listener sees only written form, rather than spoken form. The effects are clearly analyzed and understood during acoustic formation of sounds as in CV-syllable assimilation.

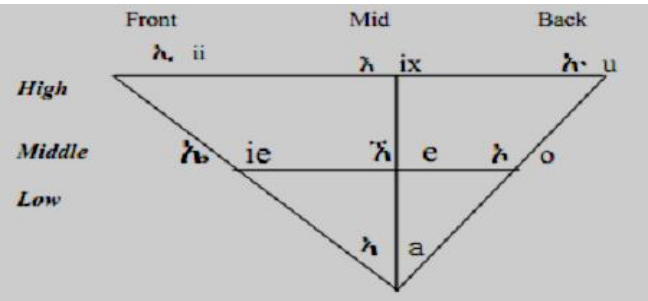


Figure3. 3:Vowels with their features of Amharic (mainly adopted from Henock(2003))

According to the above Figure 3.3 , depending on the position of the lip the Amharic vowels (*ሀ፣ ሁ፣ ሂ፣ ሃ፣ ሄ፣ ህ፣ and ሆ*) are broadly categorized into rounded *(ሁ and ሆ)* and unrounded *(ሀ፣ ሂ፣ ሃ፣ ሄ and ህ*).

3.7. Syllable structure of Amharic

Syllables are very important unit of language, without the complete knowledge of syllables there is no possibility to apply linguistic in speech applications (Hayes, 2009). Syllables are the combination of consonants and vowels. Amharic words use consonantal roots with vowel variation expressing difference in interpretation. In modern written Amharic language, each syllable pattern comes in seven different orders, reflecting the seven vowel sounds. The first order is the basic form; the other orders are called derived from which is formed from more or less regular modification using those seven vowels to drive others.

Amharic has its own non-Latin based syllabic script called “Fidel” or “Abugida”. The Ethiopian script is not strictly speaking an alphabet, but what is called a syllabary. This means that each letter or symbol usually represents a whole syllable. There are thirty-three basic shapes, represent the consonants followed by the seven vowels (a, e, i, o, u, e and ix). The basic shapes are altered in various ways to indicate a different vowel following the base consonant. The Amharic syllabary is usually presented as a grid with the vowels in the horizontal axis/columns and the consonants in the vertical axis/rows. For instance, the syllabary of the most common Amharic words as: *ሀ/ha/, ሁ/hu/ --- ሆ/ho/* order. For example, the orthographic representation of Amharic fidel “ለ” is organized into orders (derivatives) as shown in table 3.1; six of them are CV combinations while the sixth order is the consonant itself that needs further investigation. In total there are 32 consonants and 7 vowels with 7x32= 224(by removing the six order consonants which needs special attention in this study) found as CV-Syllables. In many languages, orthographic transcription has some relation to phonetic transcription. As researchers on this field states, since there are redundant sounds in Amharic orthographic that represent the same sounds, the phonemes are only 28 (by

removing the repeated characters i.e. ዐ, ዘ, ሠ, ሐ *and* ኀ). For example, the Fidel *"*በ**"** can be orthographically transcribed as:

1st	2nd	3rd	4th	5th	6th	7th
በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
B/e/	*B/u/*	*B/i/*	*B/a/*	*B/ie/*	*B*	*B/o/*

Table3. 1: Amharic Syllables structure of character "በ"

As different researchers in linguistic field stated and agreed, most of consonants can occur word-initially; any single consonant can be a syllable onset and the general structure of a syllable consists of the following segments:

- ❖ Onset (obligatory in some languages, optional or even restricted in others)
- ❖ Rhyme (final sounds/last syllable)
  - ✓ Nucleus (obligatory in most languages)
  - ✓ Coda (optional in some languages, highly restricted or prohibited in others)

Onset-Rhyme (OR) or Onset-Nucleus-Coda models of syllable structure were developed and when looking vowel and consonant assimilation from phonological perspective, a syllable is often made up of a consonant plus a vowel or a single vowel (Roelofs, 2002), there is no vowel-vowel assimilation in Amharic language. This follows the principle of maximal onset – minimal coda. The maximal onset principle states that the maximum number of consonants possible to attach a syllable onset (Côté, 2005). As seen in Figure 3.4 below, the syllable is made of rhyme and onset. Within rhyme (or core) we find peak (or nucleus) and coda. For Amharic language the internal organization of the syllables, used same structure as like other syllable based languages. The Figure 3.4 below shows the internal organization of **σ** –syllable for the case of Amharic texts.

The internal general organization of syllables is characterized as seen in Figure 3.4 below.

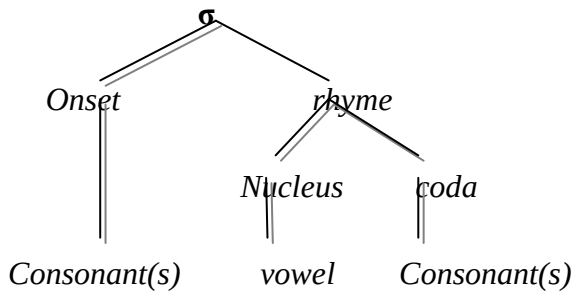


Figure3. 4: General syllable structure **σ** -syllable (Adopted from Côté (2005))

Like other languages, Amharic also has its own typical phonological and morphological features that characterize it by its own syllable structure. The following are some of the striking features of Amharic phonology that gives the language its characteristic sound when one hears it spoken: the weak, indeterminate stress; the presence of glottalic, palatal, and labialized consonants; the frequent gemination of consonants and central vowels/*e*/; and the use of an automatic helping vowel (Mulugeta, 2001). In addition, there are several factors that affect the duration of a syllable, such as it position in a word, whether it is stressed or unstressed, etc. In Amharic, stressed syllables tend to be longer than unstressed ones; it is same with other languages.

3.8. Role of gemination, syllabification and epenthesis on speech synthesis

3.8.1. Gemination

Gemination/*ጥብቻ*- (consonant lengthening) is not normally indicated in the Ge'ez script. However, since it plays a significant role in the morphology of Amharic language and is important for speech applications, it is shown in the romanized forms of words. It is distinct from stress and may appear independently of it by doubling of consonants. As in most other Ethiopian Semitic languages, gemination is contrastive in Amharic. That is, consonant length can distinguish words from one another; for example, "*ኣለ*"- *alä* means 'he said' and *"ኣለ"- allä* means "there is*"; ይመታል-"yəmätall*" means "he hits", and *ይመታል-"yəmmättall*" means "he is hit".

Gemination is not indicated in Amharic orthography, but Amharic readers seem not to find this to be a problem. This property of the writing system is analogous to the vowels of Arabic and Hebrew or the tones of many Bantu languages, which are not normally indicated in writing. The noted Ethiopian novelist Haddis Alemayehu, who was an advocate of Amharic orthography reform, indicated gemination in his novel *"Fəqər Ǝskä Mäqabər*"-*"ፍቅር እስከመቃብር*", by placing a dot above the characters whose consonants were geminated, but this practice has not caught on. It is one of the most distinctive characteristics of the cadence of the speech, and carries heavy semantic and syntactic functional weight (Mulugeta, 2001). Amharic gemination is either lexical or morphological. Gemination as a lexical feature cannot be predicted. For instance, "*ኣለ*" may be read as *alä* meaning 'he said', or *allä* meaning 'there is'. Although this is not a problem for Amharic speakers, it is a challenging problem in speech synthesis, in case of written form.

In many of Amharic words gemination occurs frequently (Nirayo, 2011), for example: - we can observe the difference between /*kixft*/ and /*kixffixtt*/ because of the geminate consonant /*f*/ (*ፍ*) and /*t*/ (*ት*) in the word "*ክፍት*" and /*gena*/ or /*genna*/ in the word "*ገና*" depending on the

context. It is possible to shows the difference word in waveform generation using waveform generation tools (like wavepad sound editor software). When /*f*/ and /*t*/ are geminated there is epenthesis vowel /*ix*/ inserted between them therefore syllabification the word becomes completely different. Without gemination the word contains only single syllable type CVCC. Therefore, the whole phoneme sequence, /*kixft*/, is taken as a single syllable. However, when it is geminated, the word will have two CVC syllables, /*kixf-fixtt*/ and similarly we can observe the missing epenthetic vowel between the two consonants /*f*/ and /***t***/ in the waveform. The Amharic writing system lacks the ability to differentiate between geminated and non-geminated consonants in the Amharic words, hence there is no simple way to know whether a consonant is read as geminated or not. This problem is solved by using special character after a consonant to show that is geminated, called apostrophe symbol (‘) and works by it doubling the consonant part of the CV-syllable before the apostrophe.

Unlike English language in which the rhythm of the speech is mainly characterized by stress (loudness), rhythm in Amharic is mainly marked by longer and shorter syllables depending on gemination of consonants, and by certain features of phrasing. In Amharic, all consonants except (*ሀ*)/*h*/ and (*ዐ*) /*ax*/ may occur in either a geminated or a non-geminated form. Amharic gemination is either lexical or morphological. As a lexical feature it usually cannot be predicted. The failure of the orthography of Amharic to show geminates is the main challenge in G2P conversion (Tadesse, 2011). A cluster of consonants is nothing but a succession of two consonants without a splitting by any vowel. In Amharic, the maximum number of allowable consonant sequences in a cluster is two (Mulugeta, 2001). Consonant clusters are not permitted at the beginning of a word. However, some speakers pronounce an initial cluster when the second element is a liquid.

Onset cluster are not permitted in Amharic. The epenthesis vowel should be inserted between the liquid and the preceding consonants. Therefore, whenever we find consonant clusters at initial position we have to insert epenthetic vowel except for the phoneme /*w*/; if it occurs next to the initial phoneme. But, final clusters of consonants are permitted in Amharic. However, if the hierarchy of sonority sequence is not satisfied, the final cluster takes an epenthesis vowel and the cluster splits.

3.8.2. Epenthesis vowel insertion

Phonotactics of a language is the permissible combinations of phonemes that can co-occur in that language. Thus, phonotactics helps us to reduce the number of syllables units that need to be covered in the unit inventory of a language, syllable based speech database. The process of epenthesis is common in Amharic. Mostly it can occur word-initially or medially. As

(Hudson, 1996), stated epenthesis is extensive in word-formation in the Ethiopian Semitic languages, since many morphemes, both roots and affixes, consists only consonants. Modeling identification of the epenthetic vowel improves speech synthesis process (Sebsbie *et al.*, 2004). We also understood from our empirical observation epenthetic vowels has great role in syllabification of Amharic words, and it is a common phenomenon for this language. It can occur at word initial, word medial, and word final position. During the process of epenthetic vowel insertion, if the epenthetic vowel is inserted in a word medial consonant cluster, it's identity is dependent on the identity of the vowel following the cluster while if the epenthetic vowel is inserted in a word final consonant cluster, it's identity is determined based on the word final consonant. Moreover, while implementing grapheme-to-phoneme conversion the written form and the spoken form is one to one except the epenthetic vowel (Tadesse, *et al.*, 2010). The rules of epenthesis will decide the presence or absence of such vowel in the spoken form of the language.

If a consonant cluster in a syllable violates the phonotactic constraints of Amharic, it is broken using epenthesis. There are two general rules concerning automatic insertion of an epenthetic vowel in Amharic. Those rules are word-initially no consonant clusters are allowed and elsewhere clusters of no more than two consonants are tolerated. For example, Hudson(1996), proposes three environments for epenthesis corresponding to the possibilities of consonants sequence word initial, medial and final position in Amharic.

1. #CC as in words like /*tsebr*/ →/*tixsber*/, /sber/ →/*sixber*/ (word initial consonant clusters are impermissible) (the (#) indicates the position of the cluster, word initial or word final).
2. CC# as in word like/*mkr*/→ /*mixkixr*/. In this case, the sonority of the final consonant, /*r*/, is greater than that of the preceding consonant, /*k*/. Thus, to split up the final cluster epenthetic vowel /*ix*/ is inserted. On the other hand, if the sonority of the first is equal or greater than that of the second consonant, epenthesis will not be applied.
3. CCC, in the case of this environment Hudson (2000) proposes three types of CCC violation where epenthesis /***ix***/ is required.
    a. CCC→ CCixC, in a word like /*fendto*/→ /*fendixto*/ "exploit"
    b. C:C→ C:ixC, in a word like /*fellgo*/→/*fellixgo*/ "want"
    c. CC:→CixC:, in a word like /sebrre/→/*sebixrre*/ "break"

The epenthetic vowel insertion/*ሰርጉ-ገብ*/ was tested by recording words and look into the acoustic evidence (waveform and spectrogram). Example we can see the effects of epenthetic vowel insertion/ix/, without and with on the word *"ትምህርት"* by transcribe into two words as

"*tmhrt*" and "*tixmhixrt*" respectively and we can mark the difference in their waveform and we can observe the missing of epenthetic vowel/*ix*/ during waveform generation of those Amharic words, especially in the six-order CV-syllables. For instance, when we see the six-order in Table 3.1, there is no vowel inserted in case of character/*ኧ/ i*s transcribed, which create problem in waveform generation of Amharic words. To solve this problem it is necessary to inculcate the epenthetic vowel/*ix*/ during conversion of words into CV-syllables. Epenthetic vowel is used to split those impermissible words according to the basic rules of the Amharic syllabification. Previous works on Amharic TTS do not see this issue as one of the factors, which affects the naturalness, and intelligibility of speech sound.

3.8.3. Syllabification

A syllable is a basic unit of word studied on both the phonetic and phonological levels of analysis. It is typically composed of more than one phoneme. No matter how easy it can be for people and even for children to count the number of syllables in a sequence in their native language, still there are no universally agreed upon phonetic definitions of what a syllable is. It is phonologically believed that syllable is a complex unit made up of nuclear and marginal elements. There is general agreement that a syllable consists of a nucleus that is almost always a vowel, together with zero or more preceding consonants (the onset) and zero or more following consonants (the coda) but determining exactly which consonants of a multisyllabic word belong to which syllable is problematic. Nuclear elements are the vowels or syllabic segments, and marginal elements are the consonants or non-syllabic segments. Standard dictionaries provide syllabification that is influenced by the morphological structure of words; it is common in such dictionaries to split prefixes and suffixes from stems (Cholin, 2004).

A syllable can be described by a series of grammars. The simplest grammar is the phoneme grammar, where a syllable is tagged with the corresponding phoneme sequence. The consonant-vowel grammar describes a syllable as a consonant-vowel-consonant (CVC) sequence. The syllable structure grammar divides a syllable into onset, nucleus, and coda (ONC). Figure 3.4, shows and example of **σ** –syllable structure above. Amharic grammar books like (Getahun, 2010 and Baye, 2010) describes the grammatical syllable structure of Amharic words. Some papers for instance, (Aster, 1981, Mulugeta, 2001) also present the syllable structure of Amharic words.

Syllables play an important role in speech synthesis and recognition apart from their purely linguistic significance. The pronunciation of a given phoneme tends to vary depending on its location within a syllable while actual implementations vary, TTS systems must have, at

minimum, following modules: a letter-to-phoneme module, a prosody module, and a synthesis module. Syllabification can play a role in all speech synthesis modules like NLP, DSP (Nirayo, 2011). Also in speech recognition, syllabification has been used to build recognizer which represents pronunciations in terms of syllables rather than phones. In addition, syllabification can help annotate corpora with syllable boundaries for corpus linguistics research (Sisay & Haller, 1987).

Amharic is a syllabic language in which every of our grapheme (character) represent (Consonant-Vowel Assimilation). However, while reading a text in Amharic, all the syllables are not uttered as expected and hence the text syllables are not the CV sequence seen in the text. This limits performance of many speech systems and other NLP applications. A great number of diverse algorithms have been proposed for syllabification in different languages and many researches has been done on other languages. In many languages, the pronunciation of phonemes is a function of their location in the syllable relative to the syllable boundaries. Location in the syllable also has a strong effect on the duration of the phone and on the temporal alignment of the fundamental frequency contour with the segmental chain , and is therefore a crucial piece of information for segmental duration and intonation models (Musa, *et al.* , 1969 ,and Tadesse, *et al*., 2010).

The complexity of syllable onset and coda structure poses serious problems for a syllabification algorithm because despite restrictions as to which consonants, or classes of consonants, may occur in any given position within the onset or coda of a syllable-ambiguous and multiple alternative syllable boundary locations are usually observed in polysyllabic words, notably in compounds (Kurian & Narayan, 2011).

It has been identified that there are six legal syllable structures (templates) in Amharic, namely *V, VC, VCC, CV, CVC* and *CVCC* for words (Sebsibe, *et al*., 2004), which is the most convenient template of syllable structure for Amharic. Though a number of examples for syllabified words belonging to each of the above structures are presented in the literature (Mulugeta, 2001 and Aster, 1981), the methodology, or grammatical rules describing how to syllabify a given word has not been presented by the researchers. A word can be syllabified in many ways retaining the permitted structures, but only a single correct combination of structures is accepted in a properly syllabified word. For example, a word having the consonant-vowel structure *VCVCVC* can be syllabified in the following different ways, retaining the valid syllable structures described in the literature: *V-CVC-VC, VC-VC-VC,* and *VC-V-CVC.* However, only one of these forms represents the properly syllabified word. As discussed on and mentioned in pervious, the six order *CV*-syllables can be read in two forms:

with and without vowel sound/*ɨ*/. The transcription becomes relatively simple when a six-order symbol comes at the end of a word. In the case of this, the sound *ɨ*-[ix**]** or [*i*] will be omitted. The problem arises when a cluster of six-order CV-syllables come at the middle of a word, since there is no straight forward way of determining which one of the six-order CV-syllables is read with the vowel sound and which is omitting. For instance, in ልብ /*lixbb*/ 'heart', the first character, ል, represents the CV sequence /*lix*/ (voweled), whereas in ስልክ /*sixlk*/ 'telephone', the same character represents the bare consonant /*l*/ (unvoweled). During transcription of the six order CV-syllable **'ል'** becomes / *lɨ*/ and /**l**/ which is stored separately, since such differences are crucial for speech synthesis, a TTS system needs access to the epenthesis rules.

3.8.4. Stress and Syllables

When we see and discussed among linguistic experts on stress assignment in Amharic, it is complex which need attention that scholars didn't agree there existence yet. However, there are some systems proposed in relation with stress and syllable structure. In many stress languages, stress is sensitive to a distinction called syllable weight. In a simple weight distinction, there are heavy and light syllables, defined as follows: heavy syllable-syllable that either ends in a consonant or has a long vowel or diphthongs and light syllable-syllable that ends in a short vowel.

Regarding the stress assignment rules of Amharic, we get the following rules form different literature. There are also other methods proposed by different scholars but the following rules have direct relation with syllables and syllable weight.

a. Stress falls on a heavy final syllable only in bi-syllabic words when the first syllable is light.
b. Otherwise, the final syllable is skipped and the right heaviest syllable is stressed.
c. In the absence of any heavy syllables, the left most of a string syllables is stressed.

Although stress assignment is beyond the scope of this thesis, once we have the syllables we can use the benefit of syllabification algorithm in order to have syllables and use rules of syllable weight assignment to assign stress based on the rules defined in relation with syllables and their corresponding weight. Therefore, having syllables and syllable weight for each syllable in the given word we can assign stress based on the rules specified.

3.9. Amharic words with their transcription

For Amharic language, before transcription is made the transliteration (representation of an alphabet with letters from a different alphabet) of each character is made by using the ASCII

value of each of the character. Transliteration is the practice of transcribing a word or text written in one writing system into another writing system or system of rules for such practice. For this work, the transliteration scheme proposed by Sebsibe (2004) is adopted (see Appendix 1.3). Transcription is the process of producing phonetic representation of a given Amharic text.

The International Phonetic Association (IPA) - responsible for standardizing representation of the sounds of spoken language defines a vowel as a sound, occurs at a syllable center and consonants as unsound which depend on the vowel. A chart depicting the Amharic vowels in the IPA representation is shown in Appendix 1.3. The IPA maps the vowels according to the position of the tongue. The vertical axis of the chart is mapped by vowel height. Vowels pronounced with the tongue lowered are at the bottom, and vowels pronounced with the tongue raised are at the top.

CHAPTER FOUR

4. DESIGN OF AUTOMATIC SPEECH SYNTHESIS ALGORITHM FOR AMHARIC

4.1. Approaches and Techniques

Based on the analysis in previous chapters on speech synthesis and Amharic phonology, a detailed description of design issues, techniques and approaches used for the Amharic TTS algorithm is dealt with in this chapter. Here, the general syllable based concatenative TTS for Amharic architecture is presented and novelty of the research is described. Moreover, this thesis present design and evaluation of concatenative speech synthesis approach and TD-PSOLA technique used for speech waveform analysis/synthesis and prosodic parameters modification (i.e. pitch and duration) is presented.

4.2. Design Goals and issues

The main goal of the study is developing syllable-based concatenative TTS algorithm for Amharic in relation naturalness and intelligibility. In order to improve the speech quality of current TTS systems in terms of naturalness, intelligibility and flexibility (Venugopalakrishna, *et al.*, 2003), three areas must be addressed: 1) improved linguistic analyses, 2) improved prosody modeling, and 3) improved speech synthesis models.

For this study, during corpus preparation some efforts done on linguistic analysis include: gemination, epenthesis and syllabification.

In general, the TTS system designed should pay attention on the basic issues that must be inculcate to come up with natural sounds. As different researches (Sarasathi & Vishalkshy, 2010, Chauhan, *et al.* , 2011, Rong-Wei, 2003, and Venugopalakrishna, *et al.*, 2007), stated in the area of speech synthesis that the quality of generating speech waveform depends on the following basic issues:

- The type of segments chosen as speech unit;
- The corpus they were extracted from and the corpus segmentation quality;
- The speech signal model to which the analysis and synthesis algorithms refers;
- The amount of degradation introduced by the speech coding phase ;
- The prosody matching efficiency, which is strongly related to the model and
- The capabilities of the segments concatenation algorithms used.

Taking the above-mentioned points clearly, the concatenative syntheses were used to synthesis the Amharic words, syllables as basic speech segment and TD-PSOLA algorithm for concatenation and speech waveform analysis-synthesis. The main idea of TD-PSOLA methods consists of the following: the initial original speech signal is multiplied by sequence of time windows synchronized with fundamental frequency (f0), for example Hanning window. The received sequence of acoustic segments is summed up and makes the required modified speech signals. During modification of duration and pitch of the speech signals the technology of repetition or elimination of some acoustic segments is used. The time domain approach was chosen within the framework of this thesis since it provides very efficient for real time implementation of synthesis systems.

In the next sections the common PSOLA and TD-PSOLA framework is presented in detail by describing the time axis warping to obtain time scale modifications during analysis and synthesis parts.

4.3. Speech Waveform Analysis-Synthesis Algorithms

4.3.1. Text-To-Speech Algorithm

In concatenative synthesis approach, the design of the cost function in relation to selection of unit is very important. However, there are some problems that arise in relation with selection of units in concatenative speech-synthesis systems. As based on Donovan (1996) analysis on concatenative synthesis methods by applying signal processing techniques to provide the following basic ideas i.e. to minimize the problems arise in speech synthesis like gap in spectrum, F0 contour, and power, which causes discontinuity at the concatenation points. The stabilization of speech quality and searching for optimal segments requires huge computational power.

To overcome the above stated problems, the concatenative method should use and apply the signal processing techniques effectively and efficiently with less amount. The TD-PSOLA method is used to enable pitch and duration transformations directly on the waveform for moderate ranges of prosodic modifications. These algorithms rely on a pitch-synchronous overlap-add (PSOLA) approach for modifying the speech prosody and concatenating speech waveforms, depending on the length of the window used in the synthesis process (Latsch & Netto, 2011).

Based on the analysis, the models and algorithms used as framework for speech analysis/synthesis and prosodic modification are: Linear Predictive Coding (LPC) Model (Markel, 1976), Hybrid Harmonic/Stochastic (H/S) Model (GRIFFIN, 1988), Hidden Markov

Model (HMM) (Plumpe, *et al*., 1998), Harmonic Pulse Noise Model (HNM), and Pitch Synchronous OverLap Add technique (PSOLA) (Moulines & Charpentier, 1990).

Each of the above models and algorithm mentioned has their own potential and drawback to provide high quality waveform generation from the synthesizer. In present time the most efficient techniques used for synthesis and speech modification is overlap-add methods (OLA). The PSOLA, is an OLA method which consists of different variants like Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) synthesis algorithm (Toma, *et al*., 2010), Frequency-Domain Pitch-Synchronous OverLap-Add (FD-PSOLA) synthesis algorithm (Conkie & Syrdal, 1998), Multi-Band Re-synthesis Pitch-Synchronous OverLap-Add (MBR-PSOLA) model (Dutoit & Leich, 1994), Synchronized Overlap-Add(SOLA) technique (Paola, *et al.*, 1985), and Overlap-Add technique based on Waveform Similarity (WSOLA) time-scale modification of speech (Verhelst & Roelands, 1993). The next sections discuss about the commonly used waveform synthesis techniques, Pitch Synchronous OverLap Add technique (PSOLA), especially TD-PSOLA algorithm in concatenative synthesis method. Since the thesis work use the Time Domain PSOLA (TD-PSOLA) technique, which is the most easy and computational efficient techniques among other types, as a result a special emphasis given for TD-PSOLA algorithm.

4.3.2. Pitch Synchronous OverLap Add technique (PSOLA)

The PSOLA is a digital signal processing technique used for speech processing and more specifically speech synthesis. It is actually not a synthesis method itself but allows pre-recorded speech samples smoothly concatenated and provides good controlling for pitch and duration (Donovan 1996). In recently developed synthesis techniques, the PSOLA technique has drawn inalienable attention because of its segmental and supersegmantal simplicity and efficiency for the concatenative synthesis.

The basic idea behind the algorithm is that it is possible to perform the duration and pitch modifications directly on continuous generated waveforms, without using any parametric model- to minimize the mismatch of prosodic parameters (Dutoit, 2001) and the modifications are performed without performing any explicit source/filter separation. The PSOLA algorithm used the Over-Lap and Add method (OLA), in which syllables are concatenated pitch synchronously in our case and modify the pitch and duration of a speech signal.

The basis of all the PSOLA techniques are to isolate pitch periods in the original signal, perform the required modification and resynthesize the final waveform through an overlap-add operation (Zervas, *et al.*, 2001). The techniques work by dividing the speech waveform in

small overlapping segments. The segments are then combined using the overlap-add technique, typically the division into segments is done using a specially modified speech recognizer set to a forced alignment mode with some manual correction afterward, using visual representation such as the waveform, pitch, pulse, intensity, formants and spectrogram. The PSOLA method consists of two major phases:

4.3.2.1. Analysis:

The analysis phase is the first phase that used to analyze the original speech waveform in order to produce an intermediate representation of the signal.

1. Determination of the pitch period or pitch mark that the original signal is divided into small blocks for which the pitch is considered constant. At the same time the pitch detection for each block is performed. The intermediate representation built from the speech waveforms consists of a sequence of short-term signals. They are obtained by multiplying the signal by a sequence of pitch synchronous analysis windows (see Figure 4. 1) or pitch shifting principle.
2. Extraction of a segment (block) centered over each pitch mark using a windowing techniques and functions (i.e. Hanning window), with the length of two pitch periods to allow for a smooth transition between the segments (fade in, fade out).

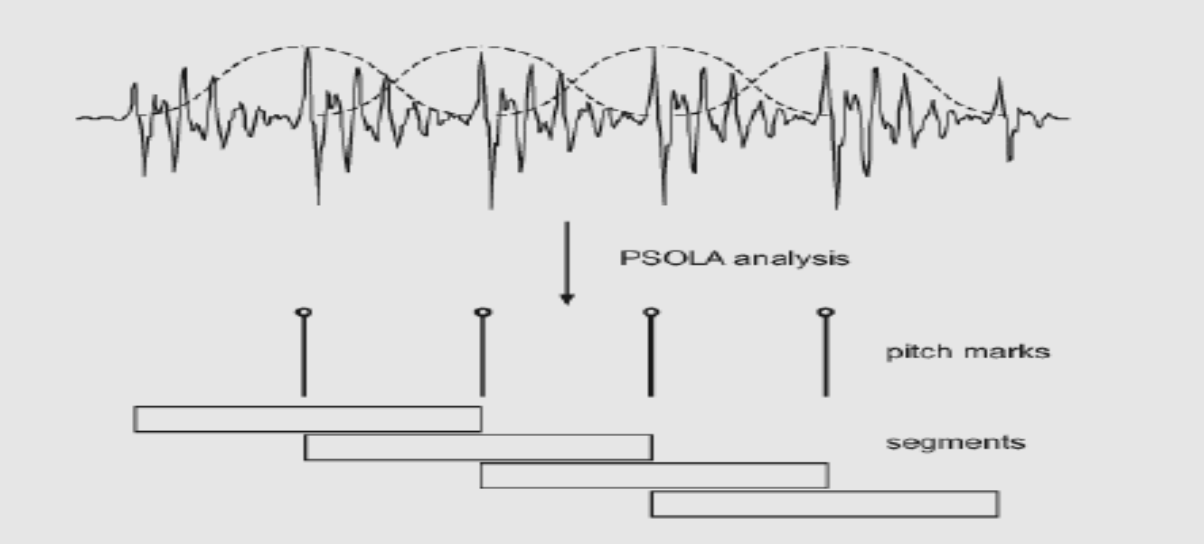


Figure4. 1: The PSOLA pitch analysis (pitch shifting)

4.3.2.2. Synthesis

At synthesis time, the best segments available to synthesize the new utterances are chosen from the corpus using a process known as unit selection. During the synthesis process, the pitch and duration of these segments may be modified to generate the desired prosody. In general, there are three steps in the PSOLA synthesis framework (Paola, *et al.*, 1985).

1. The choice of the corresponding analysis segment, which is identified by the time mark in the analysis phase.
2. Overlap and add the selected segment. At this point it is decided if the signal is going to be shrunk or stretched and repeated or deleted based on the scaling factor. If the scaling factor is less than 1, some segments will be discarded (time compression) and if the factor is more than 1, some segments will be repeated (time expansion) and
3. Determination of the time instant where the next synthesis will be centered in order to preserve pitch.

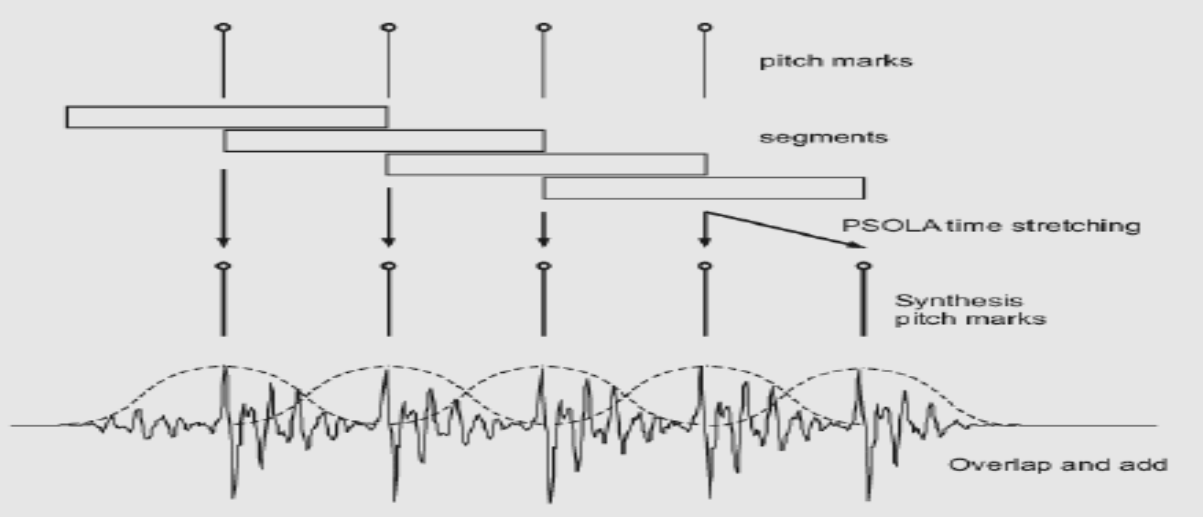


Figure4. 2: The PSOLA Synthesis (pitch shifting)

According to the PSOLA analysis-synthesis algorithm, the natural recorded speech is first divided into a number of short-term (ST) signals, done by a windowing function. That is the signal should first be divided into a number of small portions whose frequency is constant or nearly constant as an assumption. Windowing function segments a given speech signal into a number of small overlapping units by multiplying a signal and provide one for interest region and zero others. The windowing explained above causes one problem i.e. signal distortion. To minimize it we use smoother windowing functions like Hamming and Hanning. These windows are zero at the edge and rise gradually at the middle to be one. When we use those windows the edges of the signal are de-emphasised and the effects of the edge are reduced efficiently. The widowing signal and function are shown in Figure 4.3 below. The Hanning window function used to returns the N-point symmetric Hanning window in a column vector with the first and last zero-weighted window samples.

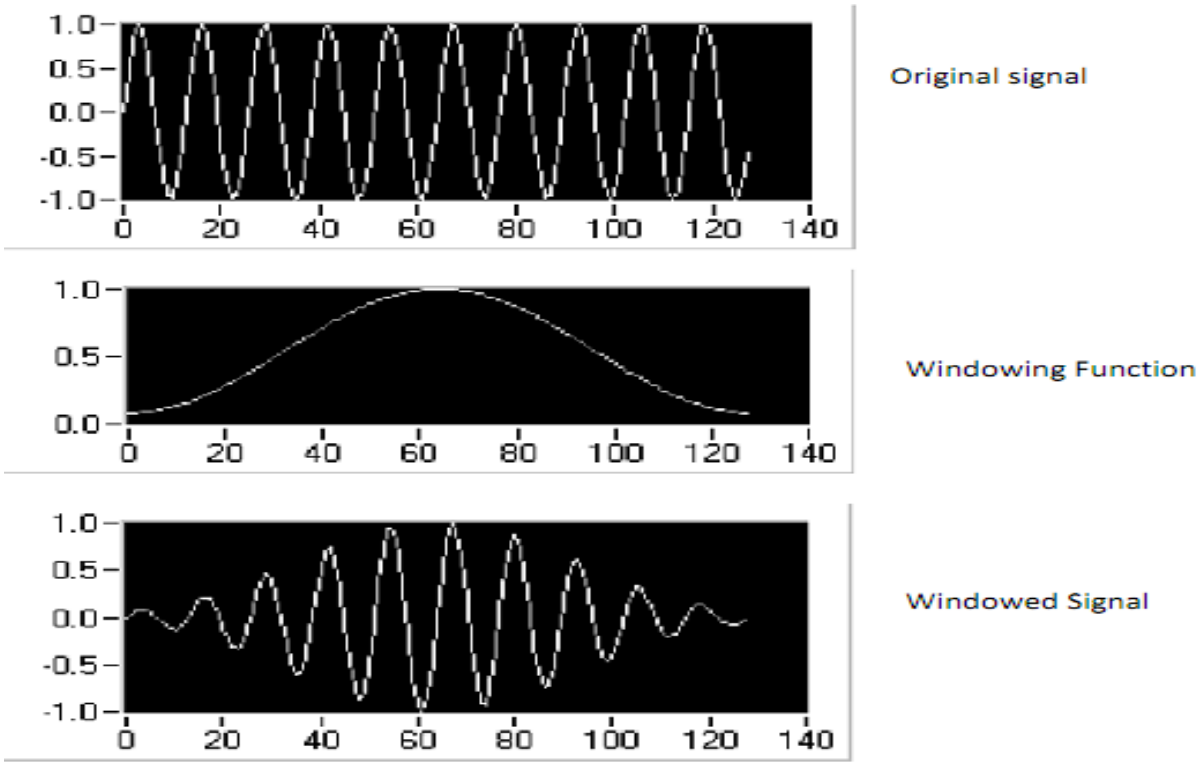


Figure4. 3: Windowing a Signal

The length of the ST signals depends on the pitch period that the pitch and duration modifications are applied on the ST signals (see Figure 4.2). For example, the pitch is raise or lower by varying distance between the short signals. The modification of duration takes place by repeating or deleting the ST signals as necessary. Provided infinite periodic signal, we are able to shift period from original T_0 to required T, by summing windowed data Si(n), originated from X(n) signal shown in equation 1,2 and 3 below..

$$Si(n) = X(n).W(n - i.To) \dots\dots\dots\dots\dots (1)$$

$$S(n) = \sum_{n=-\infty}^{\infty} \mathrm{Si(n - i(T - To))} \quad -------- (2)$$

The samples Si(n) only differ from zero on an interval dependent on recovering factor FR, defined as a ratio of size L of the analysis window w(n) by the pitch period To.

$$FR=L/T_o \qquad \text{---------- (3)}$$

In practice, we choose FR ≈ 2, when the spectrum of Si (n) signal approximates the spectrum of S (n) signal. Then the concatenation process changes the pitch without affecting the formants' frequencies.

As seen below in Figure 4.4, the original signal is divided by the windowing function into small overlapping units and detects the voiced part as interest region (one) and zero for unvoiced speech signal parts.

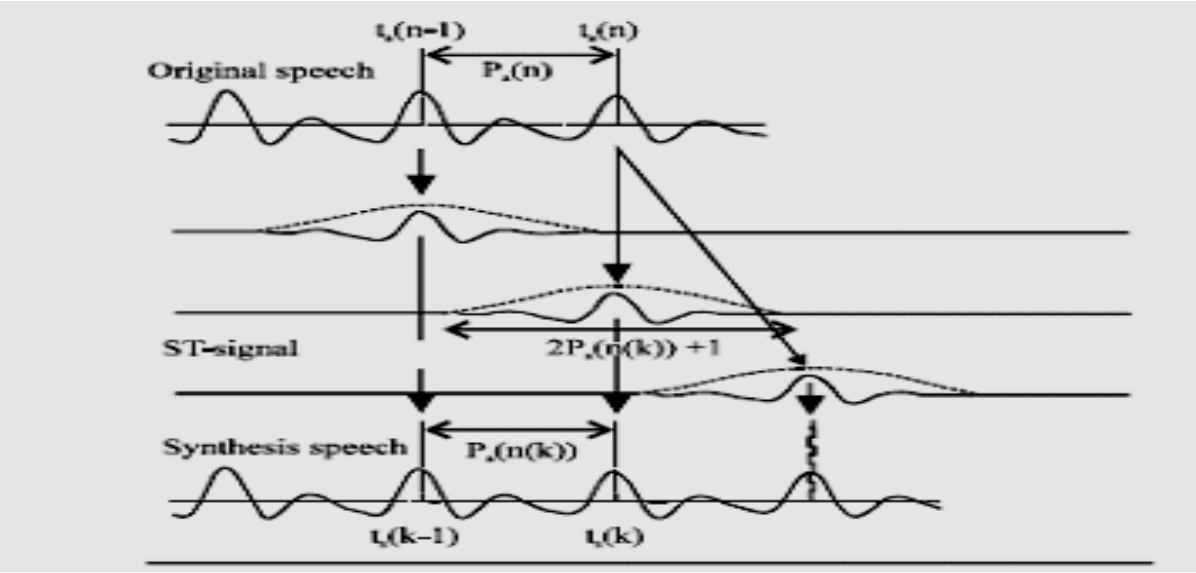


Figure4. 4: Speech Waveform Synthesis

From those variant methods of the PSOLA is the TD-PSOLA which is the widely used for speech waveform analysis-synthesis, which is used in this thesis work as the speech analysis and synthesis algorithm. For example, in the Figure 4.5 below, we can see the effects of the TD-PSOLA method that produce synthesized speech from the original signal using OLA method and windowing function (Hanning) after marking of the pitch and other prosodic attributes.

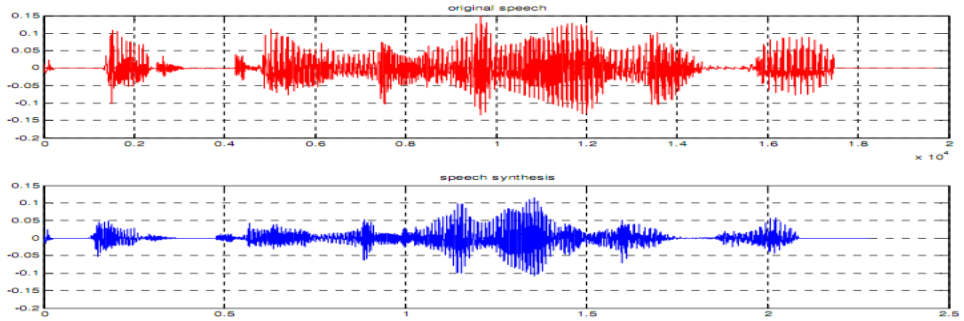


Figure4. 5: A waveform of a sound utterance and its synthesized using TD-PSOLA

4.3.2.3. Time Domain Pitch Synchronous OverLap-Add (TD-PSOLA)

The TD-PSOLA is one of the most used methods due to its simplicity and quality of speech waveform analysis-synthesis. Although it is computationally very efficient, it also requires rather large speech databases in order to produce high-quality synthetic speech. It works pith-synchronously, which means there is one analysis window per pitch period, while data reduction algorithms like Code Excited Linear Prediction(CELP) used in (Chabchoub &

Cherif, 2010) can be employed in order to reduce the size of the database, the complexity of the compression algorithm will increase the number of operations.

The concatenative TD-PSOLA is known to allow for high-quality pitch and time scale modification of speech segments for concatenative speech synthesis (Lin & Jang, 2004). The quality of TD-PSOLA-modified speech largely depends on how the speech signal is split into windowed double period segments. A Time-domain version called TD-PSOLA is the most commonly used due to its computational efficiency and simplicity of real-time implementation (Kortekaas, *et al.,* 1997). In general, the algorithm can be summarized into three broad steps as follows:

- Division of original signal into separate but overlapping ST signals
- Modification of each of the short term signals into a corresponding synthesis signal by adjusting duration and pitch (into appropriate levels).
- Recombination of the short term signals to synthesized desired speech(Synthesized the segments of signals)

As the Maheswari's (2012), discussed that TD-PSOLA algorithm consists of two basic components to perform the analysis and the synthesis of speech signal into segments and finally concatenate each segment into synthesised signal (Maheswari & Rajeswari, 2012). Those components are classified as analysis and synthesis part. The first component that used to analysis speech waveform is the analysis part, which performs mainly the following tasks:

- Detection of Voiced/Unvoiced after getting the acoustic units it is necessary to identify each acoustic unit in to voiced and unvoiced according to the linguistic analysis from the corpus. Since the pitch modification in TD-PSOLA are carried out on the voiced part of acoustic units. The voiced and unvoiced detection takes place by the use of the time domain parameters like the Root Mean Square (RMS) and Zero Crossing Rate (ZCR). The RMS is used to measure energy found in speech signals. Voiced signal have high RMS values than unvoiced, and the ZCR is measures the number of times the signal crosses the zero line per unit of time and detects voiced and unvoiced signals. The ZCR assigns its value low to voiced signals and high for unvoiced signals.
- The next task performed by the algorithm under analysis part is pitch marking. The marking Pitch pulses are the location at the energy peak of the short-term signals. According to (Lin & Jang, 2004) discussed, the pitch pulses are found under the positive or negative extremes of the pitch cycle and states that a pitch-mark (pitch

period) is defined as 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. Pitch synchronous speech synthesis algorithms require the beginning location of the pitch period (pitch-mark) for every voiced segment prior to speech synthesis. The synthesis technique is pitch synchronous, which requires information about where pitch-marks occur in the acoustic signal. Pitch-marks are especially important in prosodic modification algorithms that employ a method known as PSOLA to change the time and pitch scale of a speech signal. An essential part of TD-PSOLA is pitch marking, which tries to find the glottal closure instant (GCI) in order to perform synchronous analysis. If the result of pitch marking is not good, TD-PSOLA will produce low-quality speech. As a result, it is very important to have an efficient and effective pitch marking method, especially for real-time pitch scaling in various applications. For instance in Figure 4.6, shows two pitch-marks (pitch mark 1 and 2) on the short-time speech signal of the vowel /***u***/.

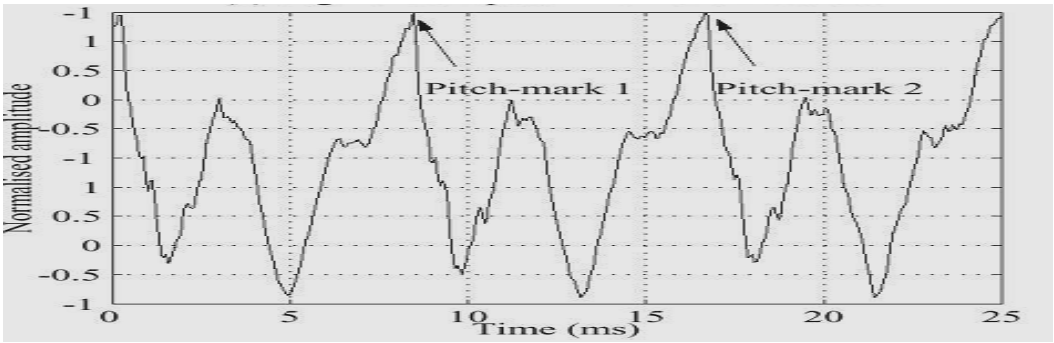


Figure4. 6: Pitch Marks on the short-time speech signal of the vowel /***u***/.

The pitch detection itself was fairly difficult task to accomplish, since the pitch detector worked pretty well on some of the sound files and may not be in other files, it may not be similar at another time in the case of vowel/***u***/ above. The two main problems while processing sound files with broad range of vocals and instruments. More than likely this was caused by numerous fundamental frequencies present in those sound files.

- Another task during analysis of speech signal is storage part, which is used to store the processed data into the database. The speech segments, pitch marks, detection of voiced/unvoiced data should be stored, and that could be used by loading into memory during synthesis part.

The second component that is performed by the TD- PSOLA algorithm is synthesis part. These components takes data stored in analysis part (i.e. required prosodic information) and

then reconstructs speech by concatenating the appropriate speech units. The following tasks are performed during synthesis part of TD-PSOLA.

- ❖ The first task during synthesis part is the selection of appropriate data units that used to concatenate each other by looking the result of the G2P conversion.
- ❖ The other task is synthesizing voiced and unvoiced part of speech. Synthesis of the voiced part involves by overlapping the windowed signal with proper displacement and then adding them properly. The target pitch is determined by the displacement that if the displacement offset is high the pitch will be low and vise versa (Minghui, 2000). In addition, the unvoiced part is synthesized simply by copying it from the database.

As shown in the Figures 4.7, and 4.8, the TD-PSOLA algorithm is done by starting from original speech waveform with epochs, Hanning windows at each epoch to create separate frames, finally use overlap-add method to merge and concatenate the separated speech signals and to find perceptual identical waveform(reconstructed) from the original.

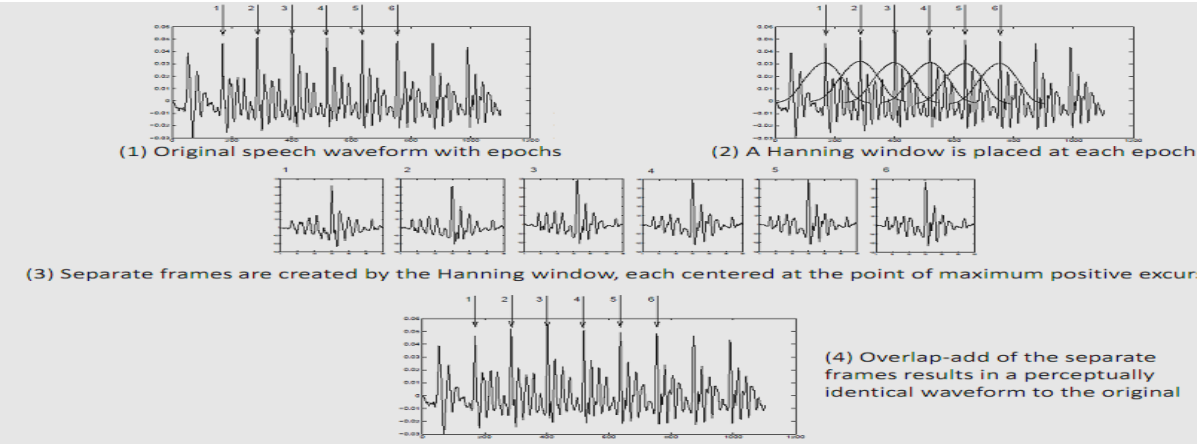


Figure4. 7: TD- PSOLA Analysis and reconstruction (mainly adopted (Taylor, 2009)

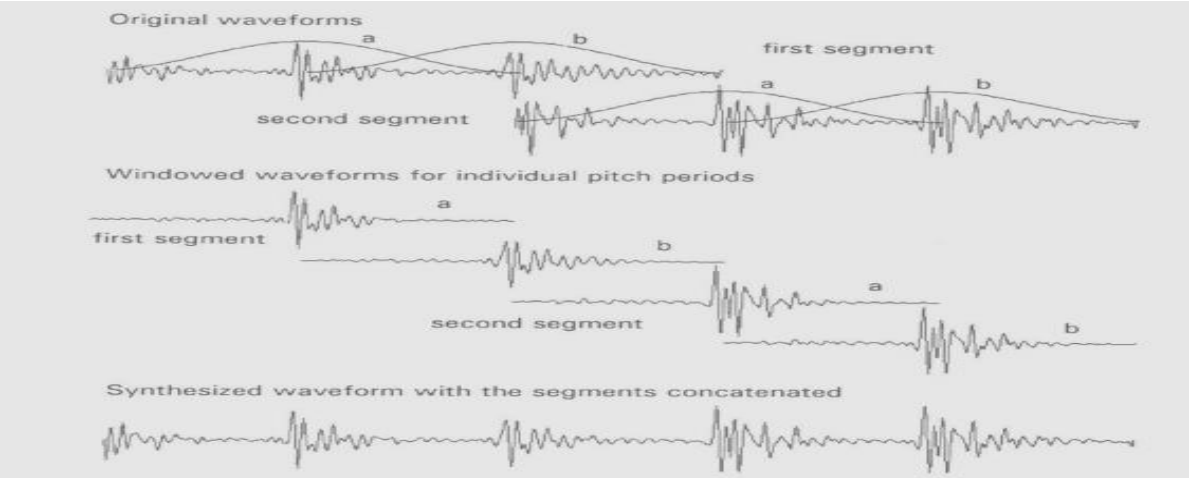


Figure4. 8: Merging two speech signal segments (Taylor, 2009)

Under TD-PSOLA method, the basis of concatenative synthesis is to join short segments of speech, usually taken from a pre-recorded database, and then impose synthetic prosody by appropriate signal processing techniques (Edgington, *et al.*, 2000). Both of these steps can introduce distortion to the synthetic speech: at the boundaries between speech segments (syllables in our case) by inappropriate selection or insufficient merging of segments, and by the prosodic modification process, due to an insufficiently robust speech modification model or creating the prosodic parameters mismatch during analysis-synthesis part.

It also reduces the noise of the sound produced during recording of the sound files, which is one of the basic potential advantages of the algorithm. Let see the effects and how it removes noise from recorded speech waveform. For instance, the Figure 4.9 and 4.10 below shows the task of TD-PSOLA analysis/synthesis before and after applying the algorithm.

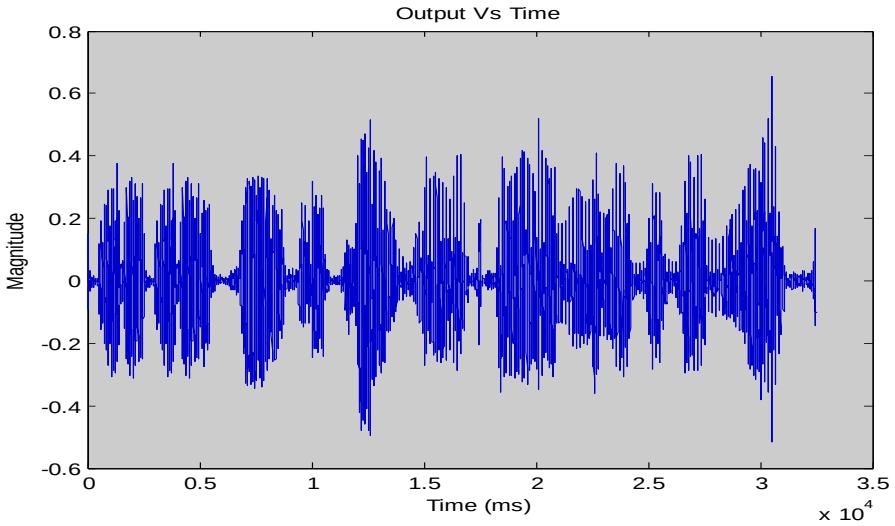


Figure4. 9: Shows before noise with in the speech waveform (“noisespeech.wav”)

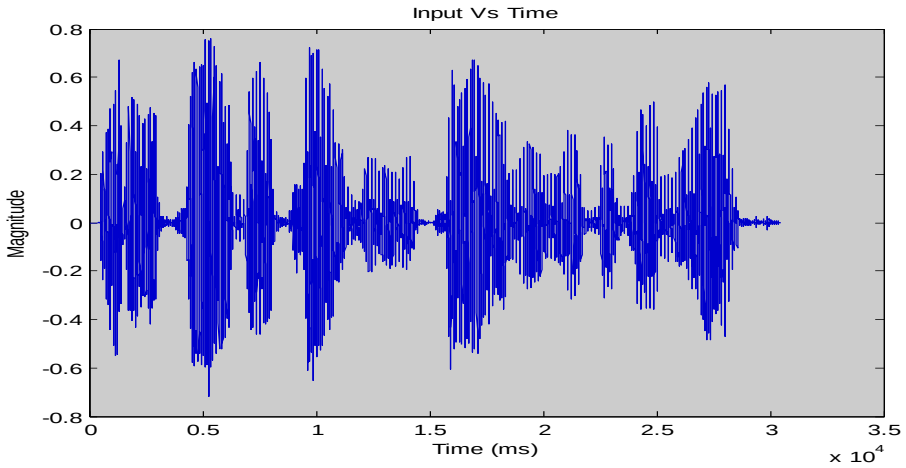


Figure4. 10: The cleaned waveform (“cleanspeech.wav”) after applying using TD-PSOLA

The techniques proposed in this study to perform prosodic modification to come up with minimum distortions of speech signals generated are the following:

- ❖ Pitch-synchronous cross-correlation cost to improve discontinuity at the concatenation points of syllables and the overall quality of the synthesized speech.
- ❖ TD-PSOLA algorithm to stabilize speech quality in prosodic modification(i.e. Time-scale modification modifies the duration of the utterance without affecting pitch and Pitch-scale modification seeks to modify the pitch of the utterance without affecting its duration) and
- ❖ Discontinuous cost function to reduce computational time required during analysis-synthesis of speech waveform.

4.4. Proposed Architecture of Speech Synthesis Algorithm for Amharic

The architecture of the proposed system is shown in Figure 4.17. The components shown are common in most of the speech synthesis systems that use unit selection concatenative approach. The system can be mainly divided into three parts: analysis (front-end), unit selection, and generation (back-end) of speech waveforms. The analysis module is responsible for producing an internal linguistic and prosodic description of the input text. This description is fed into the unit selection module as the target specification. The unit selection module uses this specification to choose the units from the speech database such that a cost function between the specification and the chosen units is minimized. The system uses an internal data structure to store the information for the text to be synthesized.

When we use concatenative speech synthesis where the segments of recorded speech are concatenated to produce the desired output, we apply prosody parameters in which it makes the synthesized speech sound more similar with natural speech. Smoothing techniques is also done to smooth the transition between segments (i.e. syllables) in order to produce continuous output and minimize the discontinuity of concatenation points between speech segment units by using PSOLA methods (Verhelst & Roelands, 1993). In general the PSLOA and its Variants TD-PSOLA method performed in the following steps:

1. The input signal is divided into overlapping segments of fixed length.
2. The overlapping segments are shifted according to a desired time scaling factor
3. The area of overlap intervals are searched for a discrete-time lag of maximum similarity. At the point of maximum similarity, the overlapping segments are weighted by a fade in or

fade out function to eliminate abrupt changes. The segments add together for an audio sample of changed time length.

4.4.1. Major Tasks of the TTS System

In TTS System, the concatenative synthesis approach is used where natural speech is concatenated to give the resulting speech output. It involves two phases, namely, the offline phase and the online phase. Offline phase includes pre-processing, segmentation and pitch marking. Online phase includes text analysis and synthesis this minimizes the memory usage. The syllable based concatenative synthesis comprises of following fundamental components: text analysis, phonetic analysis, prosodic analysis, and synthesis.

The first stage of a TTS system is the pre-processing module, called tokenization and normalization. It converts the input text into a sequence of words and symbols to be processed by the rest of the system. It identifies and makes decisions on what to do with punctuation marks and other non-alphabetic textual symbols (e.g. parentheses), identifies and expands abbreviations, acronyms, and numbers to full- blown orthographic strings. Each input line is scanned and converted into an appropriate word or sequence of words. Generally the text analysis module handles the pre-processing tasks, and responsible for producing an internal linguistic and prosodic description of the input text. After input text is analyzed and transformed into a linguistic representation containing all the necessary information needed in the subsequent TTS steps, the next stage is the phonetic and prosodic analysis, majorly the letter to sound conversion and modification of pitch and duration. At the end the concatenation and synthesis of the segments and speech waveforms is performed.

In this stage the efforts such as finding Amharic pronunciation lexicon, which is a collection of words and their pronunciations and any other word level information are prepared. The speech recordings from a speaker are processed to construct an inventory of speech units (speech corpus) as speech unit inventory. The construction of the speech inventory is an offline process. The unit inventory stores the waveforms for the units, phone identities of the units, phonetic context and prosodic annotations for the units. The waveforms in the speech corpus have been compressed for efficient storage of units using TD-PSOLA method.

4.4.2. Segmentation of speech units

Connecting pre-recorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. However, concatenative synthesizers are usually limited to one speaker and one voice and usually require more memory capacity than other methods. 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. 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. To make the phonation more natural a special algorithm based on the overlap-add (OLA) technique is called pitch synchronous OLA (PSOLA) and built by making some changes in the sequence of the identical acoustic segments, which considers the epenthesis vowel insertion, gemination and syllabification before mapping of grapheme to phoneme. The design of the system also inculcates issues mentioned above and algorithm to analyze and synthesis the prosodic parameters modification of pitch and duration. During segmentation of Amharic words into syllables, handling of epenthesis and gemination is takes place before syllabified the target texts, then after syllabification and the unique syllabified text form is stored as dataset. After wards the syllables in text form are recorded for experimentation and evaluation.

As discussed in chapter three, speech in Amharic language is based on basic sound units, which are inherently syllable units made from six CV-syllable combinations. From perceptual results, it is observed that from different choices of speech units like phone, half phone, diphone and syllable, the syllable unit performs better than all the rest and is a better representation for Amharic languages (Solomon & Menzel, 2007). Since the longer speech segments provides most natural synthetic speech and segment concatenation has also produced reasonably intelligible synthetic speech (Mohanty, 2011).

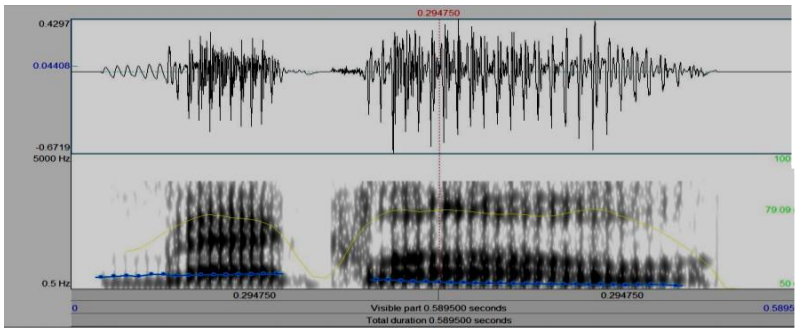
In this work, for Amharic text is segmented into syllables using the basic syllable pattern, which is the most acceptable template structure. The syllable segmentation is done based on the appropriate matching of Amharic syllable template, else further investigation is needed. Such as epenthesis vowel insertion and gemination should be handled properly to come up quality linguistic and speech corpus.

The proposed approach uses six syllables template (V, VC, VCC, CV, CVC, and CVCC) in Amharic language as the basis of segmentation and syllabification of Amharic texts into unique syllabified texts. Both the syllabification and epenthetic vowel insertion algorithm reads input from left-to-right, since any syllable requires a vowel and onset also filled up before coda (Kopecek, 1997). In the next sections, we have seen the rules of the language in relation with epenthetic vowel insertion, gemination and syllabification effects on speech synthesis to determine the pronunciation of given Amharic words based on its spelling, in the process of grapheme-to-phoneme conversion. Most of the rules and procedures of

gemination, epenthesis and syllabification used is mainly adopted from those works of (Mulugeta, 2001), (Tadesse, *et al.*, 2010), and (Nirayo, 2011).

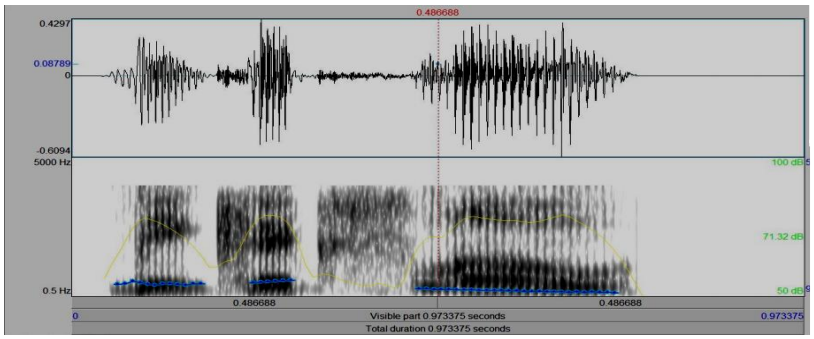
4.4.2.1. Epenthetic Vowel insertion rule

The process of epenthesis is common in Amharic and can occur word-initially or medially. As (Hudson, 2000) stated epenthesis is extensive in word-formation in the Ethiopian Semitic languages, since many morphemes, both roots and affixes, consists only consonants. Amharic epenthesis vowel may be said to provide almost all occurrences of the high central vowel /ix/ (ɨ). As seen in syllabification example below, before epenthetic vowel insertion the word በቅሎ - means "mule" becomes-"beqlo", and -"beqixlo" after epenthesised by inserting/ix/, since at word medial cluster of consonants contains the singleton in sequence, the epenthesis vowel is inserted between the two consonants. When we see the pronunciation and waveform of the above word it is strictly different from the previous version when we analysis the waveform the following information like time domain, time sampling (i.e. number of samples, sampling period and frequency) and other prosodic features (pulse, pitch and duration) is shown in the following Figure 4.11 and 4.12, that the effect of epenthesis vowel insertion in the word በቅሎ- means "mule".



- Waveform of the word-beqlo
- Formants
- Pulse
- Spectrogram
- Intensity

Figure4. 11: Waveform of the word *በቅሎ-"beqlo*", with out effect of/*ix*/



- Waveform of the word-beqixlo
- Formants
- Pulse
- Spectrogram
- Intensity

Figure4. 12: waveform of the word *በቅሎ-"beqixlo*" after epenthetic vowel is applied

In general the epenthetic vowel insertion rule for Amharic is as follows:

1. *Accept input word and scan from left to right.*
2. *If consonant cluster occurs at word initial position, insert epenthetic vowel between them.*

   ***Exception**: If the first phoneme is consonant and the next consonant is glide /w/. (Rule #1)*
3. *If three consonants are appeared in sequence word medially or word final position, insert epenthetic vowel before the third consonant.(Rule #2)*

   ***Exception**: If the middle consonant sonority is greater than the rest insert epenthetic vowel after the first consonant in the cluster.*
4. *If a cluster of consonants contains the geminate and singleton in sequence, insert epenthetic vowel after the geminated consonants.(Rule #3)*
5. *If a cluster of consonants contains the singleton and geminate in sequence, insert epenthetic vowel after the singleton consonants. (Rule #4)*
6. *If a cluster of consonants contains two different geminates in sequence, insert epenthetic vowel between the two geminate consonants. (Rule #5)*
7. *If the sonority of the final consonant is greater than that of the preceding consonant, the epenthetic vowel is inserted between the final consonant clusters. (Rule #6)*
8. *Repeat 2 up to 7 until all the phonemes are parsed in the phonemes list.*

Afterwards the epenthesis of vowel insertion and doubling of consonants (gemination) according to lengthening of it, the next issue is syllabification of those words into syllables. A syllable is a basic unit of written and spoken language. It is a unit consisting of uninterrupted sound that can be used to make up words. The segmentation of words into syllables takes place automatically; a recording is done at word level for syllabified unique text. The syllable count is reduced by finding the syllables that are common for Amharic language. The number of syllables that you hear when you pronounce a word is the same as the number of vowel sounds heard. The number of vowels sounds found is the same as the number of syllables. For instance, the word በቅሎ/*beqlo*/ meaning “mule”, which consist two syllables, because the word contains two vowels i.e. ‘*e’* and ‘*o’*. The waveform also shows as follows in Figure 4.13 below:

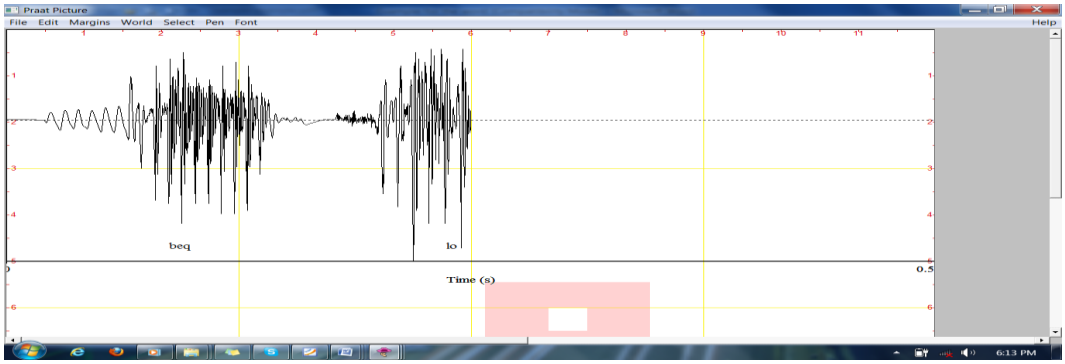


Figure4. 13: the waveform word /*beqlo*/ meaning "mule", having two syllables

4.4.2.2. Syllabification rule

The final output of the epenthesis module directly becomes the input for the syllabification model. The syllabification algorithm reads the given phonemes sequence from left-to-right and the rules are applied repeating the template matching operation for each phoneme sequence in the given word. The syllabification rule for Amharic is as follows:

1. *Accept the input from epenthesis algorithm and scan from left to right.*
2. *At word initial position if two vowels phonemes (VV) occurs in sequence, mark syllable boundary between them.*
3. *If the initial phoneme is vowel and the next two phonemes are consonant and vowels respectively; mark the syllable boundary just at the second*
4. *If (VCCV) pattern occurs at any position, mark syllable boundary between the two consonant clusters.*
5. *If (VCVC) pattern occurs at word initial position, mark syllable boundary before the second vowel.*
6. *If (CVV) type sequence occurs at any position, mark syllable boundary between the two vowels.*
7. *If (CVCCV) phoneme sequence occurs at word initial position mark syllable boundary between the middle consonant clusters (CVC- CV).*
8. *If (CVCC) pattern occurs at word final position and if there is phoneme before the first consonant mark syllable boundary before the initial consonant in this pattern.*
9. *If (CVCV) pattern occurs at any position, mark syllable boundary after the vowels, but if it occurs at word final position the syllable boundary becomes CV - CV pattern.*
10. *If (CVC_1C_1VC or CVCCVC) pattern occurs in a word mark syllable boundary between the geminated consonants. (CVC_1- C_1VC).*
11. *If (VVCC) syllable pattern occurs at word final or initial position mark syllable boundary between the two vowels.*

12. *Repeat 2 up to11 until all phonemes are parsed.*

For Amharic, based on these rules the text is split into syllables and the location of the corresponding sound files for the syllables are written in a text file as shown below in the Figure 4.14. For instance, the word -/*beqlo*/ meaning "mule" after epenthesis and syllabify becomes/ *be-qix-lo*/, having three syllables '*e*','*o*' and the epenthesis vowel/*ix*/. If we see the word above, using the waveform analysis-synthesis tools it becomes different due to the effects of the epenthesis vowel insertion as seen in Figure 4.11 and 4.12 above. The segmentation of Amharic words into syllables is seen in Figure 4.14 below by handling epenthesis and gemination after finding unique syllabified text.

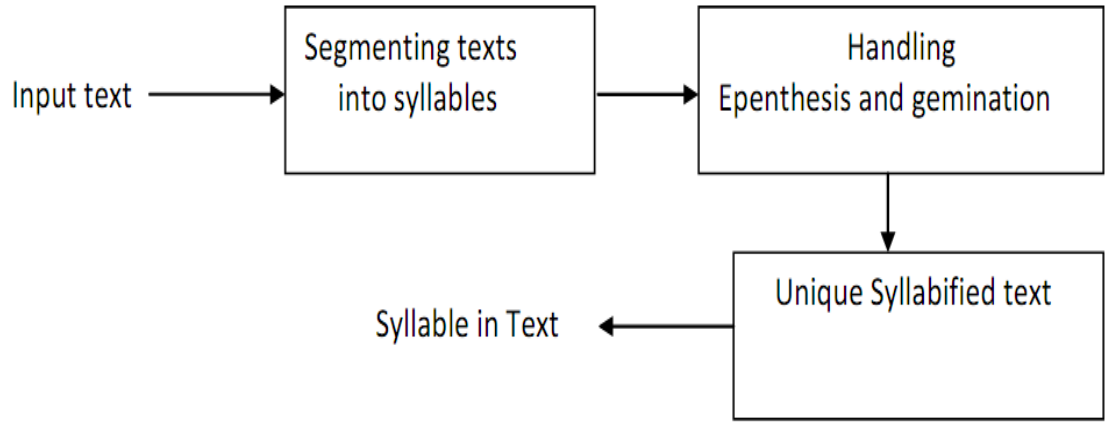


Figure4. 14: Segmentation of text into Syllables

After segmenting and recording the syllables in text, the waveform is seen in the Figure 4.15 below, that the syllables increase and becomes three including epenthetic vowel/*ix*/. The syllable structure becomes, CV-CV-CV, before that it was, and when the epenthesis, and gemination is applied CVC-VC; this shows the effects of epenthesis vowel insertion in the syllabification of Amharic texts. It is necessary to handle gemination and epenthesis to find unique syllabified text as seen Figure 4.14 above.

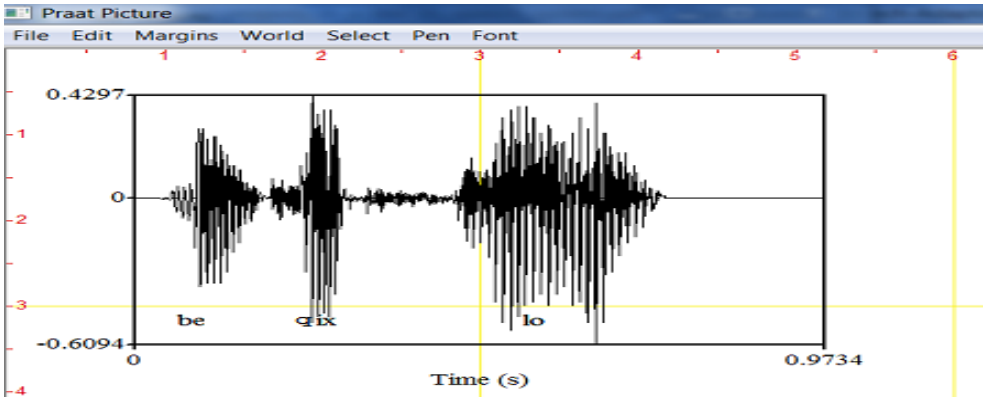


Figure4. 15: The waveform of the word: - '*በቅሎ*' becomes /*be-qix-lo*/

4.4.2.3. Concatenation

The TTS system is produced by concatenating sound units called syllables, to maximize the characteristics of synthetic speech. These sound units are drawn from a carefully designed database that has syllables in variable length with various prosodic and phonetic contexts. Selected speech units are modified according to the predicted prosody and concatenated to form a single speech file in the waveform generation module. This module of the TTS system is responsible for finding the boundary values of a syllable sound in the recorded sound file from the database. The input text that is segmented into words is passed to this module. First search in the database is made to look if the word is present in the database. If it is present, the starting and end positions of word sound in the recorded sound file, are returned from the database, otherwise, syllabification module is called to segment the word into syllables (Lin, *et al.* , 2005). After finding the unique syllabified Amharic text, the recording and storing into syllable inventory database take place. The next is modifying the prosodic parameters on the speech waveform to overcome the mismatch, discontinuity of speech signals.

The recorded syllable in text is fed into the unit selection module to select the required units from the speech database. During speech waveform analysis-synthesis, the waveforms for the required units are selected and then concatenated it in the generation module, where the smoothing of concatenation points is also handled. This module will receive a list of syllable segment that has been properly arranged according to the Amharic raw text. Based on the list of recorded syllable, in the synthesis module, the syllable concatenation is take place according to the sequence and finally plays the sound file, which we know as synthesized speech. The syllables in text form maps to sound files and then apply prosodic modification and concatenate them to find the desired speech output. The concatenation of syllable units is shown in Figure 4.16 below.

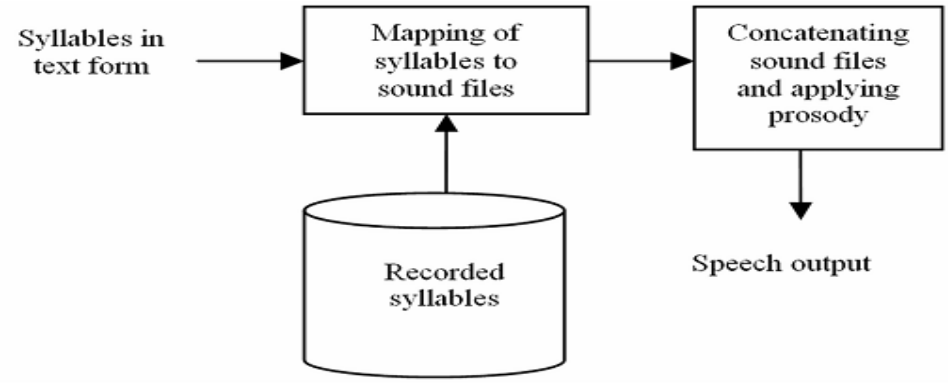


Figure4. 16: Concatenation of Syllables units

The overall structure of the proposed architecture of Amharic TTS system (see Figure 4.17), is inline with the basic functional organization of a general TTS synthesizer. It consists of several modules, each of which has its own tasks and functions. The proposed syllable based concatenative text-to-speech for Amharic is used to synthesise given Amharic text (syllabified) using the modules inculcated under each of steps to come up high quality desired speech output. The system used the syllabication, epenthesis vowel insertion and gemination procedures in linguistic and phonetic analysis, which affects the pronunciation lexicon of the Amharic language.

The capability of automatically determining the syllable boundaries in a word is useful for such applications as G2P conversion and TTS synthesis (Dell & Elmedlaoui, 1985). Since identification of syllables structure of words play an important role in speech synthesis apart from their purely linguistic significance. The pronunciation of a given phoneme tends to vary depending on its location within a syllable.

While actual implementations vary, most state of the art TTS systems must have at minimum three components: a text and phonetic analysis module, a prosody module, and a synthesis module. In phonetic analysis, the G2P conversion is handled. It is a process that converts a target word from its written form (grapheme) to its pronunciation form (phoneme) automatically.

The syllabification task plays a great role in all the above modules (Bartlett, *et al.*, 2009). Moreover, in speech recognition syllabification has been used to build recognizer, which represents pronunciations in terms of syllables rather than phonemes. In addition, syllabification can help annotate corpora with syllable boundaries for corpus linguistics research (Thomas, 2007).

Figure 4.17 below shows the proposed architecture of syllable based concatenative TTS for Amharic. The proposed architecture consists of the gemination, epenthesis and syllabification module which improves the phonetic and prosodic models apart from its linguistic significance.

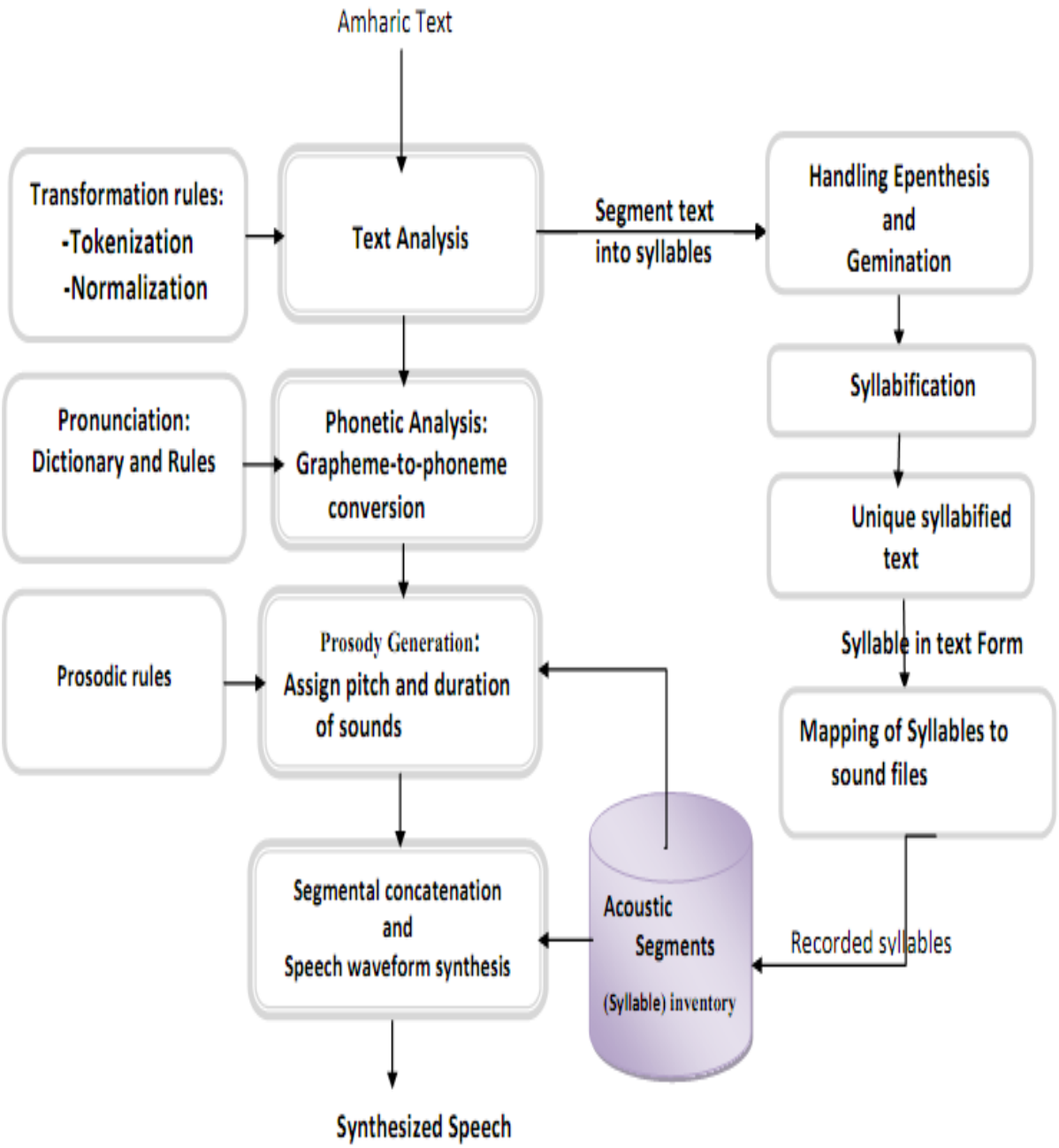


Figure4. 17: Block Diagram of proposed Syllable Based Concatenative TTS for Amharic

4.5. Novelty of the research

The proposed generalized TTS for Amharic in this thesis work handles issues which is necessary to improve the naturalness of waveform generation by the synthesizers', corpus-based concatenative synthesis with syllable speech segment and a TD-PSOLA speech waveform analysis-synthesis algorithm for prosodic modification.

The proposed system handles issues including gemination, epenthesis vowel insertion, and syllabification concepts and rules for them as major issue since they affect the linguistic, phonetic and prosodic model to find quality synthesized speech output. Some of the issues like gemination issue is considered in (Tadesse, *et al.,* 2010) and unit size selection (Sebsibe, *et al.,* 2004) work.

Having rules for those issues it is possible to find quality speech corpus, synthesis it and produce natural sound. When we handled the epenthesis vowel insertion, geminated consonant identification, epenthetic vowel insertion and sonority scale of phonemes are identified. In addition, in the case of syllabification the consonant-vowel parsing, syllable template matching, syllable boundary marking, stress assignment and marker, and syllable weight assignment are marked. Issues mentioned above are crucially important for different linguistic processing and applications especially speech processing i.e. speech synthesis and recognition.

The general quality and naturalness of synthetic voices crucially depends on the building of large databases annotated at multiple levels for the training and testing of prosodic models able to generate adequate rhythmic and intonational patterns. One of the most striking difficulties in the building of new voices in the present framework is that the type of annotation required, since it extremely time consuming and depends on different factors such as speaking style.

In this study, attempts have been made to make the speech natural and intelligible by accepting normalized Amharic texts which syllabified texts as sample speech corpus. Syllables are used as basic speech unit for concatenative speech synthesis process that helps to provide better prosody attributes.

CHAPTER FIVE

5. EXPERIMENTAL RESULTS AND EVALUATION

5.1. Introduction

This chapter discusses the procedures, experiment and evaluation result of the study by starting from speech corpus preparation to synthesis of waveforms. The research aims to design and develop text to speech synthesizer for Amharic language related with naturalness and intelligibility of synthetic sounds. The system is enabled to handle different modules with an integration of the current state of the art of the text to speech synthesis system. Since the integration of the speech and language technology provides an assistive technique for human beings to use computer and other machines. The performance of the system is evaluated using sampled datasets to check the intelligibility and naturalness of the synthesizer based on the users' acceptance test.

5.2. Test corpus description

The thesis test corpus consists of words which are selected from different Amharic dictionary and other books. The corpus prepared for experimenting and evaluating the system contains a total 1000 Amharic words and sentences as datasets for this thesis experiment and among them around 60 words and 10 sentences are selected to evaluate the system performance. Each word contains an average of three up to four syllables that makes speech synthesis of variable size unit selection. This test corpus contains around a total of 3,025 syllables, CV-syllable structure including six order CV-syllables. Corpus preparation is very important for speech analysis and synthesis since the quality the test corpus matters the quality of speech produce by the synthesizer. For speech synthesizer, text corpus of different size, syllable in text form is selected; have to be recorded, analyzed, and studied using speech analysis-synthesis tools.

Since in TTS system based on concatenative synthesis needs well arranged speech corpus. The quality of synthesized speech waveform depends up on the number of realization of various units present in the speech corpus. For concatenative TTS, the quality of speech corpus is very important because the characteristic of synthesized speech are directly related to nature of speech corpus. The accuracy of the system is evaluated in the ways of naturalness and intelligibility of the synthesized speech. In this work, a male voice is used for recoding Amharic syllabified words. The sampled speech corpus is analyzed with respect to: the quality of the synthesized speech, variations in natural prosody, and the perceptual distortion with respect to prosodic and spectral modifications.

5.3. Data Preparation

The data preparation is the first step in the experiment that is required in the development of the TTS synthesis system. The data is prepared by doing different pre-processing techniques among them the epenthesis vowel insertion in the six order CV-syllables, gemination based on lengthening of consonant, and syllabification are the one which takes place during preprocessing of the selected Amharic words. The above issues are handled automatically using Microsoft Visual Studio environment, programming language. The quality of the synthetic speech produced by a corpus-based synthesizer depends, to a large extent, upon the suitability of the speech inventory to represent the variability of the language within the target application domain.

Text input to the synthesizer can be in transliterated form or in UTF-8 form and the major selected rules are applied on it to find the required dataset, syllabified in text. The Text processing module consists of preprocessing, epenthesis, gemination, and syllabication as a module. The text in transliterated form is preprocessed to come up in to equivalent Amharic text. And also, preprocessing module adds syllable boundary and markers. The preprocessed text is further passed on to the syllabication module. This module output the syllable form of the input texts in CV-syllable assimilation.

In the development of text to speech system, Amharic speech synthesizer is one of the major tasks which need high attention during construction of syllable text form database. Since before performing to construct the syllable based database there is prerequisite which needs to handle properly like epenthesis vowel insertion, gemination, and digraph replacement. The above mentioned rules play a great role when we syllabify and pronounce Amharic texts.

Transcription systems are helpful in labeling the prosodic elements of a particular language (Hussain, 2001). They allow users to transcribe the most important aspects of prosody symbolically in a tiered structure aligned in time with the waveform of the utterance. The recorded speech data is segmented into the phonemes, syllables, and pauses and a measurement of the prosodic information like pitch, duration and pulse of each of these speech elements is made using speech analysis tool called PRAAT.

5.4. Acoustic unit inventory Design

For this research work, the acoustic units (syllables) are selected during acoustic unit inventory design process stage because syllables can easily show the linguistic and prosodic information for given Amharic text. During acoustic unit inventory, the most expressive

elements such as pitch accent, syllable duration, and stress assignment are found in syllable construction.

During syllable based database construction the input text are first transliterated in to corresponding symbols taking into consideration the epenthesis vowel/*ix*/, gemination and syllabification to find the required datasets according to the linguistic rules. As we discussed in the previous chapters, Amharic language consists of 28 consonants and seven vowels and the orthographic form found in the assimilation of the consonant with vowel to communicate with it. The combination CV gives the six orders (Laine, 1998), with totally 28*7=196 basic CV-syllables to represent the Amharic writing system. On the other hand there are about 39 phonemes in Amharic and each phoneme combination provides variable length of syllable. In the combination rule there may not be exist in Amharic language for instance there is no vowel-vowel (V-V) combination, which is found in another language.

Therefore, having the entire syllable (CV-syllable) is out of question since there is time limitation to label each and every CV-syllable combination. From total of 3,025 CV-syllable structures, around 142 are considered in this experiment for evaluation of the system accuracy and performance. Syllables are extracted from prerecorded word utterances, and concatenative speech synthesis exploits recorded speech that forms the content of the speech corpus (Szklanny & Wójtowski, 2008). The syllables constituting the acoustic inventory database used for this experiment is given in Appendix 1.1.

After the corpus is prepared and recorded, the next stage is to extract and normalize the acoustic units from the corpus. The extraction is performed using any signal analysis/synthesis tool, PRAAT is one of the most widely used to manipulate, and label waveform of the recorded words, and subsequent extraction of waveform of the recorded texts is done manually using this tool. When we recorded the sampled speech corpus the effects of epenthesis, gemination, and syllabification should handled properly, since they play a great role on the waveform generation. For instance, effects of epenthetic vowel/*ix*/ in the word *"ልብ"* when omitting it- "*lbb* " and having /*ix*/ becomes- "*lixbb***"** and gemination effect on the word "*ገና*" when omitting becomes "*gena*" means "yet" and "*genna*" means "Christmas" when it geminate the '*n*' consonant*/ተናግሯ/* create variation of waveform generated of the words(see Figure 5.1). Gemination happen and identified when a spoken consonant is pronounced longer period of time than a short consonant, and occurs frequently in Amharic texts. It is distinct from stress & may appear independently of it, i.e. doubling of consonants principle.

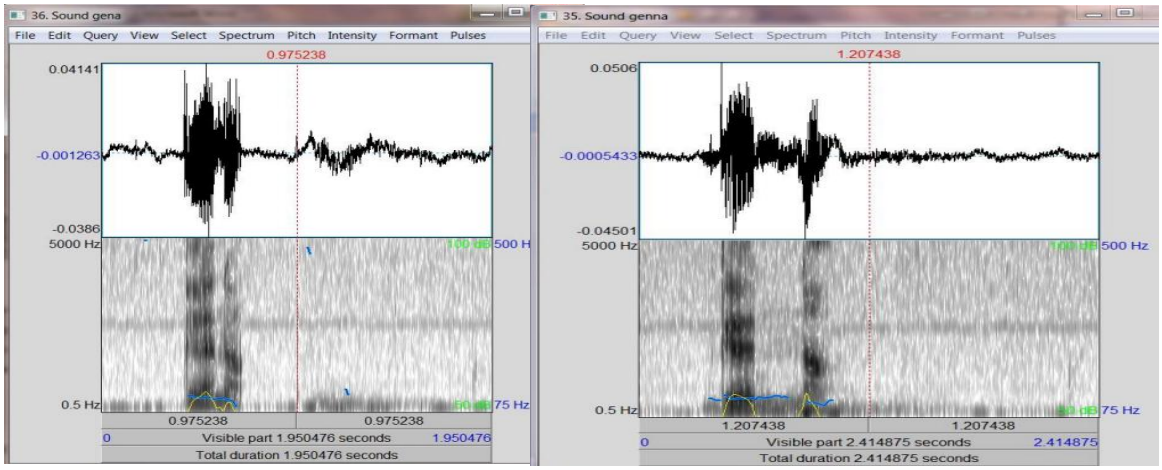


Figure5. 1: Shows the waveform of the word "*gena*" and "genna" respectively

In general such factors and effects mentioned above must be handled so as to avoid degradation effects happened which affects during analysis/synthesis of the speech waveform. Some of the steps like recording corpus, acoustic unit feature extraction, and normalization of units have been done by using PRAAT software tool.

5.5. Syllable transcription

In the transcription process, the Amharic symbols which are equivalent with the Latin are transcribed in to corresponding symbols to represent the sound of Amharic texts from equivalent phonemes and in the form of CV-syllable assimilation. During transcription of Amharic texts in to its equivalent syllables, the pre-processing tasks epenthesis vowel insertion, gemination, and digraph replacement is performed accordingly. Taking into consideration that gemination occurs when pronounce the consonant longer period of time and epenthesis is based on the consonantal cluster with respect to position of it in the words sequence. For instance, the Amharic word "*ክፍት*"-/*kift*/**,** when it is transcribed simply and after preprocessing rules it becomes- /*kixf-fixtt* / which is uniquely syllabified text; the '–' shows the syllable boundary of the text during syllabification and separate subsequent syllable representatives. When we analysis the waveform of the above word -"*ክፍት*", it is completely different without and with effects on those rules. The Figure 5.2 and 5.3 below shows the above mentioned effects briefly. Those above effects create even the meaning of the word completely for instance, as seen in the previous example the word '*ገና*' becomes "*gena*" means "yet" and "*genna*" means "Christmas".

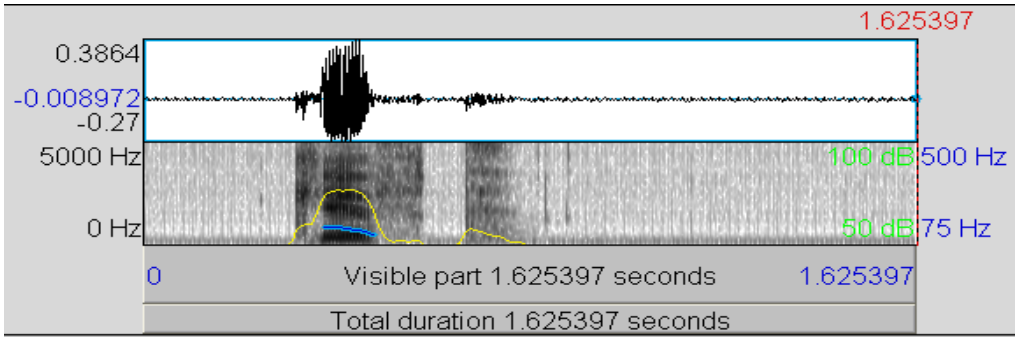


Figure5. 2: Waveform for the word *(ክፍት)* without gemination effect /*kixft*/

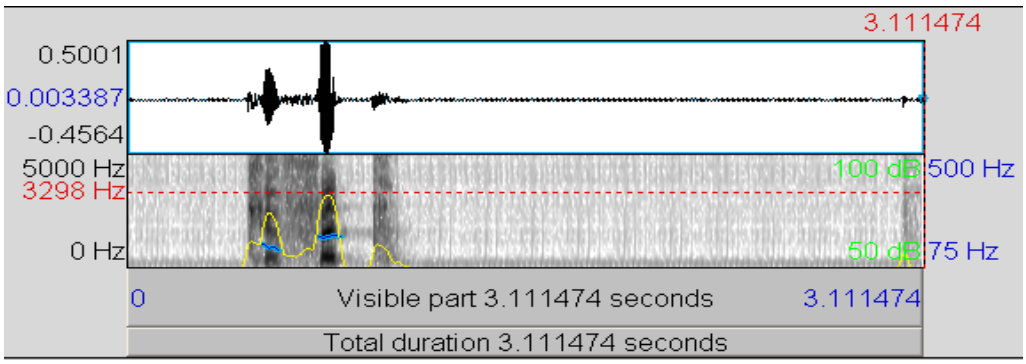


Figure5. 3: Waveform for the word *(ክፍት)* with gemination effect /*kixffixtt*/

Strictly speaking the term syllable might be more accurately applied only after transcription to phonemes. However, we shall use it here to apply to such pronunciation units described orthographically. The purpose of such analysis is to obtain information which is used by the phonetic transcription stage to make better judgments on the pronunciation of consonant and vowel clusters in particular.

Consider the Figure 5.2, which shows the sound waveform and the pitch contour epenthesized Amharic word *"ክፍት"-"kixft"***,** the sound is uttered and syllabified it become *"ክኧፍት"*, the most prominent syllable in this utterance is the syllable (*kixft*) of *kixft*. The pitch contour shows a rise that peaks at the end of */kix*/- and then falls. In the results we have seen so far, an F_0 peak or valley occurs on the accented syllable. Since, syllables are the basic prosodic elements; they carry prosodic attributes of pitch, duration, and intensity (energy).

5.6. Recording the corpus

Once the corpus dataset is prepared, pre-processed and designed the expected acoustic unit inventory, the next task is recording the corpus data from the CV-syllables that are extracted during their assimilation of the consonant-vowel. Annotation and segmentation of Amharic texts into CV-syllables is based on the following major significant rules, i. e. the epenthesis, gemination, and syllabification. The recorded speech have to be marked prosodically and for baseline voices the speech recordings are completely phonetically transcribed and manually

checked listening to the real recordings. The researcher only is record as sampled data using PRAAT tool, each instance of the units is stored along with linguistic and prosodic features, phonetic context, and syllable position in the word. The speech corpus recording, labeling, and parameterization are critical tasks related to the synthetic speech quality. The acoustic speech units have been extracted from a corpus recorded at a sampling rate of 8000 sample/second, 16 bit per sample of bit resolution and mono sound level, which is suited for speech reinforcement as they can provide excellent speech intelligibility.

In addition to recorded, labeled, and parameterized, it is possible to perform the following tasks using PRAAT, i.e. analysis of periodicity (peaks and periods), manipulation using pitch, pulse, and duration of the recorded speech waveform. For instance, the waveform labeling and segmentation the word: - ሰበረ-*"sebbere*" has shown in Figure 5.4 below.

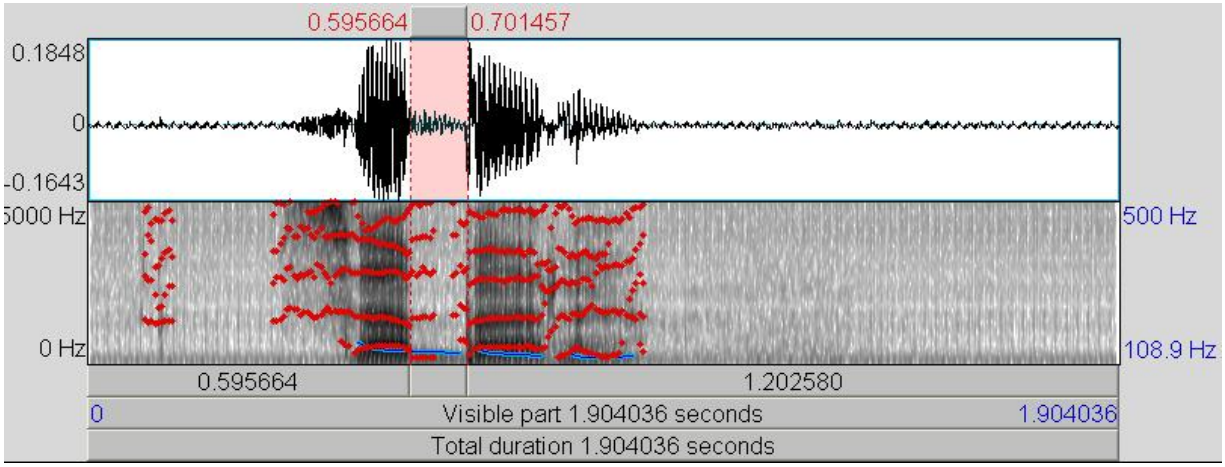


Figure5. 4: waveform labeling and segmentation the word: - ሰበረ-*"sebbere*"

5.7. Speech waveform synthesis

Once a given input text is transcribed, the next step is to synthesize the speech waveform by concatenating the appropriate acoustic units. During waveform synthesis there are parts that the selection of appropriate units from the constructed acoustic unit database and the other one is concatenating the selected acoustic speech units with the application of prosodic modifications (i.e. pitch and duration). In case of concatenative the actual short segments of recorded speech that were cut from recordings and stored in an inventory ‘‘voice database’’, either as ‘‘waveforms’’ (uncoded), or encoded by a suitable speech coding method, called TD-PSOLA.

At time of synthesis, first the system load all the acoustic speech units and other basic information like pitch mark data into pc RAM, during movement of the data from the hard disk to memory is very fast especially at the time of searching when the acoustic inventory is

very large. The speech signal can be stored in a highly compressed (i.e., coded) form so that a large voice database can be used even under tight memory limitations and an ideal speech representation.

The acoustic units are stored in the wave file format (.wav) on the hard disk. During implementation a MATLAB script is used to read the wave data to memory. During waveform analysis using MATLAB figure analysis, the vector elements are represent amplitude values of the wave file at specific times with sampled frequency and period. The required speech signals are sampled and then the arrays of sampled data are concatenated in proper order. Given the name of the file, the MATLAB wave reading script (wavread ()) returns a column vector representation of the wave file. Resultant sampled data is then converted back into speech signal, thus giving the required output in waveform. For instance, the Figure 5.5 below shows the resultant sampled data of the word ሰበረ -*"sebbere*".

To establish the speech waveform signal using Matlab it is necessary to set the number of points use to construct each signal called sample frequency (FS), sample period (Ts) calculated as 1/Fs, time window (number of samples) to know how many points are in one second of our signal, sample index T_n (time index) and establish the signal X (t) and plot the required waveform by reading from recorded speech corpus file (i.e. sebbere .wav).

The Matlab code that generated the waveform of the syllabified word ሰበረ -"*sebbere*" form the recorded speech corpus stored in the computer memory is seen in Appendix 1.4.

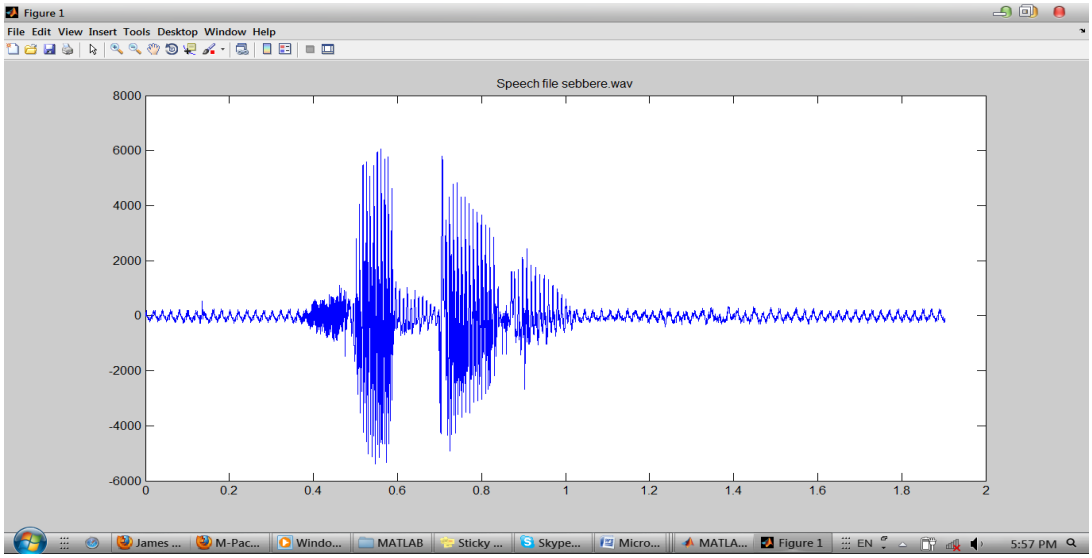


Figure5. 5: The Waveform generation of the word ሰበረ -*"sebbere"* Using Matlab

After speech waveform generated, the following information of the sound file "sebbere.wav", are also displayed: duration = 1.90404 seconds, sampling rate = 8000 samples/second, and bit resolution = 16 bits per sample.

A Matlab script has been used to extract the location of pitch marks of each acoustic unit in the acoustic unit inventory is seen in the Appendix 1.4. The script takes as an input the column vector representation of the acoustic units (*seb, be, re*) and sampling rate (Fs), and Using a Matlab function called TD-PSOLA to determine the duration and pitch of the speech signal. This MATLAB function calculates and returns the pitch marks (placed at peaks in the short-time energy function) for the input speech, that is assumed to be sampled at 8000Hz. The output is the location of the pitch marks in the form of a row vector. The pitch mark identification is one of the parts of the data preparation that location of the energy peaks of the short-term signals and also used to detect unvoiced/voiced parts of the acoustic unit, mostly voiced are vowels.

5.7.1. Acoustic unit selection

Given the speech data and the phonetic transcription of Amharic text the acoustic unit selection process is straight forward and simple. After selecting the appropriate selection of the acoustic units the next is comparing the name of each acoustic unit with each transcription. For example, the syllable based synthesis, acoustic units of the word *-ተሰበረ / te-seb-be-re/* becomes with the name **‘***te’, ’seb’, ‘be’***,** &*‘re’* are extracted and selected. Then by applying the signal processing technique on the selected units the final step is concatenate them to form syllable in text form, and become **‘***tesebbere’*.

5.7.2. Concatenation of units

The last step in the synthesis part is the concatenation of the synthesized units from the acoustic inventory. The processed acoustic units are concatenated in the form of vector representation using Matlab. The concatenated units are stored in a temporary file and finally the waveform is played whenever it is necessary. This module receives a list of syllable segment that has been properly arranged according to the raw text. Based on the list of syllable, the syllable concatenation module will concatenate the sound according to the sequence and finally play the sound which we know as synthesized speech.

TD-PSOLA is currently one of the most popular concatenation methods. Although it provides a good quality speech synthesis which are related to its non-parametric structure; spectral mismatch at segmental boundaries and tonal quality when prosodic modifications are applied on the concatenated acoustic units, even though it tries to resynthesizing voiced parts with constant phase and constant pitch to overcome the concatenation and smoothing problems.

During implementation of syllable based concatenative synthesis for Amharic the following basic elements and tasks must be performed to accomplish the TTS system. Quality of

synthesized speech which is highly dependent on the corpus of recorded speech used to create the syllable database, large number of syllables inventory with variable size needed for unit selection TTS to produce natural-sounds speech. In general the Figure below shows the overall process starting from pre-processing of the Amharic texts into syllables using syllabification algorithm up to the concatenation of the appropriate speech units to generate the speech waveform. As seen in Figure 5.6, 5.7 and 5.8 below, the segmentation of recorded speech, concatenation of individual segment units and generation of waveform respectively.

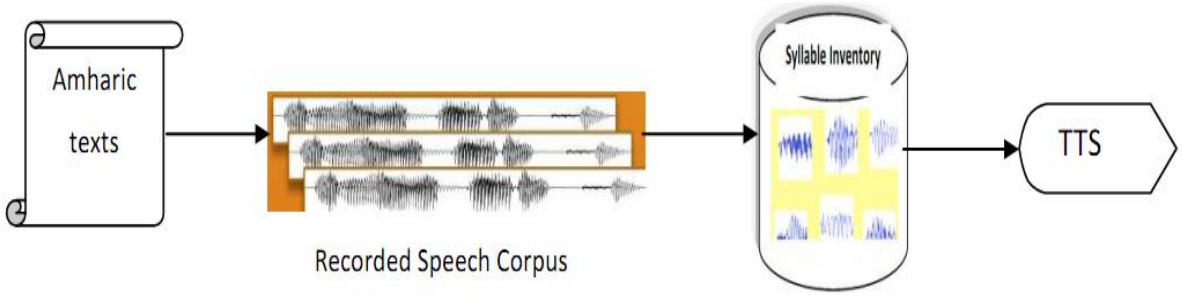


Figure5. 6: Elements of syllable based concatenative synthesis

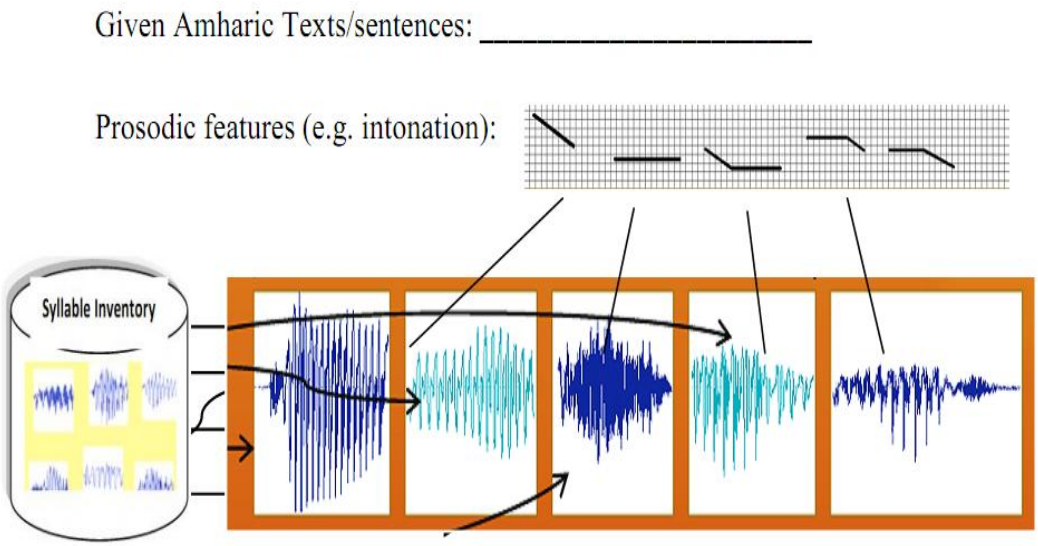


Figure5. 7: Segmentation of recorded speech into audio speech segments (units)

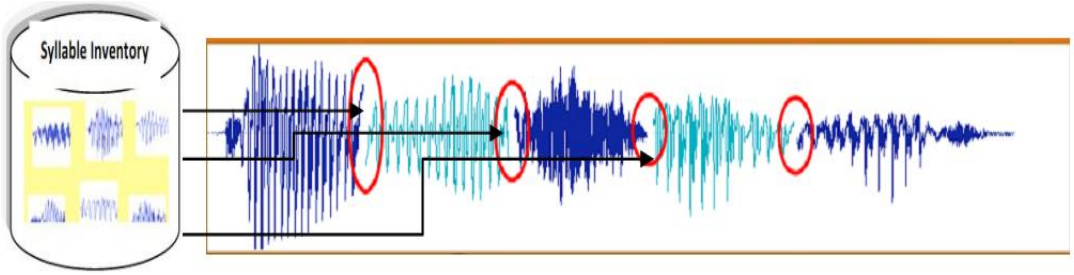


Figure5. 8: During Concatenation of Individual segment Units

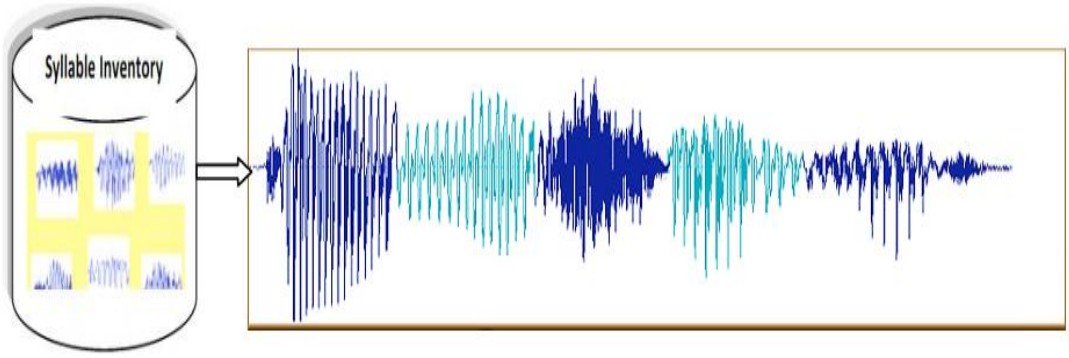


Figure5. 9: concatenating segment units and generate speech waveform

The flow chart shown in Figure 5.10 below is the proposed generalized TTS system for Amharic starting from accepting normalized texts up to synthetic waveform generation.

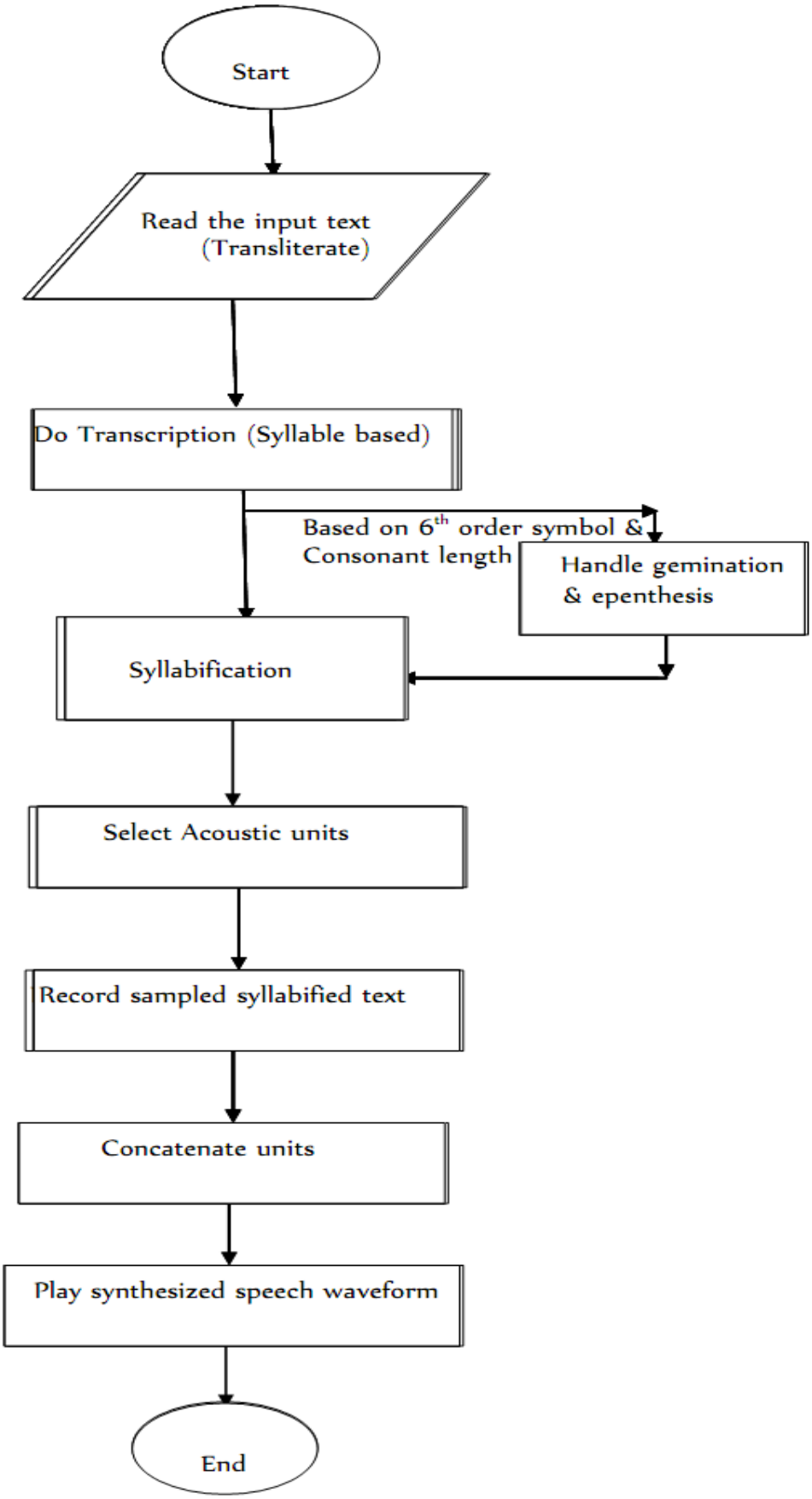


Figure5. 10: Flow chart of the Amharic TTS synthesis

5.8. Text-To-Speech System Evaluation

The main goal of this evaluation test, which was designed according to existing methods for evaluations of the TTS-systems, is to determine how much of the spoken output one can understand and become natural. It is very difficult to evaluate speech synthesis systems consistently because there is no subjective criterion and usually different organizations and projects use different speech data.

The global evaluation and the evaluation of the other modules (prosody and acoustic synthesis) mainly rely on subjective tests conducted by human judges. A typical subjective evaluation procedure is as follows: test sentences (input text) are processed by the system, resulting synthesized speech excerpts are collected, and subjective judgment tests are performed by human listeners for synthesized synthetic speech. Subjects are asked to rate the quality of the synthesized sentences they listen to, according to a series of pre-defined criteria (naturalness, intelligibility, pleasantness, etc.); the TTS systems or modules under scrutiny are compared based on these scores.

Once the Amharic speech synthesizer model is developed, the objective of the test used to test if there are any marked improvement in intelligibility and naturalness of the uttered speech. To evaluate the naturalness or quality of synthesized speech, the widely used testing mechanism called Mean Opinion Score (MOS) and Open Rhyme Test (ORT) for intelligibility or understandability of the speech waveform by playing them back. The two evaluating methods are used to assess the relative closeness of the synthesized speech utterance to natural speech.

For ORT and MOS performance test four users were selected to simply listen to the 60 transcribed only and syllabified words and 10 sentences. The four users are invite to assess the system and provide their assessment in written form by filling and marking whether they are understand/√/ or not understand/×/ for the case of ORT and scaling the quality of speech after they listen the sentences for MOS. The users' acceptance test result on selected words and sentences are given below in table 5.1and 5.2 for ORT and MOS respectively.

Words	Word without effects(transliteration)	Person1	Person2	Person3	Person4
ሰፊ	*safi*	√	×	√	√
ሸፍታ	*sxifta*	√	×	√	×
ገና	*gena*	√	√	√	√
ለጋ	*lega*	√	√	×	×
ይሰማል	*yisamal*	√	√	√	√

ክፍት	kft	×	×	×	√
ወንጌል	wengel	√	√	√	√
ትምህርት	tmhrt	×	√	×	√
ትንሽ	tnx	√	√	×	×
መንግስት	mengst	√	√	√	√
ምልክት	mlkt	√	√	√	×
ኮከብ	kokeb	√	√	√	√
ሀብታም	habtam	√	×	√	√
ገባ	geba	√	√	√	×
መለከት	mlkt	×	√	√	×
ደብር	debr	√	√	√	√
ተሰበረ	tesebere	√	×	√	√
በቀለ	bekele	√	√	√	√
ትልቅ	tlk	×	×	√	×
አህጉር	ahgur	√	√	×	×
ሁከት	huket	√	√	√	√
ትርፍ	trf	√	√	√	√
ክረምት	kremt	×	√	×	√
በላ	bela	√	×	×	√
ረዳ	reda	√	×	√	√
አድባር	adbar	√	√	√	√
መልክሽ	melkish	×	√	√	×
ድንግል	dingil	√	√	√	√
ጥበብ	tebiib	√	√	√	√
ብልሃት	blhat	√	√	×	×
ዝናብ	znab	×	√	√	√
ወደደ	wededе	√	×	√	√
አሰበ	asebe	√	√	√	√
ሃብት	habet	√	√	√	√
ልብ	lbb	√	×	×	√
ተመለሰ	temelese	√	√	√	√
ፈለገ	felege	√	√	√	√
አምላክ	axemlak	√	√	×	×
ባህል	bahl	√	√	√	√
አለመ	alleme	√	×	√	×
አሰበ	assebe	√	√	×	×
አሰረ	assere	√	√	√	√
አስተማሪ	astemarii	√	×	×	√

በር	berr	√	√	√	√
በራ	berra	√	×	×	√
ብልህ	blh	√	×	√	√
ደመረ	demmere	√	√	√	√
ድህነት	dhnnet	×	√	√	×
ፈካ	fekka	√	√	√	√
ፈለገ	fellege	√	×	√	√
ሃያል	hayyal	√	√	×	×
እስራት	ixssrat	×	√	×	√
ከዳ	kedda	√	×	√	√
ለመለመ	lemelleme	√	√	√	×
ሰባበረ	sebabbere	√	√	√	√
ጠቀመ	takkeme	√	×	×	√
ወንጀል	wenjel	√	√	√	×
ፀሎት	xxelot	√	√	√	√
ፀና	xxenna	√	√	×	×
ሰብከት	sbket	×	√	×	√
Words	Transcribed and Syllabified Words				
ሰፊ	sa-ffi	√	√	√	√
ሸፍታ	sxif-fita	√	×	√	√
ገና	gen-na	√	√	√	×
ሊጋ	leg-ga	√	√	√	√
ይሰማል	yis-sem-mal	√	√	√	√
ክፍት	kixf-fixtt	×	√	√	√
ትምህርት	tixm-hixrt	×	√	√	×
ትንሽ	tix-nixnx	√	√	×	×
መንግስት	men-gixst	√	√	√	√
ምልክት	mix-lixk-kixt	√	√	√	√
ኮከብ	ko-keb	√	√	√	√
ወንጌል	wen-gel	√	×	√	√
ሀብታም	hab-tam	√	√	√	√
ገባ	geb-ba	√	√	√	√
መለከት	mix-lix-kixt	√	√	×	√
ደብር	debr	√	√	√	×
ተሰበረ	te-seb-be-re	√	√	√	√
በቀለ	bek-ke-le	√	√	√	√
ትልቅ	tixlk	√	√	×	√
አህጉር	ah-gur	√	√	√	√
ሁከት	hu-ket	√	√	√	√

ትርፍ	tixrf	×	√	√	√
ክረምት	kix-remt	√	√	√	√
በላ	bel-la	√	×	√	√
ረዳ	red-da	√	√	√	√
አድባር	ad-bar	√	√	√	√
መልክሽ	melk-iixsh	√	√	√	√
ድንግል	din-gil	√	√	√	√
ጥበብ	te-biib	√	√	×	√
ብልሃት	bixl-hat	√	×	√	√
ዝናብ	zix-nab	×	×	√	√
ወደደ	wed-de-de	√	√	√	√
አሰበ	as-se-be	√	√	√	√
ሃብት	habt	√	√	√	√
ልብ	lixbb	√	√	√	√
ተመለሰ	te-me-le-se	√	√	√	√
ፈለገ	fel-le-ge	√	√	√	√
አምላክ	axem-lak	√	√	√	√
ባህል	ba-hixl	√	√	√	√
አለመ	al-le-me	√	√	√	√
አሰበ	as-se-be	√	√	√	√
አሰረ	as-se-re	×	√	√	√
አስተማሪ	as-te-ma-rii	√	×	√	√
በር	berr	√	√	√	√
በራ	ber-ra	√	√	√	√
ብልህ	bixlh	√	√	√	√
ደመረ	dem-me-re	√	√	√	√
ድህነት	dix-hixn-net	√	√	×	×
ፈካ	fek-ka	√	×	√	√
ፈለገ	fel-le-ge	×	√	√	√
ሃያል	hay-yal	√	√	√	√
እስራት	ixs-six-rat	√	√	√	√
ከዳ	ked-da	√	√	√	√
ለመለመ	le-mel-le-me	√	√	√	√
ሰባበረ	se-bab-be-re	√	√	√	√
ጠቀመ	tak-ke-me	√	√	√	√
ወንጀል	wen-jel	√	√	×	√
ፀሎት	xxe-lot	√	√	√	√
ፀና	xxen-na	√	√	×	√
ሰብክት	sixb-ket	√	√	√	√

Table5. 1: Result of the ORT test for simple transcribed and syllabified texts

Key:	quality of speech	mark
	Understand (Perceptible)	-------- √
	Not understand(less perceptible)	------- ×

The MOS technique used to evaluate the naturalness of synthesis system using five evaluation standards is shown in table 5.2 below. Based on the scale the users give their perception evaluation result by listened the selected sentences and average of the opinion is taken as the performance of the system (i.e. naturalness).

Sentences	*Person1*	*Person2*	*Person3*	*Person4*
አበበ መጣ	*4*	*4*	*5*	*4*
መቼ ይመጣል?	*3*	*5*	*3*	*4*
ሰይስት ኪሎ እንጎናኝ	*3*	*4*	*4*	*4*
ሰንት ሰዓት ነው	*4*	*2*	*3*	*3*
ወደ ቤት ልሂድ	*4*	*3*	*4*	*3*
በዘግታ ያሸከርክሩ	*3*	*3*	*3*	*2*
ወደ ቤተክርስቲያን ሂዱ	*3*	*4*	*4*	*3*
መጽሐፍ አምጣልኝ	*4*	*3*	*5*	*4*
ስብሰባው መቼ ነው	*3*	*4*	*3*	*4*
ገና አልደረሰም	*2*	*3*	*3*	*3*

Table5. 2: The MOS score for syllable based synthesis

Hint: to evaluate the sentences quality used the following standards

Key: quality of speech		score
Bad	----------	1
Fair	----------	2
Good	----------	3
Very Good	----------	4
Excellent	----------	5

As shown the result of MOS in the above figure the difference may be the users' perception and emotion to listen and rate the texts and the pre-processing task which takes into account the effects of gemination, epenthesis, and syllabification. This evaluation technique for system performance makes evaluation of the speech synthesizer very difficult.

5.9. Analysis of the Experimentation Results

In this section the experimentation and evaluation of the system is analyzed based on the users' acceptance using MOS and ORT test methods for the synthesizer naturalness and intelligibility. From the experimental result found that the synthesizer synthesis syllables with a better performance, naturalness and intelligibility of the synthetic sound, which means that the quality of corpus data, leads the quality of speech result. In addition, the evaluation result shows from user's acceptance response based on their perception about the sound produced taking without and with effects of the pre-processing tasks like gemination, epenthesis, and syllabification and the techniques used to analysis/synthesis the waveform generated.

During evaluation of the TTS modeled and developed system the Mean Opinion Score (MOS), which involves participants in rating (some specific aspect of) output on a scale from 1 (bad) to 5 (excellent); and the Open Rhythm Test(ORT), which evaluates the system whether the sound produced is understand by the evaluator or not.

In the ORT test, the words synthesized using syllables are the more intelligible other smaller speech units. Since those of the problems like gemination, epenthesis and syllabification are identified and handle properly during corpus preparation, that Amharic language is mostly faced with those mentioned. The recorded data itself consists of those unique syllabified words for easily identification of exact boundaries, detecting unvoiced/voiced sounds, identifying of pitch mark, modifying of pitch and duration, acoustic unit extraction and normalization and easily uttering the six order CV-Syllables. In case of gemination we incorporate those of geminated form Amharic words in the acoustic unit inventory of syllable and in the test corpus. As Mulugeta (2001), cited Musa, *et al*. (1969), and discussed that syllables have been described as thrust of chest muscles of respiration, peaks of sonority, pulse of sound energy, necessary units in the mental organization and production of speech, a group of speech movements, and a basic unit of speech perception and production. The researcher remarks that syllable based speech applications are related with good quality results.

The evaluation result from selected users' perception shows that the unique syllabified words provides more intelligible synthetic sound due to that the basic problems which creates discontinuity among syllables is minimized by handling the gemination and epenthetic vowel insertion process effectively. According to evaluation test using ORT test result, an average of 73.75% for the case of simple transcribed words had been recognized and 89.58% for syllabified words had been recognized and understand correctly by listeners in the case of

syllable based synthesis. This shows effect of epenthesis, gemination, and syllabification affects the intelligibility of the sound produced by the synthesizer.

In the MOS evaluation for the naturalness of the produced sound, listeners are asked to assess the overall quality of each sentence according to the following scale: (1) bad; (2) fair; (3) good; (4) Very good; (5) excellent. The mean opinion score (MOS) is the arithmetic mean of all the scores from each individual. All listeners used for testing of the sound files are native Amharic speech experts with no known hearing problems. According to users acceptance result, which is 3.45 out of five (good), indicates that the unit selection for speech segment and linguistic pre-processing task affects the naturalness and intelligibility of the synthesizer. However, because the number of syllables is much larger than the number of the phones, there can be degradation of intelligence caused by lack of data for some syllable, which pronounced by spelled out the characters. However, as shown in MOS evaluation, the naturalness of synthetic speech produced by synthesizer depends on the quality of the corpus prepared and recorded of syllabified texts.

CHAPTER SIX

6. CONCLUSION AND RECOMMENDATION

6.1. CONCLUSION

In this thesis work, the framework of a generalized TTS system for Amharic language using syllables as speech unit and concatenative synthesis approach is implemented and evaluated. The TD-PSOLA technique has been used to modify the prosodic features (i.e. pitch and duration) of the speech segments and analyze/synthesize the desired speech utterances. The Amharic TTS synthesizer was tested on sample data and evaluated for the quality of the speech generated. The evaluation result yields high accuracies both for naturalness and intelligibility criterion is carried out by using the MOS, and ORT techniques as being the most frequently employed TTS evaluation approaches in this field which showed that the synthesized speech from Amharic TTS synthesizer is quite satisfactory.

As seen from the experimental evaluation result, the naturalness, and intelligibility of the utterances highly dependent on the quality of the data, which shows it is necessary to handle the epenthesis, gemination, and syllabification to come up the best performance. Since when we seen result from user acceptance test on the given words and sentences the result in case of syllabified words gives better result than simple transliterated words.

The performance of the system according to users' acceptance test is 73.75% and 89.58%, using ORT test result for intelligibility in case of simple transliterated and syllabified texts respectively. In addition the mean average score of 3.45 using MOS test result for the naturalness for transcribed and syllabified texts. The result shown in Table 5.1 and 5.2, indicate that the synthesizer for syllabified text is better than simple transcribed texts. The expert compares the input phoneme sequence and the output of the algorithm and gives their remark on each grapheme-to-phoneme conversion of the language Amharic. In addition, the expert checked while evaluating the syllabification (including gemination, epenthesis vowel insertion), is 97.8% word accuracy, play a great role on speech synthesis related to naturalness and intelligibility of synthetic sounds produced from the synthesizer. The linguistic pre-processing performance and the quality of corpus which is good, that shows it possible to find improved prosody and synthesis model that the success of the synthesis approach is crucially depends on an intelligent corpus design in order to find instances of all necessary units in matching prosodic contexts of the given Amharic texts.

6.2. **RECOMMENDATION**

This research is an attempt to see the possibility of improving the quality of the synthesizer result by handling the linguistic and prosodic effects using TD-PSOLA algorithm and accepting normalized syllabified input texts. Through the development of this thesis, a number of possible future works have been identified and directed. The following points came about either as a result of the strict time constraints of the study, and hence certain features were unable to be implemented, or because it was discovered that many areas offered great depth in which much future work could continue. This is especially true for corpus-based concatenative synthesis syllables as speech unit to avoid the domain specific problems, whose usability and potential is indeed enormous.

In continuation with our efforts to build speech synthesizers for Amharic language, we discussed the development of unit selection TTS for Amharic language. As a future work we would like suggest the following directions:

- The integration of the input module (automatic syllabification and prosody generation) with the rest of the synthesis system (synthesis module) is a major task to be done in future. The system then can be extended to take directly the Amharic text as input, then syllabify, record, synthesis and play the produce waveforms automatically.
- The correct assignment of stress to words in Amharic is an important aspect of the text-to-speech synthesis system. Stress in Amharic can be defined as the relative prominence of the syllables in a word and it has an important influence the pronunciation of Amharic texts.
- Spectral continuity measures to predict the audible discontinuities of the Amharic syllable boundaries and marks, since in TTS systems based on concatenative synthesis the naturalness of synthetic speech is highly affected by the spectral continuities at the concatenation points of each syllables, in the case of waveforms generation phase.
- Speech emotion development for different type emotions like normal, happy, anger, and sad, fear and grief are some of the emotion type which make the speech output as well as waveform generation varied. Therefore, there is much work that could be carried out in this area alone. However, future work in other emotions may not produce the same results found in this thesis. This would be due to a number of reasons: more complex emotions are less understood and as a consequence of speech correlates for complex emotions are much harder to identify.

- Speaking Styles, an interesting area for future work would be the investigation of developing different speaking styles. Work in this area is quite important, as the literature is very rich in the discussion of how and why we adopt different speaking styles. For instance, Knapp (1980) shows how research in the field of paralinguistic's suggests that the way we speak changes depending on who we are talking to (e.g. a group of people or "one on one", someone of the opposite sex, someone from a different age group etc). Research has also shown that male and female speakers have different speech intonation patterns and duration length.
- Duration and intonation modeling automatically from input Amharic texts, since speech synthesizer has to consider prosody, which take into the account of specific intonation and duration of speakers by building part of speech (POS) for Amharic.

REFERENCES

Alan, O., & Mikel, G. (2007). "A Brief Introduction to Speech Synthesis and Voice Modification", Dublin Institute of Technology, Dublin 2,Ireland.

Alan W., & Lenzo, K. (2003). "Building Synthetic Voices - for FestVox 2.0 Edition", Retrieved February 12, 2012, from http://www.festvox.org/bsv.

Alessandro, N. D. (2009). "An Introduction to Text-to-Speech Synthesis", Laboratoire de Théorie des Circuits et Traitement de Signal Faculté Polytechnique de Mons. *Language and Science*.

Aster, T. (1981). "The syllable structure of Amharic and syllabification of Medial Consonant Clusters and Geminates", B.A thesis in Linguistics, Addis Ababa University, Department of Linguistics, Addis Ababa, Ethiopia.

Bartlett, S., Kondrak, G., & Cherry, C. (2009). "On the Syllabication of Phonemes", Department of Computing Science, Canada.

Baye, Y. (2006). "The Interaction of Tense Aspect, and Agreement in Amharic Syntax", Addis Ababa University, Addis Aababa, Ethiopia. *Syntax*, *1999*(i), 193-202.

Baye, Y. (2010). "አጭርና ቀላል የአማርኛ ሰዋስው": ("Short and simple Amharic Grammar"), Addis Ababa, Ethiopia.

Bender, M. L., Sydeny, H. W., & Roger, C. (1976). "The Ethiopian Writing System and Language in Ethiopia", London, Oxford University press.

Birkholz, P., Steiner, I., & Breuer, S. (2007). "Control Concepts for Articulatory Speech Synthesis." ,*Computer*, 5-10.

Chabchoub, A., & Cherif, A. (2010). "Implementation of the Arabic Speech Synthesis with TD-PSOLA Modifier", Laboratory of Signal processing Science, Faculty of unis, Tunisia. *Engineering and Technology*.

Chauhan, V., Singh, G., Choudhary, C., & Arya, P. (2011). "Design and Development of a Text-To-Speech Synthesizer System", Engineering, Rookee Haridwar, Uttrakhand, India. *Database*, *7109*, 42-45.

Cholin, J. (2004). *"*Syllables in speech production: effects of syllable preparation and syllable frequency*", te Cagliari, Italïe*.

Conkie, A., & Syrdal, A. K. (1998). "Diphone Synthesis using Unit Selection", ISCA Archive, AT&T Labs - Research, Florham Park, NJ, USA. *Synthesis*.

Côté, M. (2005). "syllabification, variation and perception", Université d' Ottawa. *Journal of Phonetics*, 2003-2004.

Daniel, A. (2006). "Amharic Speech Training for the Deaf", Addis Ababa University, Department of Computer Science, Addis Ababa, Ethiopia. *Science*, (August).

Daniel, Y. (2006). "Application of the Double Metaphone Algorithm to Amharic Orthography", International Conference of Ethiopian Studies, Addis Ababa,Ethiopia.

Daniels, P. (1997). "Script of Semitic Languages in: Robert Hetzron, editor", Proceedings of the Corpus Linguistics. 16-45.

Dell, F., & Elmedlaoui, M. (1985). "Syllabic consonants and syllabification in Imdlawan Tashihiyt Berber", Journal of African Language and Linguistics.

Donovan, R. E. (1996). "Trainable Speech Synthesis", Cambridge University, *Cambridge University Engineering Department*.

Dutoit, Thierry. (1997). "An Introduction to Text-to-Speech Synthesis", Montclair State University and AT& T Labs--Research. *Computational Linguistics, 24*(2), 322-323.

Dutoit, Thierry. (2008). "A Short Introduction to Text-to- Speech Synthesis", TTS research team, TCTS Lab. *Academic research*, 1-16.

Dutoit, Theiry, & Leich, H. (1993). "MBR-PSOLA: Text-to-Speech Based on MBE Resynthesis of the segments Database", Speech Communication, Vol. 13.435-440

Dutoit, Thierry. (2001). "High-quality text-to-speech synthesis", Faculte Polytechnique de Mons, Belgium. *Quality*.

ECSA. (1998). "The 1994 Population and Housing Census of Ethiopia: Results at Country Level. Vol.1, Statistical Report 44, Addis Ababa, Ethiopia."

Edgington, M., Lowry, A., Laboratories, B. T., & Heath, M. (2000). "Residual-Based Speech Modification Algorithms for Text-to-Speech Synthesis". *Synthesis*.

Engin, T. J. E. (2006). "A Corpus-Based Concatenative Speech Synthesis", Bogazici University, Department of Computer Engineering, 34342, Bebek, ˙ Istanbul-TURKEY. *Computer*, *14*(2), 209-223.

Ethnologue. (2004). "Ethnologue Languages of the World 14th Ed", Institute of Linguistics, Retrieved February 25, 2012, from http://www.ethnologue.com.

Fék, M., Pesti, P., Németh, G., Zainkó, C., & Olaszy, G. (2006). "Corpus-Based Unit Selection TTS for Hungarian", Department of Telecommunications and Media Informatics Budapest University of Technology and Economics, Hungary. 367-373.

Griffin, D., & Lim, J. (1988). "Multi-Band Excitation Vocoder", IEEE Trans. on ASSP, vol. ASSP-36, pp. 1223-1235.

Gambäck, B., Asker, L., Olsson, F., & Atelach, A. (2002). "Collecting , Processing and Testing an Amharic Corpus", Swedish Institute of Computer Science AB Box 1263, SE-164 29 Kista, Sweden. *Forum American Bar Association*.

Gasser, M. (2009a). "Semitic Morphological Analysis and Generation Using Finite State Transducers with Feature Structures", Indiana University, School of Informatics Bloomington, Indiana, USA. *Computational Linguistics*, (April), 309-317.

Gasser, M. (2009b). "AMMORPHO 1.0 User's Guide", Indiana University, School of Informatics and Computing, India. *Computing*, 1-35.

Gasser, M. (2011). "HornMorpho : a system for morphological processing of Amharic, Oromo, and Tigrinya", Indiana University Bloomington, Indiana, USA. *Technology*, (May), 2-5.

Getahun, A. (2010). "ዘመናዊ የአማርኛ ሰዋስው በቀላል አቀራረብ": ("Modern Amharic Grammar in a simple approach.", Addis Ababa, Ethiopia.

Habamu, T. (2006). "Diphone based text-to-speech synthesis system for Amharic", MSc Project, Addis Ababa University, Faculty of Informatics, Information Science Department, Addis Ababa, Ethiopia.

Hasim, S. (2004). “A corpus-based concatenative speech synthesis system for Turkish“, Bogazici University.

Hayes, B. (2009). “Syllabification in English”, Chapter 13A, Department of Linguistics UCLA. *English*.

Henock, L. (2003). “Concatenative TTS for Amharic Language”, Addis Ababa Univesrity, faculty of Informaatics, School of Information Science, Addis Ababa, Ethiopia.

Honda, M. (2003). “Human Speech Production Mechanisms.” *Ntt Technical Review*, *1*.

Huang , X.. (2001). “Speech Synthesis”, Prentice Hall PTR. Phonetics, 53-59.

Hudson, G. (1996). “Phonology of Ethiopian Languages: The Handbook of Phonological Theory”, Blackwell Publishing.

Hunt, A., & Black, W. (1996). “Unit selection in a concatenative speech synthesis system”, ATR Interpreting Telecommunications Research Labs. *Synthesis*, 373-376.

Hussain, S. (2001). “Letter-to-Sound Conversion for Urdu Text-to-Speech System”, Center for Research in Urdu Language Processing, National University of Computer and Emerging Sciences , Faisal Town Lahore, Pakistan. *Processing*.

James L. F. (1965). “Speech Analysis: Synthesis and Perception”, Springer: Berlin.

Jilei T. (2006a). “Modular design for Mandarin text to speech synthesis“, TC-STAR workshop on speech to speech translation, Barcelona, Spain.

Kenny, N. (1998). ’’Survey of data driven approaches of speech synthesis’’, Massachusetts institute of technology.

Klatt, D. (1987). ” Review of TTS Conversion for English”, Journal of Acoustic society of America 82(3), 737-793.

Kopecek, I. (1997). “Syllable Based Speech Synthesis”, Proceedings of SPECOM’97, Cluj-Napoca.

Kuhn, M. (2009). “Digital Signal Processing (DSP)”, University of Cambridge. *Digital Signal Processing*.

Kurian, A., & Narayan, B. (2011). "Indian Language Screen Readers and Syllable Based Festival Text-to-Speech Synthesis System", IIT-Madras, India. *Computational Linguistics*, 63-72.

Laine, B. (1998). "Text-to-Speech Synthesis of the Amharic Language," MSc Thesis, Addis Ababa University , Faculty of Technology, Addis Ababa, Ethiopia.

Latsch, V. L., & Netto, S. L. (2011). "Pitch-Synchronous Time Alignment of Speech Signals for Prosody Transplantation", Federal University of Rio de Janeiro, Brazil. 2405-2408.

Lemmetty, S. (1999). "Review of Speech Synthesis Technology", Master's Thesis, Helsinki University of Technology. *Science*.

Leslau, W. (1996). "Amharic Language Reference Grammar: A revised edition", University of California, Los Angeles.

Lewis, E. (2001). "Automatic Segmentation of Recorded Speech into Syllables for Speech Synthesis", Proceedings of Eurospeech '01, 1703-1707. Aalborg: International Speech Communication Association, 1-5.

Lewis, E., Tatham, M., Building, M. V., Road, W., & Park, W. (2000). "Word and Syllable Concatenation in Text-To-Speech Synthesis", Department of Computer Science, Merchant Venturers Building, Woodland Road. *Science and Language*, 1-4.

Liberman, M. (1992). "Text Analysis and Word Pronunciation in Text-To-Speech Synthesis."

Libossek, M., & Schiel, F. (2000). "Syllable Based Text - To- Phoneme Conversion For German", Bavarian Archive for Speech Signals (BAS). *System*.

Lin, C., & Jang, J. (2004). "A Two-Phase Pitch Marking Method for TD-PSOLA Synthesis", Department of Computer Science, National Tsing Hua University, Taiwan. *Search*.

Lin, C.-yuan, Jang, J.-shing R., & Chen, K.-ting. (2005). "Automatic Segmentation and Labeling for Mandarin Chinese Speech Corpora for Concatenation-based TTS", The Association for Computational Linguistics and Chinese Language Processing. *Computational Linguistics*, *10*(2), 145-166.

Lyons, R. (2004). “Understanding Digital Signal Processing“, Addison-Wesley, Upper Saddle River, 2nd edition.

Maheswari, U., & Rajeswari, K. (2012). “Prosody Modeling Techniques for Text-to-Speech Synthesis Systems” - A Survey, international Journal of Computer Applications, TamilNadu, India. *International Journal*, *39*(16), 2010-2013.

Markel, J. D., & Gray, A. H. (1976). “Linear Prediction of Speech”, Springer-Verlag, Berlin, Heidelberg, New York.

Minghui, D. (2000). “Speech Synthesis Techniques”. Singapore: National university of Singapore, school of computing.

Mohanty, S. (2011). “Syllable Based Indian Language Text To Speech System”, Department of Computer Science and Application, Utkal University, Bhubaneswar. *International Journal*, *1*(2), 138-143.

Monaghan, A., & Keynes, M. (2001). “A Brief Outline of Aculab TTS : Multilingual TTS for Computer Telephony.” *Architecture*.

Morais, E., & Violaro, F. (2005). “Data-Driven Text-to-Speech Synthesis”, School of Electrical and Computer Engineering, University of Campinas, São Paulo, Brazil. *Science and Language*, 04-08.

Moulines, E., & Charpentier, F. (1990). Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones'', Speech Communication 9. p. 453.

Mulugeta, S. (2001). “The syllable Structure and Syllablification in Amharic”, Masters of philosophy in general linguistic thesis, Department of Linguistics, Trondheim, Norway.

Musa, H., Kadir, R. A., Azman, A., & Abdullah, M. T. (1969). "Syllabification Algorithm based on Syllable Rules Matching for Malay Language". *World Journal Of The International Linguistic Association*, 279-286.

Möbius, B. (2000). “Corpus-Based Speech Synthesis: Methods and Challenges”, University of Stuttgart, AIMS 6 (4), pp. 87-116.

NLPA-Phon2. (2007). "Natural Language Processing & Applications Speech Synthesis and Recognition". *English*, *2*, 1-10.

Nadew, T. (2008). "Formant Based Speech Synthesis for Amharic Vowels", Graduate Studies of Addis Ababa University, Department of Computer Science, Addis Ababa, Ethiopia. *Quality*.

Nirayo, H. (2011). "Modeling Improved Amharic Syllabification Algorithm", Addis Ababa University, Department of Computer Science, Addis Ababa, Ethiopia. *Language*.

Obin, N., Rodet, X., & Lacheret-dujour, A. (2009). "A Syllable-Based Prominence Detection Model Based on Discriminate Analysis and Context-Dependency Analysis-Synthesis", Speech and Computer team, Paris, France. *Framework*, 1-4.

O'Shaughnessy, D. (2007). "Speech Communication Human and Machine", IEEE Circuits and Systems Magazine, IEEE, INRS-EMT, Montreal, Canada.

Paola, A., Dorran, D., Lawlor, R., & Coyle, E. (1985). "High quality time-scale modification of speech using a peak alignment overlap-add algorithm (PSOLA)", Dublin Institute of Technology, National University of Ireland, Maynooth. *Synthesis*.

Plumpe, M., Acero, A., Hon, H., Huang, X., & Way, O. M. (1998). "HMM-Based Smoothing for Concatenative Speech Synthesis", Washington 98052, USA. *Synthesis*, 1-4.

Raj, A. A., & Sarkar, T. (2007). "Text Processing for Text-to-Speech Systems in Indian Languages", 6th ISCA Workshop on Speech Synthesis, Bonn, Germany. *Building*, 188-193.

Rashad, M. Z., & Mastorakis, N. (2003). "An Overview of Text-To-Speech Synthesis Techniques", Department of Computer Science, Faculty of Computer and Information Systems, Mansoura University, Egypt. *Text*, 84-89.

Roelofs, A. (2002). "Syllable structure effects turn out to be word length effects", Language And Cognitive Processes, 2002. *Structure*, *17*(1), 1-13.

Rong-Wei, J. (2003). "Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis", Massachusetts Institute of Technology, Doctor of Philosophy in Electrical Engineering and Computer Science. *Development*.

Ruth, K. (2008). “The Amharic Definite Marker and the Syntax-Morphology Interface”, Ruth Kramer University of California , Santa Cruz, 1-39.

Sarasathi, S., & Vishalkshy, R. (2010). “Design of Multilingual Speech Synthesis System” , Department of Information Technology, Pondicherry Engg., College, Pondicherry, India. *Intelligent Information Management*, (January), 58-64.

Scelta, G. (2001). “The Comparative Origin and Usage of the Ge’ez writing system of Ethiopia”, Arts of Africa, AH 215.

Sebsibe, H., Kishore, S., Black, A., Kumar, R., & Sangal, R. (2004). “Unit Selection Voice for Amharic Using Festvox”, Language Technologies Research Center International Institute of Information Technology, Hyderabad. *Synthesis*, 103-107.

Shah, A., Ansari, A., & Das, L. (2004). “Bi-Lingual Text to Speech Synthesis System for Urdu and Sindhi”, Institute of IT, University of Sindh, Jamshoro, Pakistan. *System*, *1*, 126-130.

Sisay, F., & Haller, J. (1987). “Application of corpus-based techniques to Amharic texts”, Institute for Applied Information Sciences, University of Saarland, Saarbrücken, Germany.

Solomon, T., & Menzel, W. (2007). “Syllable-Based Speech Recognition for Amharic”, Hamburg, Germany. *Computational Linguistics*, (June), 33-40.

Szklanny, K., & Wójtowski, M. (2008). “Automatic segmentation quality improvement for realization of unit selection speech synthesis”, Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw.

Thomas, S. (2007). “Text to speech Based On Syllable-Like Units”, Department Of Computer Science and Engineering Indian Institute of Technology Madras. *Computer*.

Tadesse, A. (2009). “Development of an Amharic Text-to-Speech System Using Cepstral Method”, ICT Development Office, Addis Ababa University, Ethiopia. *Computational Linguistics*, (March), 46-52.

Tadesse, A., Takara, T., & Kim, D. (2010). “Modeling Of Geminate Duration In An Amharic Text-To-Speech Synthesis System”, ISCA Archive, Penang, Malaysia. *Computer*, 122-129.

Tadesse, A., Gasser, M. & Yoon, K. (2011). "Grapheme-to-Phoneme Conversion for Amharic Text-to-Speech System ", Ajou University, Graduate School of Information and Communication, South Korea. *Technology and Communication*, (May), 2-5.

Tesfay, Y. (2004). "Diphone based TTS synthesis system for Tigrigna Language”, MSc Thesis, Addis Ababa University, Faculty of Informatics, School of Information Science, Addis Ababa, Ethiopia.

Toma, Ş., Târşa, G., Oancea, E., Munteanu, D., Totir, F., & Anton, L. (2010). “A TD-PSOLA Based Method for Speech Synthesis and Compression”, Military Technical Academy, Bucharest, Romania. *Time*, 123-126.

Utama, R. ., & Syrdal, A. K. (2006). “Six Approaches to Limited Domain Concatenative Speech Synthesis”, Interspeech, Pittsburgh, Pennsylvania.

Venugopalakrishna, Y., Thomas, S., & Bommepally, K. (2007). “Design and Development of a Text-To-Speech Synthesizer for Indian Languages”, Indian Institute of Technology Madras, Chennai, India. *Technology*.

Venugopalakrishna, Y., Vinodh, M., Murthy, H., & Ramalingam, C. (2003). “Methods For Improving The Quality Of Syllable Based Speech Synthesis”, Department of Computer Science and Engineering, Indian Institute of Technology Madras. *Techniques*.

Verhelst, W., & Roelands, M. (1993). “An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech”, Faculty of Applied Science, dept., Belgium. *Evaluation*, (1), 2-5.

Weber, A. (2005). “Stop Epenthesis at Syllable Boundaries”, Department of Linguistics University of Arizona, Tucson, Arizona, and Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands. *Production*, *1*.

Yamagishi, J., Kobayashi, T., Renals, S., King, S., Zen, H., Toda, T., & Tokuda, K. (2007). “Improved Average-Voice-based Speech Synthesis Using Gender-Mixed Modeling and a Parameter Generation Algorithm Considering GV.” *Challenge*.

Yibeltal, T. (2008). “Formant-Based Speech Synthesis: A Case of Amharic Words”, School of Graduate Studies, Faculty of Informatics, Department of Computer Science,Addis Ababa, Ethiopia. *Science*.

Zervas, P., Potamitis, I., Fakotakis, N., & Kokkinakis, G. (2001). “A Greek TTS Based on Non Uniform Unit Concatenation and the Utilization of Festival Architecture”, Wire Communications Lab, Department of Electrical & Computer Engineering University of Patras, 26500, Rion, Patras, Greece. *Challenges*.

Zhang, J. (2004). “Language Generation and Speech Synthesis in Dialogues for Language Learning", Msc thesis, Massachusetts Institute of Technology.

APPENDICS

Appendix1. 1: Some of transliterate and syllabified test corpus

Evaluated by: Solomon Getahun
Linguistic department, AAU

Transliterated Input word	Syllabification by the system	Remark on G2P	Remark on Syllabification
alleme	al-le-me		
amlak	am-lak		
ammeme	am-me-me		
asar	a-sar		
assebe	as-se-be		
assere	as-se-re		
assese	as-se-se		
astemarii	as-te-ma-rii		
astewale	as-te-wa-le		
asteyayet	as-te-ya-yet		
bahl	ba-hixl		
begena	be-ge-na		
berie	be-rie		
berkatta	ber-kat-ta	Missed epenthesis	be-rix-kat-ta
berr	berr		
berra	ber-ra		
berrede	ber-re-de		
bet	bet		
blh	bixlh		
blhat	bixl-hat		
Bltxgna	bixl-txixg-na	Missed epenthsis	bixl-txix-gix-na

cenkar	cen-kar		
cxelleme	cxel-le-me		
cxereqa	cxe-re-qa		
debr	de-bixr	Incorrect epenthesis	debr
demmere	dem-me-re		
dgr	dix-gixr		
dhnnet	dix-hixn-net		
dldy	dixl-dixy		
dmmera	dixm-me-ra		
dmmet	dixm-met		
dnber	dixn-ber		
drsan	dixr-san		
dubbe	dub-be		
duro	du-ro		
fana	fa-na		
fecxcxe	fecx-cxe		
fegegta	fe-geg-ta		
fekka	fek-ka		
fellege	fel-le-ge		
fellgewal	fel-lix-gix-wal		
Fellgo	fel-lix-go		
Flat	fix-lat		
flfl	fixl-fixl		
fllagot	fixl-la-got		
flqlq	fixl-qixlq	Missed epenthesis	fixl-qix-lixq

foqqeqe	foq-qe-qe		
fth	fix-tixh		
ftxnet	fixtx-net		
gehannem	ge-han-nem		
gena	ge-na		
genna	gen-na		
gennet	gen-net		
gerreme	ger-re-me		
gobez	go-bez		
gobezazt	go-be-zazt		
gommen	gom-men		
habt	habt		
habtam	hab-tam		
habtamnnet	hab-ta-mixn-net		
hayyal	hay-yal		
hdar	hix-dar		
huket	hu-ket		
hullumm	hul-lumm		
hxxanat	hix-xxa-nat		
ixbamm	ix-bamm		
ixbiita	ix-bii-ta		
ixgeliit	ix-ge-liit		
ixgziiabher	ixg-zii-ab-her		
ixnglt	ixn-gixlt		
ixnkokko	ixn-kok-ko		

ixnkoy	ixn-koy		
ixssrat	ixs-six-rat		
ixssratie	ixs-six-ra-tie		
jemmer	jem-mer		
kbriit	kixb-riit		
kedda	ked-da		
kefaffete	ke-faf-fe-te		
keffete	kef-fe-te		
kellele	kel-le-le		
kemiil	ke-miil		
Kenafr	ke-na-fixr		
kereddede	ke-red-de-de		
kesebat	ke-se-bat		
kesekkese	ke-sek-ke-se		
kesost	ke-sost		
kessiitta	kes-siit-ta		
ketegenebu	ke-te-ge-ne-bu		
keteketelut	ke-te-ke-te-lut		
kewakbt	ke-wak-bixt		
kffl	kixf-fixl		
kfftt	kixf-fixtt		
kflfay	kixf-lix-fay		
kft	kixft		
klkl	kixl-kixl		
kokeb	ko-keb		

lbb	lixbb		
legenna	le-gen-na		
leixntxotxo	le-ixn-txo-txo		
lemelleme	le-mel-le-me		
lemezzege	le-mez-ze-ge		
lemlamie	lem-la-mie		
lemmanxnet	lem-ma-nxix-net		
leslassannet	les-las-san-net		
maaxbel	maax-bel		
maaxkel	maax-kel		
maaxqeb	maax-qeb		
maaxzen	maax-zen		
mabreja	mab-re-ja		
mamonxnxa	ma-monx-nxa		
manoriiya	ma-no-rii-ya		
maqwaqwamiiya	maq-waq-wa-mii-ya		
masgebatun	mas-ge-ba-tun		
mastaweqiiya	mas-ta-we-qii-ya		
mastawesxa	mas-ta-we-sxa		
mazawer	ma-za-wer		
mazoriiya	ma-zo-rii-ya		
mebrat	meb-rat		
mebrathayl	meb-rat-hayl		
Mehal	me-hal		
mehaym	me-haym	Missed epenthesis	me-ha-yixm

mehayyem	me-hay-yem		
mehonacnen	me-ho-nac-nen		
mekkelakeya	mek-ke-la-ke-ya		
melaaxkt	me-la-axkt		
melak	me-lak		
melaktenxoc	me-lak-te-nxoc		
melkam	mel-kam		
mlkkt	mix-lixk-kixt		
Mncet	mixn-cet		
mngllat	mixn-gixl-lat		
mnm	mixnm		
mnnxnxa	mix-nixnx-nxa		
mnxotuna	mix-nxo-tu-na		
mnzarii	mixn-za-rii		
nefsie	nef-sie		
neger	ne-ger		
neh	neh		
nehasie	ne-ha-sie		
nen	nen		
nenx	nenx		
neqaqqele	ne-qaq-qe-le		
nerrere	ner-re-re		
nesennese	ne-sen-ne-se		
nessa	nes-sa		
nesxsxetxe	nesx-sxe-txe		

nssha	nixs-six-ha		
nuro	nu-ro		
posta	pos-ta		
postenxnxa	pos-tenx-nxa		
pxapxpxas	pxapx-pxas		
pxaqumie	pxa-qu-mie		
pxepxpxese	pxepx-pxe-se		
qbaqddus	qix-ba-qixd-dus		
qddasie	qixd-da-sie		
Qddst	qixd-dix-sixt		
qddusnnet	qixd-du-sixn-net		
qelatxie	qe-la-txie		
Qellal	qel-lal		
qenecxcxebe	qe-necx-cxe-be		
qenxnx	qenxnx		
qerrere	qer-re-re		
Qlqql	qix-lixq-qixl		
qltxfftxf	qixl-txixf-fix-txixf		
Qntxot	qixn-txot		
qntxtxabii	qix-nixtx-txa-bii		
qtxr	qix-txixr		
Qullff	qul-lixff		
Radde	rad-de		
raq	raq		
rbrb	rixb-rixb		

rebbadda	reb-bad-da		
reta	re-ta		
rguz	rix-guz		
riesa	rie-sa		
robna	rob-na		
same	sa-me		
samii	sa-mii		
saniitiesxn	sa-nii-tie-sxixn		
saran	sa-ran		
sayasdebedbwat	sa-yas-de-bed-bix-wat		
saybela	say-be-la		
saytenxa	say-te-nxa		
sbket	sixb-ket		
seaxeliinnet	se-axe-liin-net		
sebabbere	se-bab-be-re		
Sellatie	sel-la-tie		
selletene	sel-le-te-ne		
sew	sew		
seyyeme	sey-ye-me		
siihiedu	sii-hie-du		
siiqeldubet	sii-qel-du-bet		
siiwetxu	sii-we-txu		
siiyasgeddew	sii-yas-ged-dew		
siiyayu	sii-ya-te-yu		
sl	sixl		

slbabot	sixl-ba-bot		
slenekubacew	six-le-ne-ku-ba-cew		
sltxanat	sixl-txa-nat		
Sltxun	sixl-txun		
smmnet	sixm-mix-net		
taggese	tag-ge-se		
tajjebe	taj-je-be		
takkeme	tak-ke-me		
tammeme	tam-me-me		
tassere	tas-se-re		
tataqii	ta-ta-qii		
taxgst	tixax-gixst		
tbiitenxnxa	tix-bii-tenx-nxa		
tdar	tix-dar		
tebazxii	te-ba-zxii		
tebejajje	te-be-jaj-je		
tebekakkele	te-be-kak-ke-le		
tebesabbese	te-be-sab-be-se		
tebetxatxasx	te-be-txa-txasx		
tebetxtxese	te-betx-txe-se		
tebiib	te-biib		
tebiiban	te-bii-ban		
teblecxellecxe	teb-le-cxel-le-cxe		
tecellese	te-cel-le-se		
tedeladdele	te-de-lad-de-le		

tedeladele	te-de-la-de-le		
tedemarii	te-de-ma-rii		
tedenaggere	te-de-nag-ge-re		
tederragii	te-der-ra-gii		
tefetelleke	te-fe-tel-le-ke		
tefexxeme	te-fe-xxe-me		
tegafetxe	te-ga-fe-txe		
tegbarawii	teg-ba-ra-wii		
tegbaroc	teg-ba-roc		
tegebiiwn	te-ge-biiwn	Missed epenthesis	te-ge-bii-wixn
tegelxxo	te-gel-xxo		
tegojiiwoc	te-go-jii-woc		
tegsaxxx	teg-saxx		
tejemmere	te-jem-me-re		
tejj	tejj		
tekebbere	te-keb-be-re		
tekefaffele	te-ke-faf-fe-le		
tekeffetu	te-kef-fe-tu		
tekenawnwal	te-ke-naw-nix-wal		
tekessete	te-kes-se-te		
tekettebe	te-ket-te-be		
tekettele	te-ket-te-le		
tekl	te-kixl	Incorrect epenthesis	tekl
tekl	te-kixl		
tekliil	tek-liil		

tekunesennese	te-ku-ne-sen-ne-se		
tekuwasx	te-ku-wasx		
teleqalleqe	te-le-qal-le-qe		
telixko	te-lix-ko		
tenafaqii	te-na-fa-qii		
tenefafaqi	te-ne-fa-fa-qix		
tenfwaffwa	ten-fix-waf-fix-wa		
tengedgaj	ten-ged-gaj		
tengefeggefe	ten-ge-feg-ge-fe		
tentxefetxtxefe	ten-txe-fetx-txe-fe		
tentxefetxtxefe	ten-txe-fetx-txe-fe		
tenzazza	ten-zaz-za		
teqoranxnxe	te-qo-ranx-nxe		
teraggetxe	te-rag-ge-txe		
tereggetxe	te-reg-ge-txe		
termetxemmetxe	ter-me-txem-me-txe		
tesebbere	te-seb-be-re		
tesebbere	te-seb-be-re		
tgray	tixg-ray		
thtnna	tixh-tixn-na		
tjja	tixj-ja		
tlalq	tix-lalq	Missed epenthesis	tix-la-lixq
tlenxnxa	tix-lenx-nxa		
tlk	tixlk		
tllqnnet	tixl-lix-qixn-net		

tmhrt	tixm-hixrt		
tmhrtawii	tixm-hixr-ta-wii		
tmlml	tixm-lix-mixl		
tnannsx	tix-nan-nixsx		
tnnsx	tixn-nixsx		
tnnx	tixnnx	Incorrect (missed epenthesis)	Missed epenthesis (tix-nixnx)
Trf	tixrf		
wanza	wan-za		
wdddr	wixd-dix-dixr		
wedaj	we-daj		
weddese	wed-de-se		
wefcxo	wef-cxo		
wehynii	weh-yix-nii		
wendnnet	wen-dixn-net		
wengel	wen-gel		
wenjel	wen-jel		
weqqesa	weq-qe-sa		
werrede	wer-re-de		
wesdalet	wes-da-let		
weyzazrt	wey-zaz-rixt	Incorrect epenthesis	wey-za-zixrt
weyzeriit	wey-ze-riit		
weyzero	wey-ze-ro		
wgenxnxa	wix-genx-nxa		
xxdat	xxix-dat		
xxehafii	xxe-ha-fii		

xxelot	xxe-lot		
xxelotenxnxa	xxe-lo-tenx-nxa		
xxenna	xxen-na		
xxensa	xxen-sa		

Appendix1. 2: Amharic Abugida System

	ä	u	i	a	e	ï	o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
l	ለ	ሉ	ሊ	ላ	ሌ	ል	ሎ
h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
m	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ
s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ
r	ረ	ሩ	ሪ	ራ	ሬ	ር	ሮ
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
sh	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ
q	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
ch	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
ñ	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
a	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ

	ä	u	i	a	e	ï	o
h	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ
w	ወ	ዉ	ዊ	ዋ	ዌ	ው	ዎ
a	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ
z	ዘ	ዙ	ዚ	ዛ	ዜ	ዝ	ዞ
zh	ዠ	ዡ	ዢ	ዣ	ዤ	ዥ	ዦ
y	የ	ዩ	ዪ	ያ	ዬ	ይ	ዮ
d	ደ	ዱ	ዲ	ዳ	ዴ	ድ	ዶ
j	ጀ	ጁ	ጂ	ጃ	ጄ	ጅ	ጆ
g	ገ	ጉ	ጊ	ጋ	ጌ	ግ	ጎ
t'	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
ch'	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
p'	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
s'	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
s'	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ

Appendix1. 3: Amharic Phonetic List, IPA Equivalence and its ASCII transliteration(Sebsibe, et al., 2004)

IPA	**Transcription**	**Amharic equivalence**
	Consonants	
[p]	[p]	ፕ
[t]	[t]	ት
[k]	[k]	ክ
[ʔ]	[ax][A]	ዕ
[b]	[b]	ብ
[d]	[d]	ድ
[g]	[g]	ግ
[p’]	[px][P]	ጵ
[t’]	[tx][T]	ጥ
[c’]	[cx][C]	ጭ
[q]	[q]	ቅ
[f]	[f]	ፍ
[s]	[s]	ስ
[ʃ]	[sx][S]	ሽ
[h]	[h]	ህ
[s’]	[xx] [X]	ጽ
[tʃ]	[c]	ች
[g’]	[j]	ጅ
[m]	[m]	ም
[n]	[n]	ን
[n’]	[nx] [N]	ኝ

[l]	[l]	ል
[r]	[r]	ር
[j]	[y]	ይ
[w]	[w]	ው
[v]	[v]	ቭ
[z]	[z]	ዝ
[z’]	[zx] [Z]	ዥ
Vowels		
[E]	[e][E]	ኧ
[U]	[u][u]	ኡ
[I]	[ii] [I]	ኢ
[A]	[a][a]	ኣ
[e]	[ie][e]	ኤ
[ˆ]	[ix][i]	እ
[o]	[o][o]	ኦ

Appendix1. 4: Speech waveform generation of the word ሰበረ -"*sebbere*" using Matlab

```
%-------------------------------------------------------------------------
% Get a waveform from recorded sampled speech corpus
%-------------------------------------------------------------------------
speechSignal = input('Enter file name: ','s');
    clc;        %clears the command window
    close all;
    clear all;    %Clear variables and functions from memory
  %-----------------------------------------
fid = fopen(speechSignal,'r');  % Open the text file for reading
   first = fgets(fid);  % In this case, scan in the first line and ignore
   [F,count] = fscanf(fid, '%f%f%f');  % Scan the data into a vector
%-----------------------------------------
% Set the other parameters
  %-----------------------------------------
winLen = 301;
winOverlap = 300;
Fs=8000;                    %The sampling rate
Ts=1/Fs;                  % sample period
Ts=1/Fs;          %sample period
n=0:3.5*Fs-1;             %sample index (fs=samples per second), 3.5 seconds
ms10=floor(Fs*.01);
ms30=floor(Fs*0.03);
ncoeff=2+Fs/1000
%-----------------------------------------
% sample Hanning window is chosen
%-----------------------------------------
wHann = hanning(winLen);
% Framing and windowing the signal without for loops
sigFramed = buffer(speechSignal, winLen, winOverlap, 'nodelay');
sigWindowed = diag(sparse(wHann)) * sigFramed;
% Short-Time Energy calculation
energyST = sum(sigWindowed.^2,1);
% Time in seconds, for the graphs
```

```
t = [0:length(speechSignal)-1]/Fs;
%---------------------------------------------
% Plot the Waveform of the File given with background
%---------------------------------------------
subplot(1,1,1);
plot(t, speechSignal);
hold on;
%---------------------------------------------
% Short-Time energy is delayed due to low pass filtering. This delay is
% compensated for the graph
%---------------------------------------------
delay = (winLen - 1)/2;
figure('Color',[1 1);
plot(t(delay+1:end - delay), energyST, 'r');
%---------------------------------------------
% Assign legend of the speech Wavefrom as time domain
%---------------------------------------------
legend('Waveform', speechSignal);
xlabel('Time (s)');
ylabel('Amplitude');
% process in chunks of 30ms to determine peaks and length of formants
pos=1;
fm=[];      % formant peaks
ft=[];      % formant times
%---------------------------------------------
% Iterate till last chunks of the Sound file
%---------------------------------------------
while (pos+ms30) <= length(x)
   y=x(pos:pos+ms30-1);
   y=y-mean(y);
%---------------------------------------------
   % find TD-PSOLA  filter
%---------------------------------------------
   a=tdpsola(y,ncoeff);
   % find roots
```

```
    r=roots(a);
    r=r(imag(r)>0.01);
    ffreq=sort(atan2(imag(r),real(r))*fs/(2*pi));
    for i=1:length(ffreq)
       fm = [fm ffreq(i)];
       ft = [ft pos/fs];
    end
    pos=pos+ms10;
end;              % end of while loop

%-----------------------------------------------
% Information about the speech signal generated
%-----------------------------------------------
fprintf('Information of the sound file "%s":\n', speechSignal);
fprintf('Duration = %g seconds\n', length(y)/Fs);
fprintf('Sampling rate = %g samples/second\n', Fs);
fprintf('Bit resolution = %g bits/sample\n', nbits);
%-----------------------------------------------
   % Write the output, plot and listen the sound
   %-----------------------------------------------
  output = output/max(output);
  wavwrite(output,sf,16,'C:\matlab\samplespeech.wav'); % path the output to   be stored
  plot(output)
  sound(output)                        % Let's hear it the synthesized sound
  disp('program finished');
hold off
  %-------------------end of program----------------------------
```