ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

Text-to-Hymn Synthesis for St Yared Hymn Notations

By: Girma Zemedu

A Thesis Submitted to the School of Graduate Studies of Addis Ababa University in Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Science

March 2014

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

COLLEGE OF NATURAL SCIENCE
DEPARTMENT OF COMPUTER SCIENCE

Text-to-Hymn Synthesis for St Yared Hymn Notations

By

Girma Zemedu

APPROVED BY

EXAMINING BOARD:

1. Dr. Yaregal Assabie, Advisor _____

2. _____ _____

3. _____ _____

Table of Contents

List of Figures

List of Tables

Dedicated

To Ethiopian Traditional schools and their scholars

ACKNOWLEDGEMENT

First of all I would like to thank the God of St Yared, who gave him the bread of Angels. Next to this my gratitude goes to Dr. Yaregal Assabie, who gave me his full support of advice starting from the day of topic selection up to this point.

I would also like to thank Memhir WoldeGebriel who supports me by giving his full dedication when I visit him in his office at Mahibere Kidusan. In this work his polar value while recording and marking Notes has a great place.

I wish to thank Merigeta feriesbhat and his colloquies for their support in the early time of this thesis work.

My special respect is goes to my brothers and sisters who gave me moral and financial support. They got a special place on my first research work.

At last but not least I owe a special gratitude to my wife and child who are always with me by supporting in all ups and downs; I thank you for your patience.

ABSTRACT

Speech synthesis, a process of making artificial speech, has been one of the research areas for the last many decades. In addition to this singing voice synthesis has also started and a lot has been done internationally for western songs. But nothing has been done for the St Yared hymn notation which is used for nearly 1500 in most traditional music and inside Ethiopian Orthodox church. On the other hand, the number of scholars decreased dramatically for different reason and it is scared to be lost in the future. Thus, to preserve these notations text-to-speech synthesis for St Yared hymn notation is required.

To this effect, we used a unit selection concatenative approach to synthesize the notations. Initially, different hymn lyrics were collected and transcribed properly by marking the notes. After that voice of an expert of the hymn was recorded for the transcribed lyrics. Segmentation of the note from the recorded voice is done manually and saved in a dataset with its context information of the left and right for target note. In the synthesis phase the selected unit waves are synthesized by making certain changes on the pitch and intensity to smoothly concatenate the units. A prototype was developed to implement this model.

The system was evaluated using a test data collected from various hymn lyrics. These lyrics were tested with a simple concatenation of notes without note markers and with a note marker. Mean Opinion Score (MOS) for both Naturalness and intelligibility, which was taken after an assessment made by a number of users, shows a promising result. For naturalness and intelligibility of unmarked notes we achieved a performance result of 2.3 and 2.6 respectively. For naturalness and intelligibility of marked notes we achieved a performance result of 3.9 and 3.4 respectively

Key words, St Yared note, concatenative speech synthesis, St Yared Hymn synthesis

Chapter 1

Introduction

1.1 Background of the Study

Natural language is used as a means of communication for so many years since the era of mankind on the planet. One means of natural language communication using is speech among those who are able to communicate by voice. As technology advances the way human beings live and communicate also change. The change made on written text is accelerated due to the electronics advancement with memory capacity and processing speed. This intern tends to make a significant change on human behavior once again from voice–to-text to human-machine – understanding (artificial language).

Nowadays many of natural language parts are on significant processing speed to meet the requirement for human-machine-understanding scheme. Natural language processing is an interdisciplinary field of study dealing with computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. Text-To-Speech (TTS) Synthesis is a field that converts written text to speech. Nowadays it is not difficult to get application more than a dozen that uses TTS as means of tools to give different services [1].

As Sasirekha *et al* [2] states the evolution of TTS starts in 1779, when the first models of the human vocal tract that could produce the five long vowel sounds. These are [a], [e], [i], [o] and [u]. The next TTS developed was a system which implemented Pharyngeal Cavity, used for singing. It was controlled by keyboard. After the great change in electronic and computer transformation, the first speech unit concatenation was developed by 1970's. In 1980's and 1990's TTS show a great progress in Unit Selection and Diphone Synthesis.

According to the manual intervention speech synthesis is classified as rule-based synthesis and data-driven synthesis. Where a set of rules are produced to drive the synthesizer in case of rule-

based and typically this approach is called a formant synthesis. Formant synthesis mainly depends on Parameters such as fundamental frequency, voicing, and noise levels are varied over time to create a waveform of artificial speech. The other approach is concatenative synthesis where concatenation of recorded speech segments is produced to output the most natural – sounding speech.

Even though research works on local language for TTS is less than two decade old, there are a lot of papers written in local language text to speech synthesis. The effort of Sebsibe *et al* [3] is an encouraging one as shown on their paper. They defined a transliteration scheme to work with Amharic scripts and incorporated Amharic phone set, syllabification rules, letter to sound rules into Festvox. By selecting the prompt-list from various sources and built a unit selection voice for Amharic. On the other hand, Eyob and Yaregal [4] proposed a method that represents intonation variation according to punctuation marks. They also introduced a set of rules that can be used to locate phones that need to be stressed in sentences for Amharic language.

Since Ethiopia is one of the oldest countries in the history of early civilization, there are scripts written which are used for religious purpose and for social values. One of such script is digua, which is Ethiopian Orthodox Church script. It has its own hymn (zema) with notation indicator signs. It was originally written by St Yared. The hymn is also called as St Yared hymn by the name of the inventor. St Yared is among top scholars of Ethiopian Orthodox Church Priests. He was born early 6^{th} century in Axum, Ethiopia. It is believed that he got the three main hymn scores from three birds. Yared named them as *Geez*, *Izil*, and *Araray* [5]. Yared hymn is mostly used by Ethiopian Orthodox church, which has a follower of substantial size of the population of Ethiopia.

Hymn produced by St Yared has ten musical notations. The notation has their own name which is stated as follows: *Yizet, Deret, Rikrik, Difat, Cheret, Qenat, Hidet, Qurt, Dirs*, and, *Anbir*. It was believed that the final two (*Dirs* and *Anbir)* are added after Yared. Alphabet symbols are also added to more simplify hymn which they called it *sirey* (root)[6]. Table 1.1 shows the notations with their name in English and Geez.

Table 1-1 Yared Hymn note symbols with their name

The English transcription	The Geez name	Sign
Yizet	ይዘት	
Deret	ደረት	
Rikrik	ርክርክ	
Difat	ድፋት	
Cheret	ጭረት	
Qinat	ቅናት	
Hidet	ሂደት	
Qurt	ቁርጥ	
Dirs	ድርስ	
Anbir	አንበር	

1.2 Motivation

Hymn of St. Yared is being used for about 1500 years since he first articulated it. Since then, different notations were added on the hymn to simplify the study of Digua. One can study these Hymn from traditional schools inside churches that teaches students, which are called *Abinet temari* or *kolo temari*. The signs have different sound when put over the character of Geez. Even for one sign it shows different sound signal as illustrated below in Figure 1.1. The yellow line shows the intensity of the signal and the blue shows the pitch of the signal. As indicated on the Figure there is a difference of pitch for the same rikrik note with a pitch of 354 Hz in case of (a)

and a pitch of 221.9Hz in case of (b). This makes it difficult to learn in a few months or year. The fact that such a synthesizer is not yet developed for such language is an inspiration of this research.

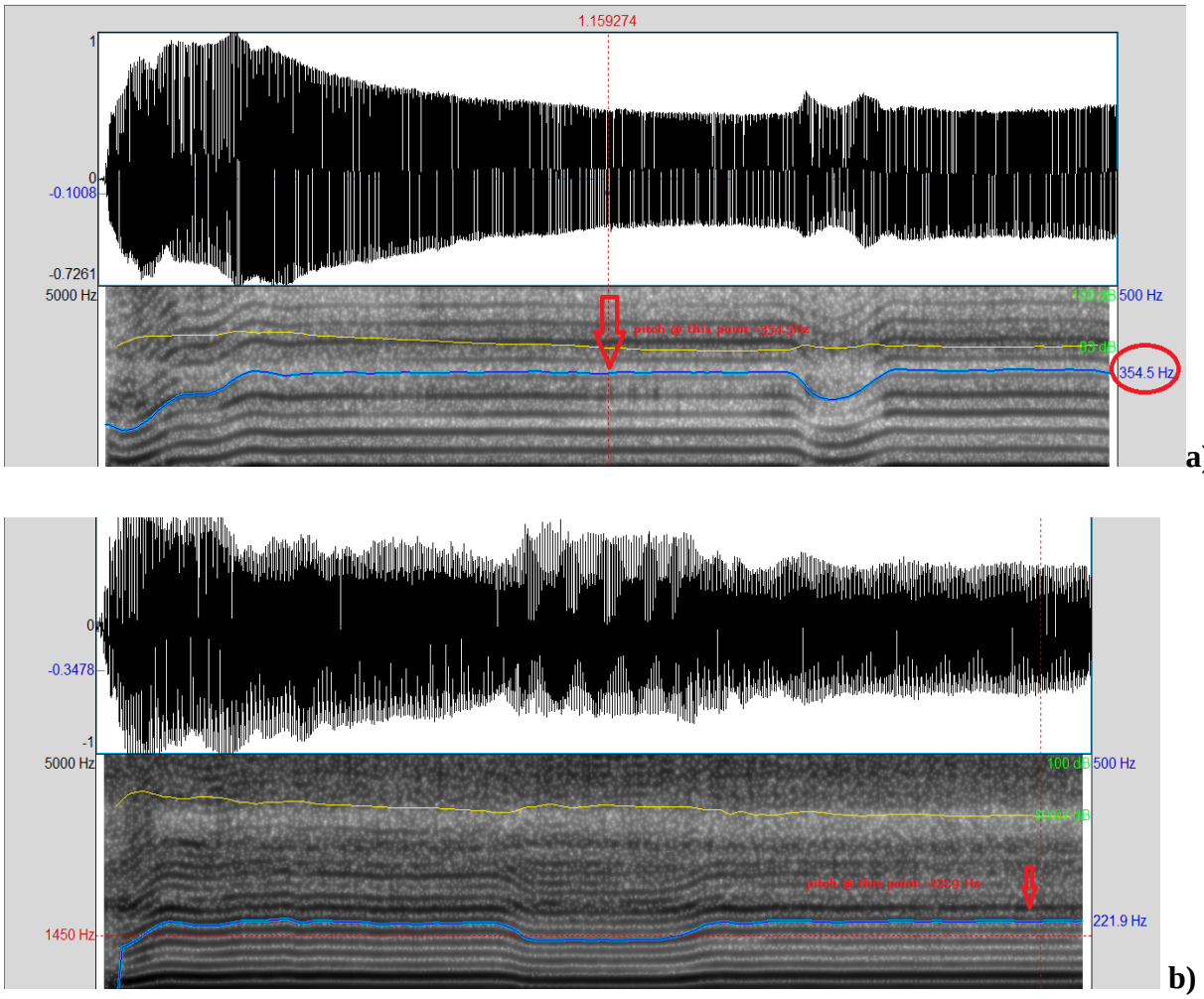


Figure 1.1 Signal wave for *rikrik* note: (a) at pitch of 354Hz, (b) at pitch of 221.9Hz

1.3 Statement of the Problem

Until recently it was normal for boys to go and study in the traditional schools. Today boys do not obey this tradition blindly; however a few young boys who follow modern education in state schools now follow, at the same time, traditional education in church schools. The lives of elder teachers of traditional schools are getting difficult, for that matter they are migrated to city.

When we came to the technology a lot of work is being done for converting musical notation to synthetic voice for western music. But for St Yared hymn nothing has been done till now, and this is the inspiration for this thesis work. In this proposal, by using different approaches for capturing and analysis the dynamicity of each notation a text-to –hymn (melody) is hypothesized as a solution.

1.4 Objective of the Study

Objective of the research are stated as general and specific as the following

1.4.1 General Objective

The general objective of this study is to develop a model for text to hymn synthesizer for St Yared hymn notations.

1.4.2 Specific Objectives

The specific objectives of the research are to:

- Conduct literature review on TTS and notations
- Collect acoustic and text data that contains St Yared hymn notations
- Design a model for St Yared hymn notation Synthesis
- To develop a prototype system for notation synthesizer
- Test the performance of the system

1.5 Scope and Limitation

Digua has other notations beside the ten signs for hymn which are used for guidance for *Akuakuam,* scared dance accompanied by drum, taw cross, and drum and sistra combination with hymn. The scope of this research is to model a synthesizer only for the basic notations of St Yared hymn.

1.6 Significance of the Study

The proposed research will have significance in the following applications

- It simplifies the study of St Yared’s Hymn
- It shortens the duration for Digua and other studies that have hymn

- It can be used as bridge by decreasing the generation gap of students of *Abinet temari* and today's modern school students

1.7 Methodology

The main activities that are conducted to achieve the objective are stated as follows:

Literature review

In order to get sufficient knowledge to model text-speech-synthesis for St Yared hymn different documents were reviewed. The literature reviewed can be classified as works done on western song voice synthesis and local text to speech synthesis for Amharic languages with different techniques. In addition to this different approaches of text to speech synthesis techniques are reviewed in detail.

Data collection

For the modeling of the new synthesis for St Yared hymn notations a representative sample lyrics for each note with their acoustic detail was collected from different scripts that has hymn notation. The collected lyrics are transcribed to the basic 8 notations of St Yared hymns manually by the experts of digua. The transcribed lyrics are recorded with a voice of the experts to get the acoustic data for the synthesis. This was used to analyze the variation of speech sound of notations at different context.

Tools

Praat software is used for recording the voice with a sampling frequency of 44100 Hz and 16 bit resolution of data representation. Also extraction of notes from acoustic wave signals are segmented and saved to audio data set. In addition to that praat was also used for pitch and duration extraction.

Java programming language is used to process the input text. It used as a tool for a natural language processing steps in the text to speech synthesis: these are transcription of the input text, pattern generation of the transcribed text, searching a candidate unit for a given pattern based on the context it has, selecting appropriate unit from temporary candidate data set and selecting audio file from the audio data set for the selected units.

Matlab program codes are written for the synthesis of the wave signal for the concatenated units to give an appropriate pitch and intensity.

Testing

The system developed by the above tools was tested by conducting qualitative test using Mean Opinion Score (MOS) for questionnaires stated at appendix G. the testers are selected based on the knowledge they have in St Yared hymn. The same lyrics selected are assessed by giving them with voices produced by simple concatenation of notes regardless of their note marker and with notes that are synthesized using note marker.

1.8 Organization of the Thesis

The remaining part of this thesis is organized as follows. Chapter 2 deals with the literature review which includes the human speech production system and different techniques used for synthesizing speech. Chapter 3 deals with the parameters of music or song and introduction of St Yared hymn notes. Chapter 4 presents the design of model for St Yared hymn notation synthesis in particular the text analysis, song analysis, unit selection and wave synthesis. Experimental results are presented in chapter 5. Conclusion and recommendations are forwarded in chapter 6. References and appendices are given at the end.

Chapter 2

Literature Review

2.1 Human Speech Production System

Speech is produced when air flows from lung to exterior through mouse and noise. There are a plenty of physical components of speech production in human organs, mainly the following are listed. These are lung, trachea, larynx, pharynx, oral and nasal cavity and vocal folds as shown on the Figure 2.1. Human speech begins with the vocal cords (folds). Air forced up by the lungs passes over the vocal cords, causing them to vibrate at certain frequencies, depending on the force of the air and the position of the vocal cords. At this point the fundamental frequency of the speech is formed and then modified by the soft palate, tongue, lips, and other parts of the vocal tract, filtering out some frequencies and creating additional frequencies which the integral product of the fundamental frequency and it is known as formant frequency.

Depending on the movement of vocal folds speech is divided as voiced and unvoiced. When the air pressure passes through the vocal vibrates the folds a sound with high energy produced and actually this is known as voiced sound. While unvoiced signal cannot vibrate the vocal folds rather they produced by constriction of vocal tracts. In spoken phone there are vowels and consonants. Voiced signals are vowels and consonants are unvoiced.

The most energetic resonant frequencies of speech are called formants, and they are critical for vowel enunciation. Nearly every vowel can be characterized by two main formants that can be expressed as a numerical ratio. (The frequency of the first formant is between 200 and 1,000 hertz and the frequency of the second formant are between 800 and 3,000 hertz, depending on the vowel.)

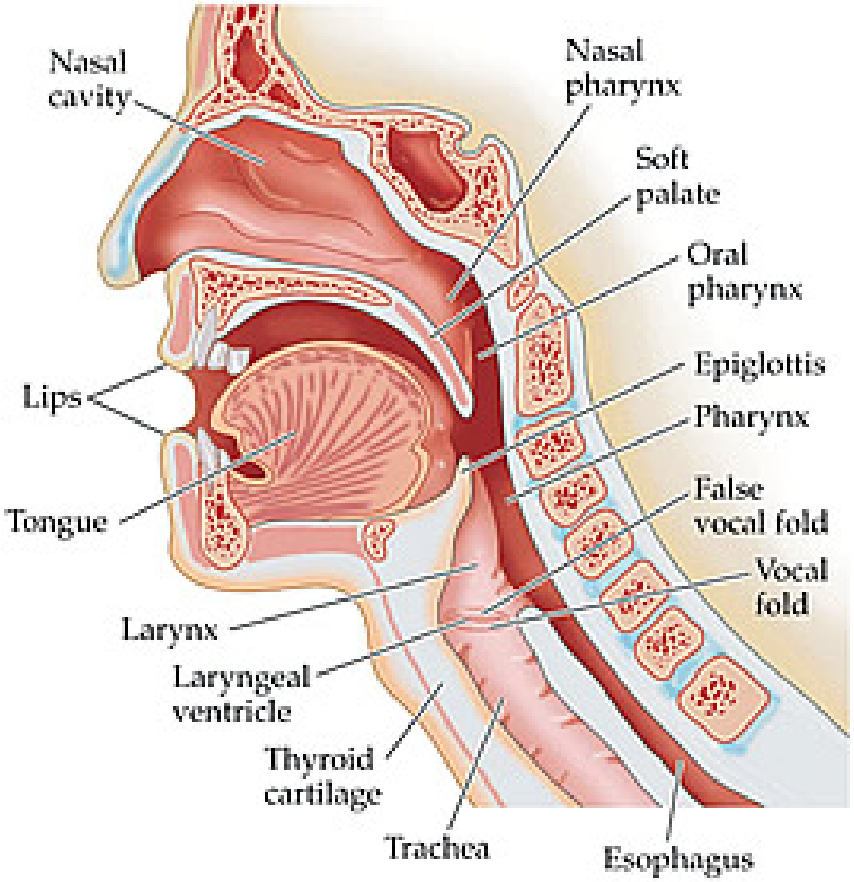


Figure 2.1 Human speech production organs (adopted from [29])

2.2 Text-to-Speech Synthesis

A process of converting written text to speech has majorly two components. These are Natural Language Processing (NLP) and Digital Signal Processing (DSP). In NLP phase the input text signal is converted to an intermediate output for the next DSP phase by transcribing and marking certain intonation and prosodic effects of the text. NLP is mainly subdivided into three major groups namely, Text Analysis, phonetic analysis and prosodic analysis as shown in the Figure 2.2 below.

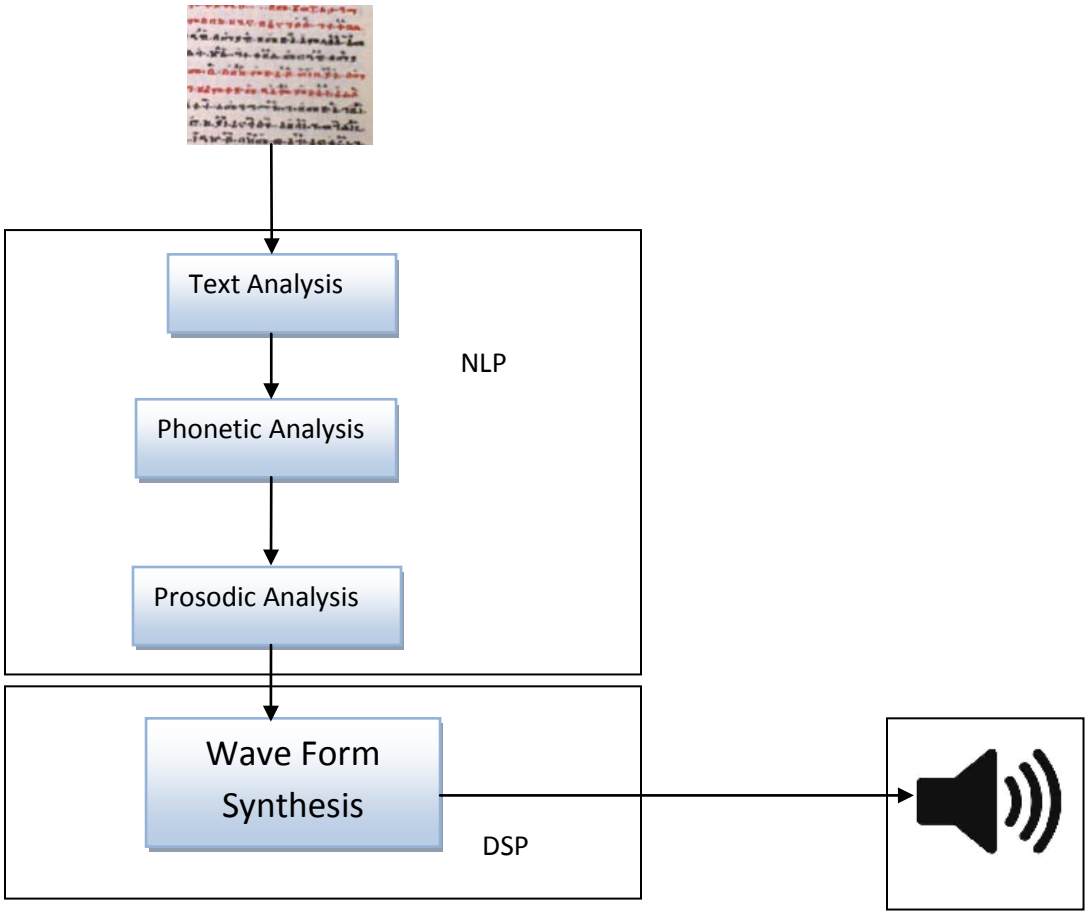


Figure 2.2 General Architecture of text to speech synthesis [2]

2.3 Natural Language Processing

The main activities in the natural language processing phase are text analysis, phonetics analysis and prosodic analysis. A brief description about each of them is given in the following sections.

2.3.1 Text Analysis

In this phase, the input text to the speech synthesis system needs to be processed and converted to a linguistic representation, which should be in a suitable form for the subsequent operations to work on. In addition to this, text analysis phase is responsible for indicating all knowledge about the text or message that is not specifically phonetic or prosodic in nature.

The first phase text analysis is the pre processing of input text. Input Text consists of numbers, abbreviations, acronyms and idiomatic and transforms them into full text when needed. Digits are converted to full text. For example in English 125 is transformed to "*one hundred and twenty*

five”, 1750 to *Seventeen Fifty* (if year) and *one thousand seven hundred and fifty* if (measurement). Like this also abbreviations has different equivalent full text representation. For example St. John Smith St., in this sentence the word St. had two full word representations and can be rewritten as Saint John Smith Street.

In preprocessing things like abbreviation and others all are to be converted to their corresponding form. In Geez language special in Degua it is common to write word in abbreviations, for example in book of mass (*kidassie*) the word *yika* (ይካ) is used for *yibel kahin* (ይበል ካህን). In the text analysis phase the raw input text is converted to its correct normalization form.

2.3.2 Phonetic Analysis

A process of changing orthographical symbols into phonological ones using a phonetic alphabet is known as phonetic analysis. Basically it is known as “grapheme-to-phoneme” conversion. Phone is the smallest sound unit, which has definite shape as a sound wave. A collection of phones that constitute minimal distinctive phonetic units are called Phoneme. Languages in the world have differed in their number of phonemes. For example the Number of phonemes in English is 44 of which 19 vowels and 25 consonants [7]. But for other languages such as Geez the number of phonemes is around 26*7=182 of which 7 are vowels and the rest are consonants.

In Phonetic analysis each letter is mapped to its equivalent sound phoneme. This can be done looking on the dictionary pronunciation in the simplest case for those words in the dictionary and had no more prosodic information. It seems easy to convert letters to sound by looking simply on the dictionary, but in reality it is not as such for different reasons. Among the major reasons the following are listed [8]:

- 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), which therefore have to be dealt with by a specific component of phonology, called morph phonology
- Pronunciation of words are strongly depends on the morphological analysis of the word. A word with the same letter character can have different pronunciation depending on the class of the word it used in the sentence, as for example the word “record” has two pronunciations when it used as verb and Noun.

- Sometimes consonants are omitted in the pronunciation in a sentence, as in ‘softness’ [s ow f n ih s] in which [t] fuses in a single gesture with the following [n].

- To simplify the aforementioned problems in Pronunciation of word, based on its spelling, has two approaches to do in a speech synthesis as:

- A dictionary is kept were It stores all kinds of words with their correct pronunciation; it is a matter of looking in to dictionary for each word for spelling out with correct pronunciation. This approach is very quick and accurate and the pronunciation quality will be better but the major drawback is that it needs a large database to store all words and the system will stop if a word is not found in the dictionary.
- A rather different strategy is adopted in rule-based transcription systems, which transfer most of the phonological competence of dictionaries into a set of letter-to-sound (or grapheme-to-phoneme) rules. This time, only those words that are pronounced in such a particular way that they constitute a rule on their own are stored in an exceptions dictionary. Notice that, since many exceptions are found in the most frequent words, a reasonably small exceptions dictionary can account for a large fraction of the words in a running text. In English, for instance, 2000 words typically suffice to cover 70% of the words in text [9].

By its nature Geez is a phonetic language, meaning that a simple grapheme-to-phoneme conversion is possible for most of the words due to close relationship with Geez orthography and phonology. Unlike languages like English, which are non phonetic, Amharic orthography does not have a difficulty as there is more or less, a one-to-one correspondence between the sounds and the graphemes.

2.3.3 Prosodic Analysis

The final stage in NLP is prosody generation. Prosody is a property of the speech signal that varies when speech uttered by a speaker on its pitch, loudness and duration of Syllable. In other way the concept of prosody is a combination of stress pattern, rhythm and intonation in a speech. Intonation describes how the speech signal varies during speaking. Pitch is used to convey this variation by using the fundamental frequency for expressing the speaker’s emotion on happens,

question raising and emphasis a part of a text. Intonation basically used as a measurement parameter for the output conversion of a text as how well it is intelligible and natural the synthesized speech is. Unfortunately, written text usually contains very little information of these features and some of them change dynamically during speech. In Figure 2.3 some factors that change the prosody of the speech are shown in category.

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 supra segmental features.

In the other hand, on timing at sentence level or grouping of words into phrases correctly is difficult because prosodic phrasing is not always marked in text by punctuation, and phrasal accentuation is almost never marked. If there is no breath pauses in speech or if they are in wrong places, the speech may sound very unnatural or even the meaning of the sentence may be misunderstood. For example, the input string "John says Peter is a liar" can be spoken as two different ways giving two different meanings as "John says: Peter is a liar" or "John, says Peter, is a liar". In the first sentence Peter is a liar, and in the second one the liar is Jo.

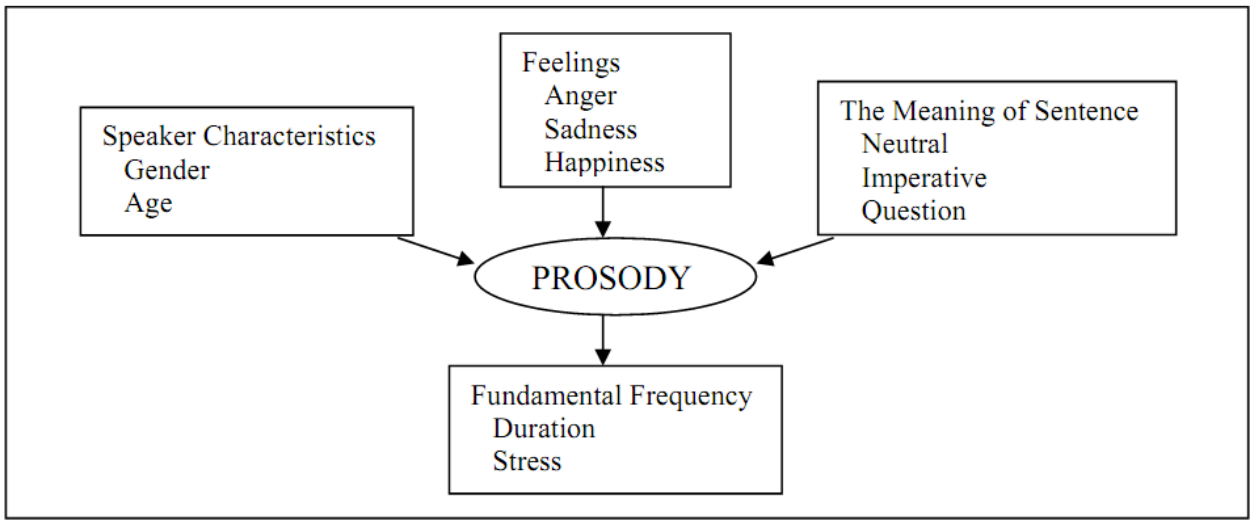


Figure 2.3 Prosodic dependencies (Lemmaty, 1999)

2.4 Digital Signal Processing (DSP)

Waveform generation is the process of using outputs of the linguistic text analysis and prosody models to generate an intelligible, natural-sounding speech waveform. This is an "ill-posed problem" in the sense that the linguistic feature input specifies only a tiny fraction of the information needed to specify the waveform. The rest must come from basic knowledge and models of a particular talker's speech based on speech production and perception principles and/or examples found in a speech corpus. Designing a waveform generation system involves specifying a speech model that can faithfully represent any speech sound, and specifying sets of model parameters to synthesize a particular utterance.

Traditionally, speech models have been lumped into three broad classes – articulatory models, formant models, and concatenative models – described below. Articulatory and formant synthesis are both considered parametric methods, which have the advantage of a compact representation for a wide range of voices. Concatenative techniques are less flexible, but tend to be less expensive and are a more natural approach for mimicking a particular voice [10].

On the other hand speech synthesis can be classified in to two as rule based and data-driven synthesis. Rule based synthesis is mostly in favor with phoneticians and chronologists, as they constitute a cognitive, generative approach of the phonation mechanism. The broad spreading of the Klatt synthesizer [11] for instance is principally due to its invaluable assistance in the study

of the characteristics of natural speech and synthesized speech. In rule based phonologists takes some rules to produce artificial sound where as in data-driven speech corpus are stored in the database and during speech synthesis part of speech waves are selected upon the phonetic expression is connected together. The former is used for formant synthesis and the later used for concatenative synthesis.

2.4.1 Articulator Synthesis

Articulatory synthesis is a phenomenon in which human physiological parameters are modeled to produce artificial sounds. It tries to model the human vocal organs as perfectly as possible in such a way that each synthetic speech will be similar to the natural speech produced by each vocal organ. The technique uses basically the following area of lip opening, constriction formed by the tongue blade, opening to the nasal cavities, average glottal area, and rate of active expansion or constriction of the vocal tract. The technique produces the best synthetic sound that resembles the natural sound even though construction of the parameters is too difficult.

2.4.2 Formant Synthesis

A type of rule based speech synthesis which simulates human voice by using source filter models. In formant synthesis basically three formant frequencies are used, to get high quality output one can use five formants. A formant is a concentration of acoustic energy around a particular frequency in the speech wave. Each formant corresponds to a resonance in the vocal tract. By modeling these formant frequencies one can simply produce speaker-dependent synthetic sound by using formant rule database.

Each formant is usually modeled with a two-pole resonator which enables both the formant frequency (pole-pair frequency) and its bandwidth to be specified [12]. There are basically two types of structures these are cascade and parallel where formant frequencies and bandwidth are connected to achieve the signal wanted to produce.

Formant Synthesis uses the rules to modify the pitch, formant frequencies, and other parameters from one sound to another while maintaining continuity present in physical systems like the human production system.

2.4.3 Concatenative Synthesis

Concatenative Synthesis is where a prerecorded utterance of speech is kept in a speech database is used to form a new sound by simple connecting of the selected utterances. So it sometimes called data-driven synthesis. It needs a huge database to accommodate big volume of languages utterance. This is in fact the drawback of this method of synthesis. In other hand it is known by its production of natural sound. So that most of simple application for telecomm-unction and announcement prefers this method. Concatenative synthesis can be also subdivided into diaphone base and Unit selection.

Diphones

Diphones are defined to extend the central point of the steady state part of the phone to the central point of the following one, so they contain the transitions between adjacent phones. That means that the concatenation point will be in the most steady state region of the signal, which reduces the distortion from concatenation points. Another advantage with diphones is that the coarticulation effect needs no more to be formulated as rules. In principle, the number of diphones is the square of the number of phonemes (plus allophones), but not all combinations of phonemes are needed. For example there are 43^2=1849 hypothetically possible diphone combinations for English. Not all of these diphones can actually occur. For example, the rules of English phonotactics rules out some combinations; phones like ‘h’, ‘y’, and ‘w’ can only occur before vowels. In addition, some diphone systems don’t bother storing diphones if there is no possible coarticulation between the phones, such as across the silence between successive voiceless stops. The AT&T 43-phone has only 1162 diphones rather than the 1849 hypothetically possible set [13].

There are certain common rules for diaphone synthesis to concatenate diphones. These Diphone syntheses can be characterized by the following steps:

Training:

1. Record a single speaker saying an example of each diphone.
2. Cut each diphone out from the speech and store all diphones in a diphone database.

Synthesis:

1. Take from the database a sequence of diphones that corresponds to the desired phone sequence.

2. Concatenate the diphones, doing some slight signal processing at the boundaries.
3. Use signal processing to change the prosody (f0, duration) of the diphone sequence to the desired prosody.

Unit Selection

Even though Diphone waveform synthesis has advantages for low memory requirements and time for recording about 1500 diphones. But it suffers from two main problems. First, the stored diphone database must be modified by signal process methods like PSOLA to produce the desired prosody. Any kind of signal processing of the stored speech leaves artifacts in the speech which can make the speech sound unnatural. Second, diphone synthesis only captures the coarticulation due to a single neighboring phone. But there are many more global effects on phonetic realization, including more distant phones, syllable structure, the stress patterns of nearby phones, and even word-level effects.

So in current time commercial Synthesis used unit selection synthesis is coming to eliminate the aforementioned problems of diphone waveform synthesis. Units used are usually words, syllables, demy-syllables, phonemes, diphones, and sometimes even tri-phones. There is a tradeoff between selections of unit of data for concatenation. With longer unit high naturalness, less concatenation points and good control of co-articulation 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.

The process of building the unit inventory incorporates three main steps. First, the natural speech must be recorded so that all used units (phonemes) within all possible contexts (allophones) are included. After this, the units must be labeled or segmented from spoken speech data, and finally, the most appropriate units must be chosen. Gathering the samples from natural speech is usually very time-consuming. However, some of this work may be done automatically by choosing the input text for analysis phase properly. The implementation of rules to select correct samples for concatenation must also be done very carefully. There are several problems in concatenative synthesis compared to other methods which are listed below.

- Distortion from discontinuities in concatenation points, which can be reduced using diphones or some special methods for smoothing signal.

- Memory requirements are usually very high, especially when long concatenation units are used, such as syllables or words.
- Data collecting and labeling of speech samples is usually time-consuming. In theory, all possible allophones should be included in the material, but trade-offs between the quality and the number of samples must be made.

Since some of the problems are alleviated with methods described below the use of concatenative method is increasing due to better computer capabilities.

2.5 Pitch Synchronous Overlap Add Technique (PSOLA)

Assume we have two diphones to concatenate, what will occur if the waveforms of the two diphones edges across the juncture are very different? We hear a perceptible click from the generated result. That is why we need some techniques to avoid unwanted signal during concatenation. For that one can apply a windowing function to the edge of both diphones so that the samples at the juncture have low or zero amplitude. Furthermore, if both diphones are voiced, we need to insure that the two diphones Joined pitch-synchronously. This means that the pitch periods at the end of the first diphone must line up with the pitch periods at the beginning of the second diphone. Otherwise the resulting single irregular pitch period at the juncture is perceptible as well.

Now given our sequence of concatenated diphones, how do we modify the pitch of prosodic requirements? The PSOLA technique is solution for this problem. 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 [12].

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 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 units are concatenated pitch synchronously 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 re-synthesize the final waveform through an overlap-add operation [14]. 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: the analysis and synthesis

Analysis phase:

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 2.4) 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).

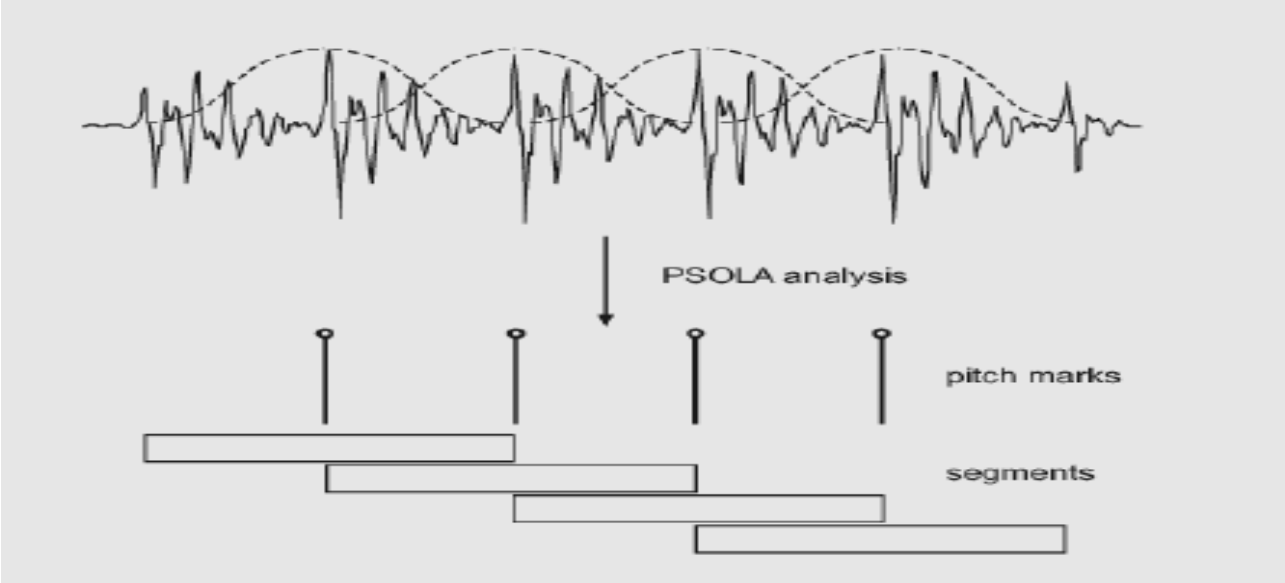


Figure 2.4 The PSOLA pitch analysis (adopted from [23])

Synthesis Phase

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 [23]. Figure 2.5 shows the steps in speech synthesis using PSOLA technique.

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.

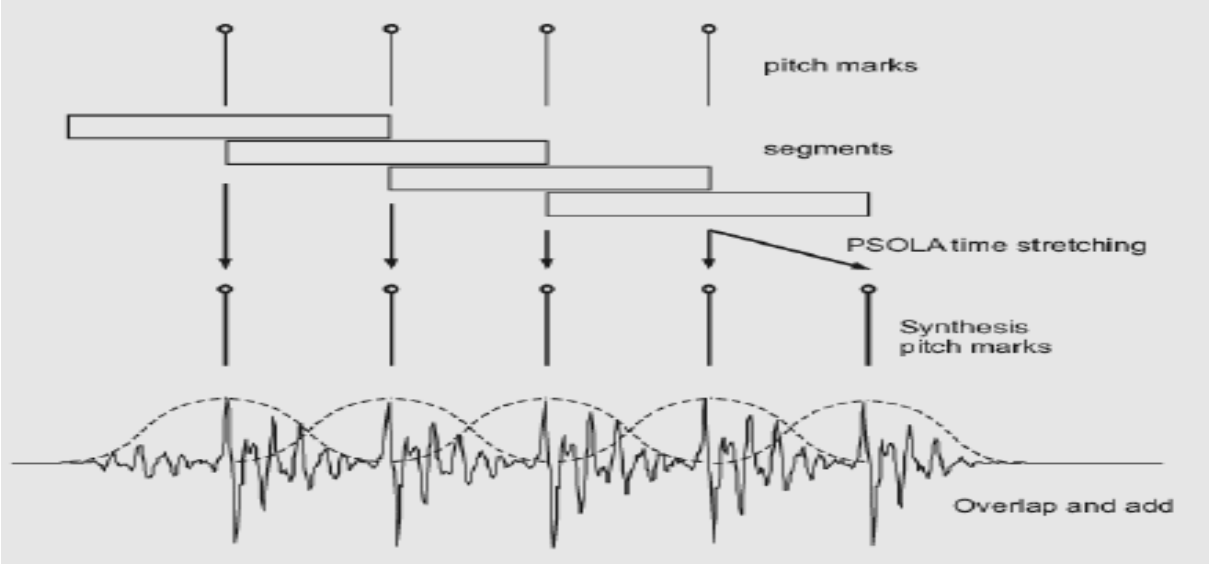


Figure 2.5 The PSOLA Synthesis (adopted from [23])

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 smoother windowing functions like Hamming and Hanning is applied. 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-emphasized and the effects of the edge are reduced efficiently. The widowing signal and function are shown in Figure 2.6 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.

2.6 HMM Based Synthesis

One of the drawbacks of concatenative synthesis is large memory capacity needed for database and it is only for a single talent or a few speakers. In addition to this it has a problem with units

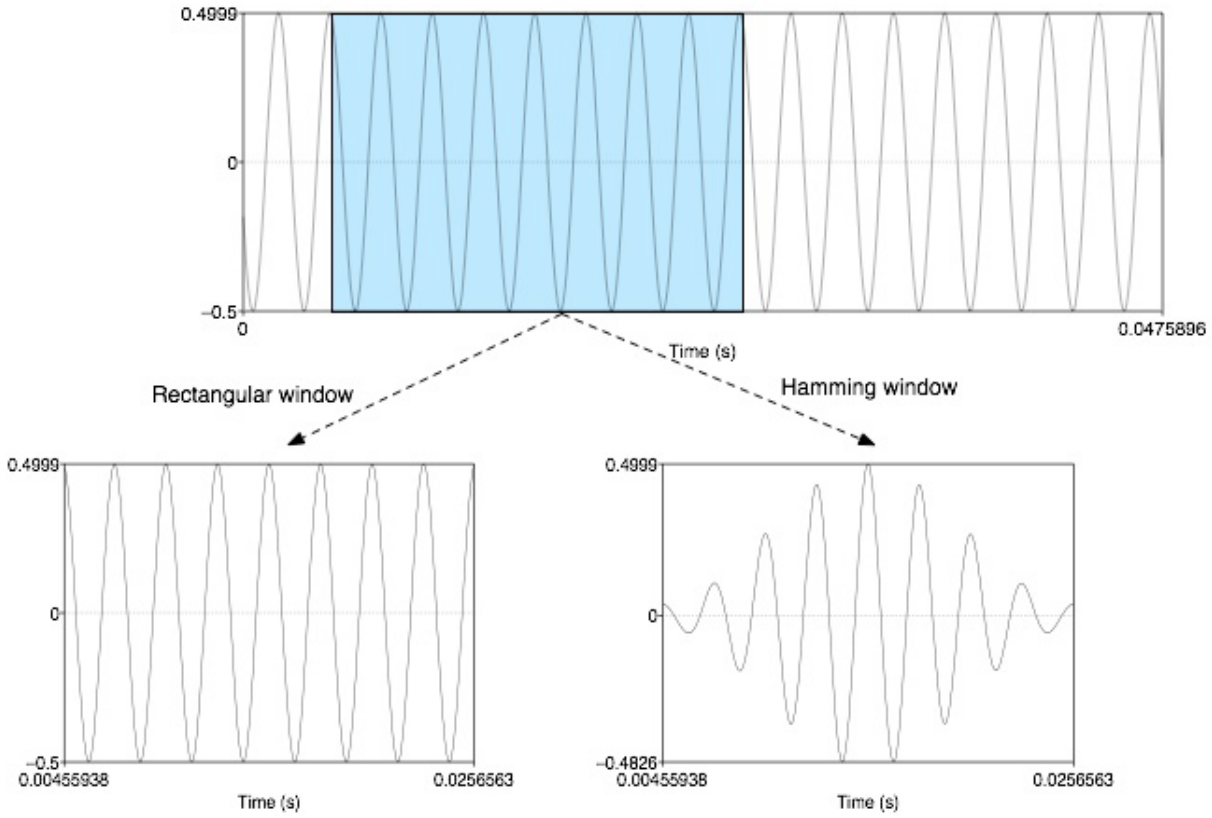


Figure 2.6 Windowing function

which are not in a database. An alternative is to use statistical models, sometimes called as machine learning techniques. This and the concatenative approach can both be described as data-driven. In the concatenative approach, the data are directly used; whereas in the statistical approach we are attempting to learn the general properties of the data. Two advantages that arise from statistical models are that initially it requires less memory to store the parameters of the model than memorize the data. Secondly, one can modify the model parameters in various ways so as to get a voice with different melody and prosody [15].

HMM-based speech synthesis is one of the most powerful statistical methods for modeling speech signal [16]. In this method, the frequency spectrum (vocal tract), fundamental frequency (vocal source), and duration (prosody) of speech are modeled simultaneously by HMMs and also the speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion [17].

Speech synthesis using HMM works by considering the internal speech production to be a sequence of hidden states, where each state models segment of a speech, and the resulting sound to be a sequence of observable states that at best approximates the (hidden) states [17]. However, there are so many different combinations of hidden states that results in a given observation so selecting the one that is the appropriate one is an issue.

2.7 Related Works

A lot of work has been done on the topic of text to speech synthesis since the early known mechanical experiment of Wolfgang von Kempelen's Speaking Machine. It is a manually operated speech synthesizer that began development in 1769. The machine was constructed to resemble human vocal tract organs. It works by a combination of several keys, one for each letter. The sounds were produced by a common bellows that fed air through various pipes with the appropriate shapes and obstructions needed to produce that letter.

The second major work on this area came after the technology shift to electronics, where VODER (Voice Demonstrator), a Bell Telephone laboratory product invented by Homer Dudley in 1938. It works as operator sitting behind the console, using keyboard and pedal shown in Figure 2.7 below. Phrases resembling human speech could be demonstrated to the audience, although the produced sounds were often difficult to understand [7].

When we came to our narrow topic, things similar to this thesis report topic has been conducted by different authors on different sides of the world on singing voice synthesis. Among these the following are selected to be reviewed.

The research conducted by Jhon McNuty [18], which is a model to capture live performance for singers to background with synthesized songs. The previous product namely Vocaloid is known to its Synthesized Singing voice but limited to studio works. In this research they used to run their application on Macintosh OS X by using an embedded speech synthesizer as a TTS engine. They took MIDI melody and score text as input for aligning intelligently the phoneme of the text with the MIDI notes. The system produces both speech and song. They use a diphone concatenating technique to produce the output. To produce a song they add pitch and duration information of the note on the TTS.

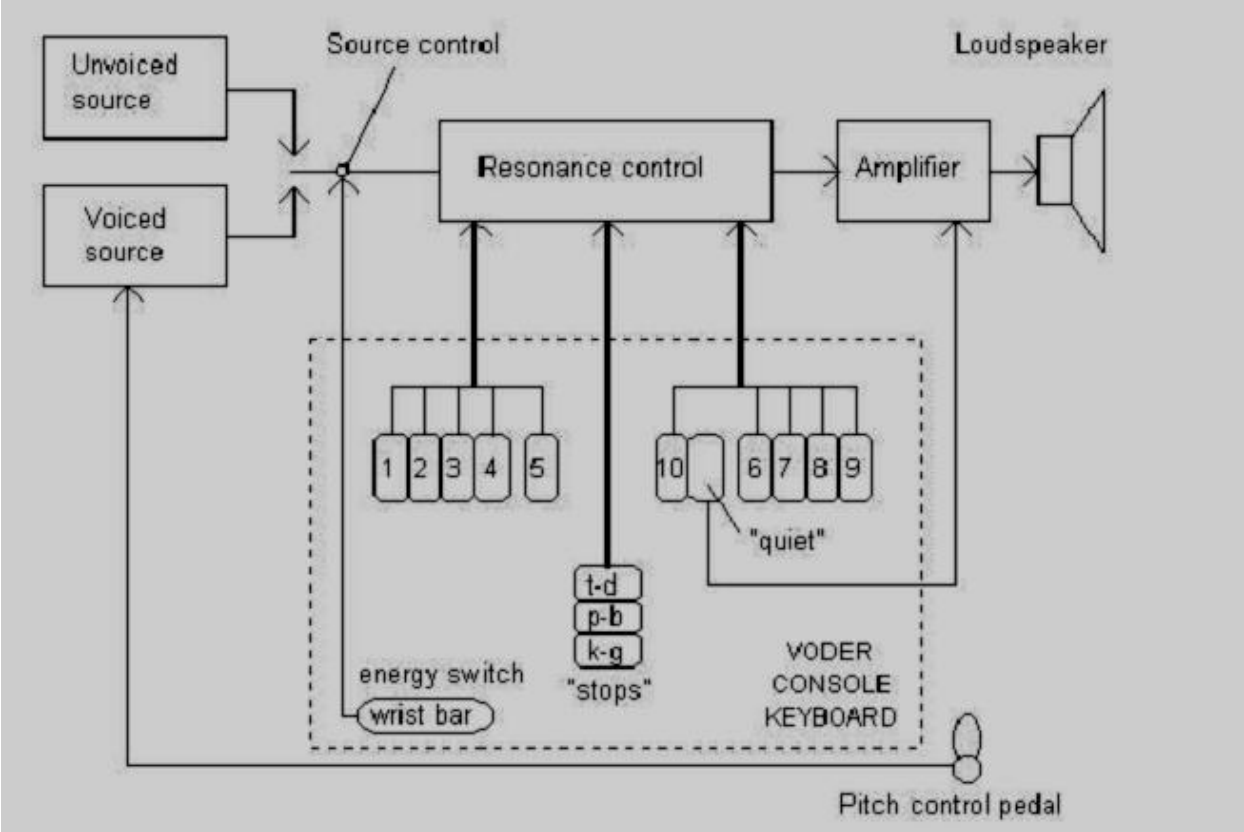


Figure 2.7 VODER Machine block diagram (from [7])

They develop an algorithm for the Phonetic Pitch and Timing Reassignment (PPTR). The goal of the Algorithm is to take speech phonemes derived from the TTS engine and then to Intelligently assign new pitches and durations to each, the reassignment being based on the MIDI melody.

Spanish singing voice synthesis using TD-PSOLA (Time Domain- Pitch Synchronous Overlap and Add) is described on their paper [19]. The researcher took a Mexican singer voice with Spanish language and constructs a syllable database for testing. Matlab software was selected to accomplish their work. They used **tdpsola** function to mark pitch and calculate distance between pitchs. By doing Syllable modification on tone and duration they confirmed their hypothesis which says “a syllable is the best suited units for concatenation singing voice synthesis”.

On the work of [20] an effort taken to synthesize Greek songs from written score to song was described. They used MBROLA (Multi-Band Re-synthesis Over Lap Add), A software product that used for speech synthesis by accepting diphone as input. They wrote their music on score editor and saved as midi file. But for the sake of MBROLA doesn't accept midi file they construct a converter from midi-file to phonetic-file expression.

They add a conversion module between midi file and phonetic file description by using lyrics messages, followed by Note On and Note Off messages to obtain the Syllable from Lyrics message and note number and delta time from Note On and Note Off messages. Issues regarding long vowels, portamento and pause are considered for singing voice. They used *mbredit* tool, which is provided by MBROLA Project, for modifying pitch, fundamental frequency and duration. In order to achieve a better SVS result they create a new diphone database for Greek language because the one that occur in MBROLA project is constructed for text to speech Synthesis. Evaluation for the new database scores MOS value above average.

When we come to local research works done so far on TTS we can get a lot of papers done on local languages especially in Amharic language. In other hand up to the knowledge of the writer there is no work done on local language Singing voice Synthesis. Since the topic of this research can be classified in a middle rank between Text-to-Speech and Singing voice Synthesis for its behavior for sharing most of the NLP phase analysis in addition to certain synthesis parts. So it is helpful to look certain research papers written. For this reason the following research works are selected.

Nadew Tademe [21], makes a great effort towards synthesizing Amharic Text-to-speech system by considering the vowels from the spoken words. They took different sounds for each vowel in a context dependent manner to model their new system. For this purpose they collect 500 words and use 12 persons to spoke it. From these words they select 800 vowels for their formant speech database.

They construct a database for storing information of specific vowels depending on the context they are in. They consider a context as for each vowel they took two neighbors from left and

right side of the target. After that they took three formants for the given vowel for a period of 20ms signal with their respective bandwidth then the fundamental frequency and sampling frequency are parameters to be saved. In the synthesis phase they took the appropriate vowel parameters from the database depending on the CART result for the transcribed phonetic input, it generates a new waveform for the given vowel.

Berket Kasaye [22], in his work proposed the ability of Hidden Markov Model based speech synthesis for Amharic language. The system they presented has two parts: The training and synthesis part. The training part includes data preparation, language modeling, feature extraction, and building the HMM. Whereas, the synthesis part includes preparing labeled text from the text input, selecting appropriate HMMs, extracting speech parameters from HMMs, and finally generating the speech waveform from the speech parameters. By considering the transcribed text they adjust it by looking on the wave form since there are certain vowels which are omitted in transcription. In other words what we call ephentic vowels for the six order letter for Amharic. They prepare a possible question set for decision tree clustering while making HMM decision tree.

Mulat Shiferaw [23] stated in his work that Amharic texts can be synthesized by using syllable based unit selection for concatenative speech synthesis method. They consider the way to handle epenthesis and gemination before segmentation of Amharic words into syllables during text analysis phase. Mainly their work handles issues which is necessary to improve the naturalness of waveform generation by the synthesizers', corpus-based concatenative synthesis with Syllable speech segment. They prepare a database for storing unit information. These are waveform of units, phone identities of the units, phonetic context and prosodic annotations for the units. A TD-PSOLA speech waveform analysis-synthesis algorithm is used for prosodic modification.

As mentioned earlier, we saw text to speech synthesis for local language using different synthesis techniques. No research is conducted for St Yared hymn synthesis from text to song. This work is possibly the first attempt towards development of the synthesizer for St Yared hymn.

Chapter 3

Hymn and Musical Notations

3.1 Introduction

A hymn is a type of song, usually religious, specifically written for the purpose of praise. These hymns are used early days in a church is accompanied with some classical instruments. The way it is used differs from one church to others. Since hymn as a broad classified under music so it is necessary to get some basic understanding of music. These are Composition of music, how to write musical Score and basic music parameters.

There are three major parameters of musical tones. These are Pitch (frequency), Tone quality and intensity. The Pitch of musical note refers to how high or low the note in the overall pitch register. Bass notes are lower in the pitch register than treble sounds are. Pitch is the ear's perception of the wavelengths of the sounds being produced.

Intensity sometimes called the volume, or how loud or soft the sound is. While the frequency is governed by the length of the sound waves, the intensity is governed by their height. The wave height can also be referred to as the wave amplitude. And the third parameter is the tone quality which is called the tone color or timber; it is the property that enables the ear to distinguish between the sounds of, say, a flute and a violin playing the same note. The tone of violin has richness and a warmth compare to the tone of the flute, which is smooth and less complex [24].

3.2 Music Notes

Modern music notes are developed from the early musical notes of French denoted as below and known as *solfeggio*. And the original phrases where the notes are formed are listed below. And the **Ut** was replaced by **Do** and **Si** replace by **Ti** to make it easier to say on the Romance language as "Do-Re-Me-Fa-So-La-Ti".

- **Ut** queant laxis
- **Re** sonare fibris

- **Mi** ra gestorum
- **Fa** muli tuorum
- **Sol** ve poluti
- **La** bi reatum
- **S** ancte **I** oannes

For representation of the conventional seven notes an English letters from **A** to **G** is assigned. The assignment of letters to the note has two variants of Sol-fa: Fixed doh and Movable doh. On a fixed Doh scale the **A** note is assigned to **La** because of the precision of the frequency(440 Hz) which corresponds to such note and end up with a correspondence as follows [25]. The seven notes of solfege with its frequency with the Key of G as shown in Table 3.1.

Table 3.1 The seven notes of solfege with its frequency in the key of G

key	G	A	B	C	D	E	F	G
name	DO	RE	*MI*	FA	SO	LA	TI	DO
Frequency (in Hz)	392	440	*494*	523	587	659	740	784

In movable-doh, you can pick a different pitch to start on and sing Do-Re-Mi-… starting at the note. What was doing is that singing the major scale in different keys. If we sing in C major Do-Re-Mi-Fa-So-LA-Ti becomes C-D E-F-G-A-B. In general we can say the Sol-fe starts with the letter corresponding to selected key major. We can say fixed-doh is A major since it starts with key A.

In the standard musical writing system there is a Staff as shown in the Figure 3.1. It is the fundamental latticework of music notation, upon which symbols are placed [30]. The five stave lines and four intervening spaces correspond to pitches of the diatonic scale– which pitch is meant by a given line or space is defined by the clef. In staff notes form a series to handle different pitches into music. So it repeats itself in a series as shown in Figure 3.2:

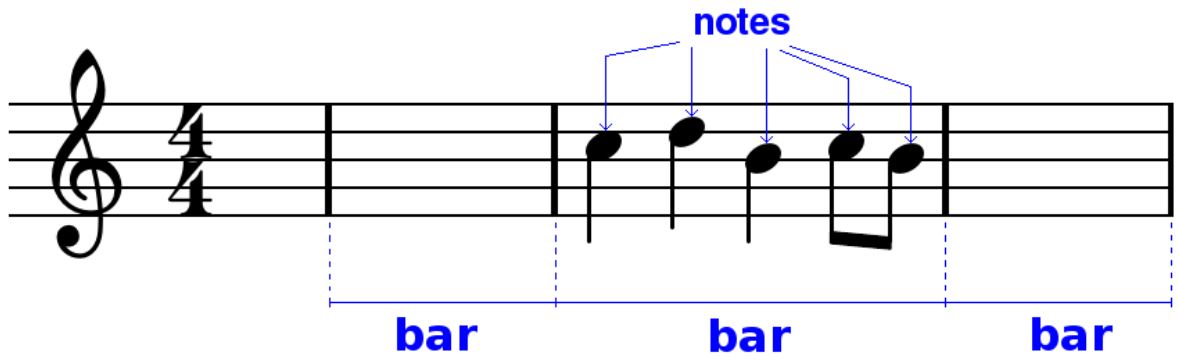


Figure 3.1 Basic staff components

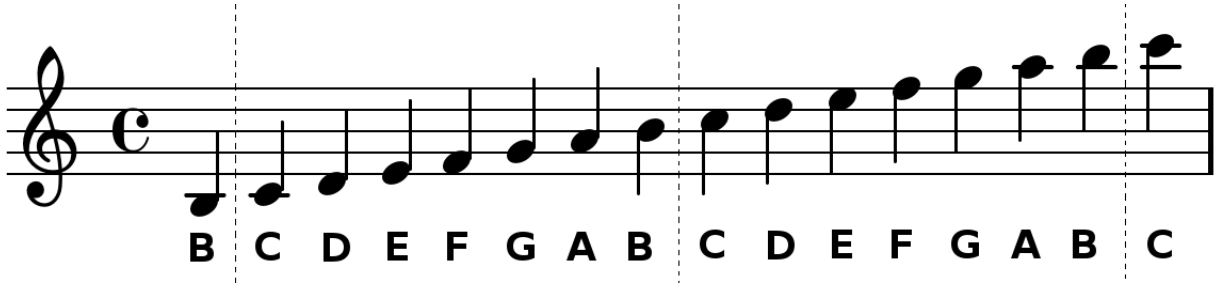


Figure 3.2 Series of notes

Since there are several A's in the score it is difficult to know about which A note we are talking, so it is necessary to add additional sign to differentiate, for example when we got A4 it is easy to say this note is A note in the fourth octave. There are 10 octaves from octave 0 to octave 9. Each octave contains one series of note that is from C4 to B4. Figure 3.3 shows four octaves from octave 3 to octave 6. Octave in a music regarding to frequency means a double in a pitch(frequency), for example A5 is 880 Hz which is two times A4 (440 Hz)

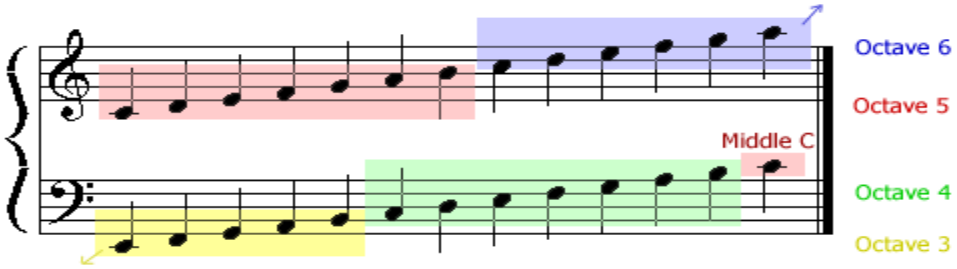


Figure 3.3 Part of octaves of the notes

The other parameter of music is duration of the notes. Duration tells as the period for the specific note to be sustained in the pitch specified. Actually notes are represented by its duration rather than note name. The durations are symbolically shown as in the Figure 3.4

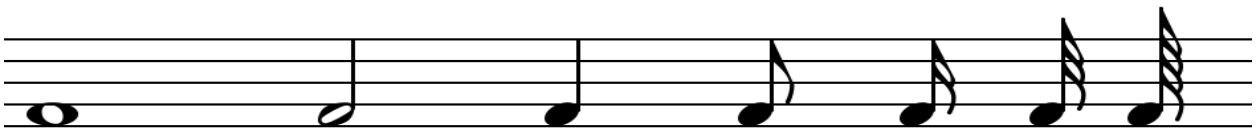


Figure 3.4 Duration of the note

Description of duration from left to right

- Whole note
- Half: it lasts half of the whole time
- Quarter: it lasts ¼ of the whole time
- Eighth: it lasts for 1/8 of the whole
- Sixteenth: lasts for 1/16 of whole
- Thirty-second: lasts for 1/32 of whole
- Sixty-fourth: lasts for 1/64 of whole

3.3 St Yared Hymn

St. Yared, the great Ethiopian scholar, was born on April 5, 501 A.D. in the ancient city of Aksum. His father was Yisak (Isaac) and his mother, Kristina. Yared received educational and Moral guidance from his uncle Gedewon who was then reputed to be a scholarly priest. Moreover, it is claimed that Yared was taken to Heaven where he was taught by three Holy Spirits, the arts of vocal performance, composition, poetry, versification and improvisation. Yared arranged and composed hymns for each season of the year, for summer and winter and spring and autumn, for festivals and Sabbaths, and for the days of the Angels, the Prophets, the Martyrs and the Righteous Yared often sang for Emperor GebreMeskel. "And when they heard the sound of his voice, "the king and the queen, and the bishop and the priests, and the king's nobles, ran to the church, and they spent the day listening to him." And one day St. Yared sang in front of Emperor Gebre Meskel accompanied by drums, sistra, and male priests. Mesmerized by

the music, the Emperor accidentally dropped his spear into the flat part of Yared's foot. The Emperor was grieved by the pain he had inflicted on his spiritual friend. He said: "Ask me whatever reward you wish in return for this thy blood which hath been shed." Yared promised the Emperor not to refuse his requests. Having accomplished that, Yared asked and was reluctantly granted permission to live in solitude and to dedicate his life to prayer, meditation, and to his music. He departed from Axum and went to the Semien Mountains where he lived until his disappearance [26].

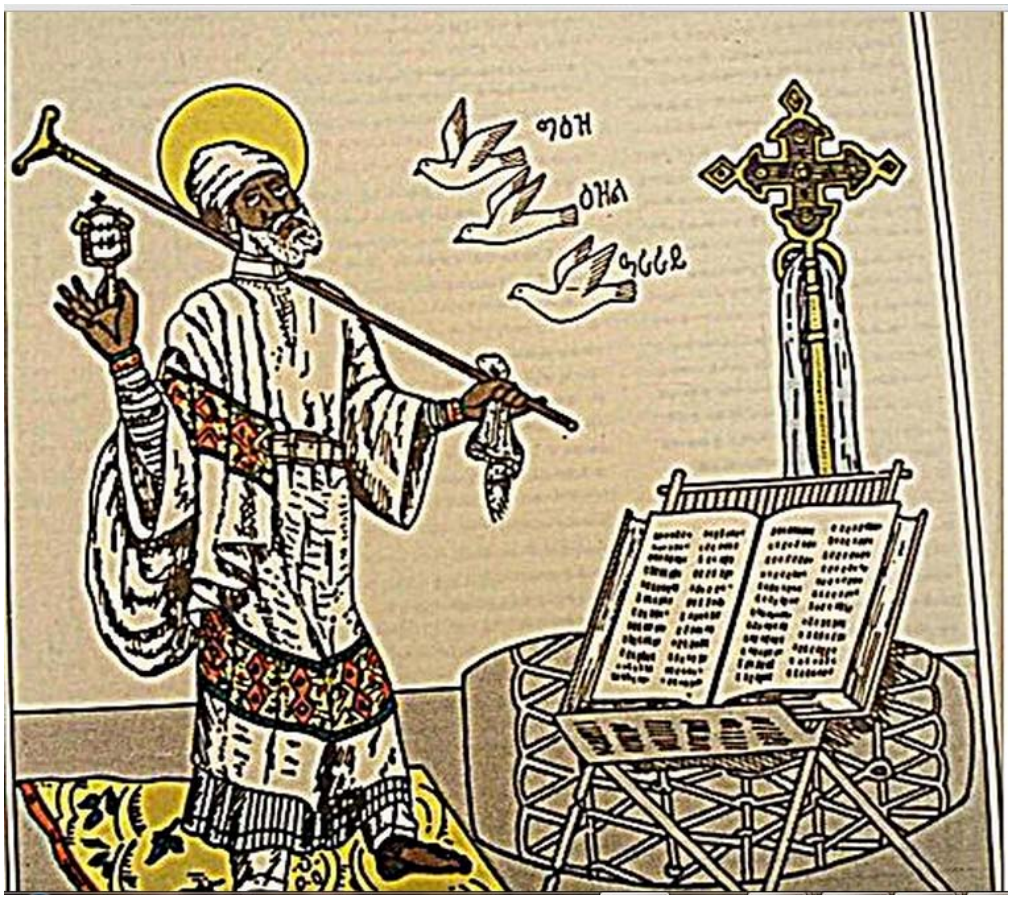


Figure 3.5 An artist rendering of St Yared while chanting *Zema* accompanied by sistrum, tau-cross staff, adopted from [5]

3.4 St Yared Hymn Notations

St Yared hymn as stated earlier in chapter 1 has eight notes in number. Namely *yizet, deret, rikrik, difat, chiret, kinat, hidet* and *Kurt. Rikrik* and *hidet* has two forms which can make the symbols ten as shown on Table 3.2. Basically these symbols are the core notes of the St Yared hymn which had the capability to describe all his songs when they appear on the letters. Actually the real note of St Yared describes the way the sounds are produced from our articulatory organs. Each note has its own meaning in religion. The table shown in Table 3.3 describes the notes with their description.

Table 3.2 St Yared hymn Notes

Name	yizet	deret	rikrik	rikrik	Difat	chiret	kinat	Hidet	kurt	hidet
Symbol	ꙮ	ꙭ	꙯	꙯	ꙫ	Ꙫ	ꙩ	Ꙩ	ꙧ	Ꙧ
Unicode	1390	1391	1392	1393	1394	1396	1395	1398	1399	1397
Code	1	2	3	4	5	6	7	8	9	10

The note has a context dependent sound. That is why it needs additional descriptors for correctly interpreting. As the western music described in the above sections has seven notes A to G and have additional descriptor which is the digits from 0 to 9 for respective octaves. St Yared hymn notes have also seven descriptors which are namely: *Ilete birhan* (ዕለተ ብርሃን), *Aynu zerigb* (ዓይኑ ዘርግቦ), *Bubay* (ቡባይ), *BeInku senpier* (በዕንቁ ሰንፔር), *Selasa* (ሠላሳ), *Burikt* (ቡርክት) and *Tsega* (ጸጋ).

It is these descriptors which told the singer to produce the correct pitch of the song. At this point we can say that notes are used to decide how to produce the sound in its articulatory organ while the descriptors used to produce the amount of vibration for the produced sound, we call this fundamental frequency and in music the pitch of the sound.

So for the sake of writing the score of song we can use the following parts of each word as shown in the Table 3.4. Here the table shows also the equivalent marker for English by using the first letter of each word in caps lock. For example the following part of song script is written as, "ወ◌ፔር◌ቡበልዴ" where ⍰ had two notes deret (◌) and difat (◌) with marker of ፔር and ቡበ respectively.

Table 3.3 St Yared notes with their description

No.	Note	Description
1	Yizet(ꙵ)	When letters or words are emphasized with louder chant in another wise regular reading form of chant
2	Deret (ᑌ)	Is a form of chant that comes from chest
3	Rikrik (፧)(፡)	Is layered and multiple chants conducted to prolong the chant
4	Difat (ᑐ)	Is a method of chanting where the voice is suppressed down in the throat and inhaling air
5	Kinat (ᓭ)	Is the highlighted last letter of a chant
6	Chiret(ᕽ)	It highlights with louder notes letter or words in between regular reading of the text. The highlighted chant is conducted for longer period of time.
7	Hidet (ᓵ)(ᓴ)	It is a chant by stretching one's voice; it is resembled to a major highway or a continuous water flow in a creek
8	Kurt (ᑋ)	Is a break from an extended chant that is achieved by withholding breathing

Table 3.4 Root words from hymn with their part for tagging for Geez and English

No.	Words		Parts for Geez	Parts for English
1	**ዕለተ ብርሃን**	***I**lete birhan*	ዕለ	I
2	**ዓይኑ ዘርግቦ**	***A**ynu zergb*	ዓይ	A
3	**ቡቢይ እርዳቢይ**	***B**ubay irdabey*	ቡቢ	B
4	**በዕንቁ ሰንፔር**	*Binku sen**P**ier*	ፔር	P
5	**ሠላሳ**	***S**elasa*	ሧ	S
6	**ቡርክት**	*Bu**R**kt*	ቡር	R
7	**ጸጋ ዘተውህበ**	***T**sega zetwuhibe*	ጸጋ	T

The following guide line described in Table 3.5 is used for correct transcription of notes in English. It shows the possible combination of notes with markers which is eight notes with

seven markers gives 56 combinations as sample of them shown in Table 3.5 whereas the full representation is shown in appendix A

Table 3.5 Sample of note marker combination

Marker	Note	Symbol	name	description
፫	▪	S1	Selasa yizet	S for Selasa and 1 for Yizet
፫	◡	S2	Selasa deret	
፫	⁝	S3	Selasa rikrik	
፫	◠	S5	Selasa difat	
፫	ˏ	S6	Selasa chiret	
፫	‿ˏ	S7	Selasa kinat	
፫	⌣_	S8	Selasa hidet	
፫	⊢	S9	Selasa kurt	
ፒር	▪	P1	Pier yizet	P for Pier and 1 for Yizet
ፒር	◡	P2	Pier deret	
ፒር	⁝	P3	Pier rikrik	
ፒር	◠	P5	Pier difat	
ፒር	ˏ	P6	Pier chiret	
ፒር	‿ˏ	P7	Pier kinat	
ፒር	⌣_	P8	Pier hidet	
ፒር	⊢	P9	Pier kurt	

The part of a song described above can be expressed in Latin (English) by using equivalent code for notes and equivalent marker for parts as follows: weP2B5ld for □◡ፒር◠ቡበልድ and roS1S6maS1nA2 for ሮ·፫ˏ፫ማ·፫ን◡ዓይ.

3.5 Types of St Yared hymn

Yared completed these compositions in nine years. All his compositions follow the three musical scales (*kegnit*), which he used to praise, according to Ethiopian tradition, his creator, who revealed to him the heavenly chants of the twenty-four heavenly scholars. Each of these

categories are further classified with three musical scales (*Kegnitoch*) that are reported to contain all the possible musical scales:

- Geez, first and straight note. It is described in its musical style as hard and imposing. Scholars often refer to it as dry and devoid of sweet melody.
- Izel, melodic, gentle and sweet note, which is often chanted after Geez. It is also described as affective tone suggesting intimation and tenderness.
- Ararai, third and melodious and melancholic note often chanted on somber moments, such as fasting and funeral mass.[5]

Chapter 4

Model for Hymn Notation Synthesis

4.1 Introduction

In order to model hymn notations it is necessary to get full understanding of music as stated in previous chapter. In this chapter each steps done to design the model for St Yared hymn notations is described in detail in the following sections.

4.2 Methodological Approach

To design a model for the St Yared hymn notation synthesis, the following important issues are taken into consideration. The initial point for this model is collection of hymn. Hymn is collected from different religion books especially *digua* and mass (*kidassie).* A great care is taken while collecting the texts, which are thought to be important for extracting the required notes. By taking random selection of the hymn words that considered to be representative to model the system. The following ways are considered to perform the aforementioned task. The first mechanism is to select songs that hold the notes of *araray* hymn in their written form. And the second way is selection of songs in Geez Hymn. This is done to accommodate the seven note markers which are *Bubay, Burikt, Aynu, Selasa, Pier*, *Tsega* and *Ile*. But, this was a bit difficult point of the data collection, because songs written in a digua has not specified all notes needed rather it puts special tags. Tags are a kind of code which had one or two letters and used as an index for the other word in different song, which has a more described note on it. For example we may get ወ◌ ልድሰገእጉ◌የጢስ. Where ሰገ ጢስ stands for ሰገደ ጢስ and the notes on them become: ሰገ for ▪◌◌◌◌◌ and ጢስ for ▪.

The second core point is differentiating among the note marker of the given note. That is an individual note on a single letter had its class of note marker as described in chapter 3. This task was a bit difficult, since it needs to song each note in a repeatedly fashion to classify its marker. For the above note sequence, which simply identified by ሰገ on the Geez character ድ as ድሰገ, their classification of note marker is as shown in Table 4.1. After collecting a sufficient representative sample of the songs for text files, the phone dataset of the words is prepared. Extraction of

features, representation of the extracted features, training, and testing is made accordingly on the collected speech file.

Table 4.1 Classification of note marker for ሰገ

symbol	▪	ᴖ	ᴗ	ˏ	ᴗ	ᴖ	ᴗ	ˏ
class	*Aynu*	*Ile*	*Aynu*	*Bubay*	*Bubay*	*Aynu*	*Bubay*	*Selasa*
combination	A1	I5	A2	B6	B2	A5	B2	S6

4.3 System Architecture

Currently, there are many techniques that are used to synthesize speech artificially. Most of these techniques use small units, like phonemes, of the speech to synthesize artificial speech. But a given phoneme cannot have the same sound when it appears in different context of a given word. For example, if you take the notes found in the word “weld” [weᴗᴖld▪ ᴖᴗˏᴗᴖᴗˏ] (the transcription of words is discussed in the following section), [ᴗ] at the beginning and [ᴗ] at the end of the word do not have same context (neighboring) information. This difference brings about a variation in the acoustic parameters generate from the text. Indirectly, this has an impact on the song to be synthesized. So, one should ask how to synthesize a given note from different contexts of a given word.

In this work, the mechanism of synthesizing notes from different contexts is presented. This is made possible by best selection of parameters in a given context. A specific note can, therefore, not be expected to have the same utterance in different contexts. To address these issues the architecture of the system is modeled into analysis and synthesis phase. The analysis phase is divided into two: *text analysis* and *speech analysis* as shown in Figure 4.1. Once the analysis phase is over the synthesis follows. Each of these phases is discussed in the following sections.

4.3.1 Text Analysis

To achieve a more natural voice quality for the notes, which is taken from songs listed in Appendix D and others, the contexts of a word is taken into account. The position of the notes in different contexts usually leads to the variation in pitch. To handle this variation, we have to get information about the notes and the surrounding phonemes. The first step towards this is, to

perform the transcription of the given word. While handling the transcription, one phoneme contexts from the left and one phoneme contexts from the right of the target note are considered. Since as practically proved that the target note is mainly depends on the previous note rather than the note before previous one. In the other hand, collecting all phonemes with its note variation is getting more complicating for the reason stated above, that is to differentiate notes with its note marker. The solution made for this hard problem is to collect the Geez alphabet phones with a relatively constant pitch and intonation, see appendix C, as just for speech purpose.

Table 4.2 Sample transcription of notes

Word	Transcription	Previous	Target	Next
ሃ^ሁሊ /halie/	Ha B2 **A5** B2 P6 lie S2 P5 S6	B2	A5	B2
ሃ^ሁሊ^ሁ /halie/	Ha B2 A5 B2 P6 lie S2 **P5** S6	S2	P5	S6
ሉ^ሏያ /luya/	Lu **A1**ya	Lu	A1	Ya
ኀ^ቀለ^ሁ/hekle/	He S1 q le **A2**	le	A2	#
ነ^ሁበ^ረ /nebere/	Ne **A2** be S1 re S1	ne	A2	be

In Table 4.2 a few example of text analysis is shown. On the first two rows the words are the same and the target note to describe is also the same (*difat*) but the notes to be transcribed have different context and note markers. The # on the above Table denotes a null value, where we do not have any context parameter point. For instance, in the row of word “hekle” has two notes [S1] and [A2]. The target note [A2] does not have any succeeding parameter. So for such cases # is assigned for the particular context.

4.3.2 Song Analysis

There are two basic activities in this part: cutting off the required notes and phonemes from the song and extraction of the pitch and duration of the units. The song analysis process is majorly

conducted by using Praat software tool which is usually used for speech analysis. The segmentation process for the units is conducted manually by looking at the time wave and listening to the segmented unit by the help of knowledge expert. Even if hand segmentation is usually closer to the correct value, it is labor intensive

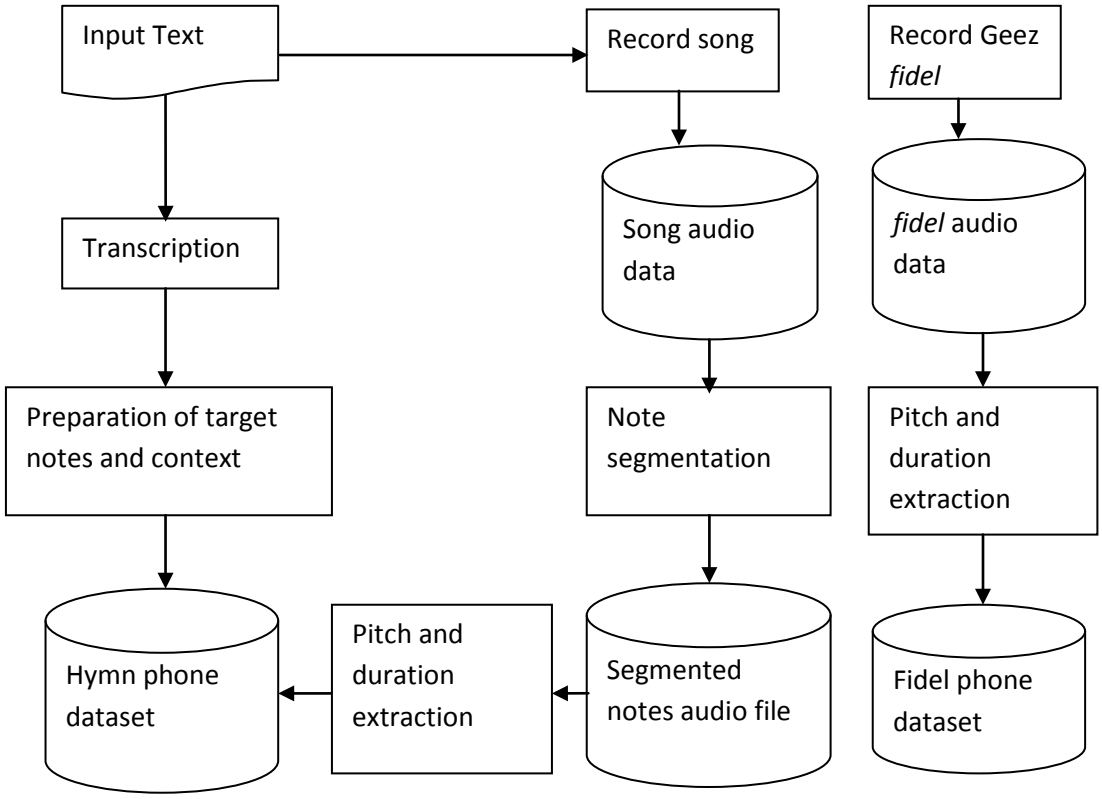


Figure 4.1 Analysis phase for text and song

4.3.3 Audio Dataset Preparation and Segmentation

For the purpose of this work three different songs are selected to incorporate the seven note markers. Around three hundred units that incorporate notes at different context are recorded. In addition to this the whole Geez letters are also recorded. The recordings are done by one person who has been an expert on teaching hymn. The advantage of recording one person is to alleviate the problem arises from different mismatch occurs because of the vocal contraction of different person gives different voice. In nature song is differ from speech by sustaining vowels for a long time [28], and we know that vowels are voiced signals so it needs special consideration. And that

is why we agree to record from a single male voice that had a tangible expert knowledge on the field of St Yare hymn.

The recording was done in area which had a less interruptions of external voice and electromagnetic waves. The parameters used to record the voices are stated as: sampling frequency of 44100 Hz, bit size of 16 per sample and saved in a Microsoft wave format.

The units are extracted by following step by step approaches. In the praat software there is a tool used to label the sound selected. First each word in a lyric is separately labeled as a word in a one labeling row for the specific song on the praat grid. On the second iteration the words are divided into unit labels on other layer of the grid. Here when we say unit it stands for notes or phonemes. The Figure 4.2 shows one of the labeling for the phrase "*Semaye bekewakibt*" (ሰማየ በከዋክብት) in praat environment.

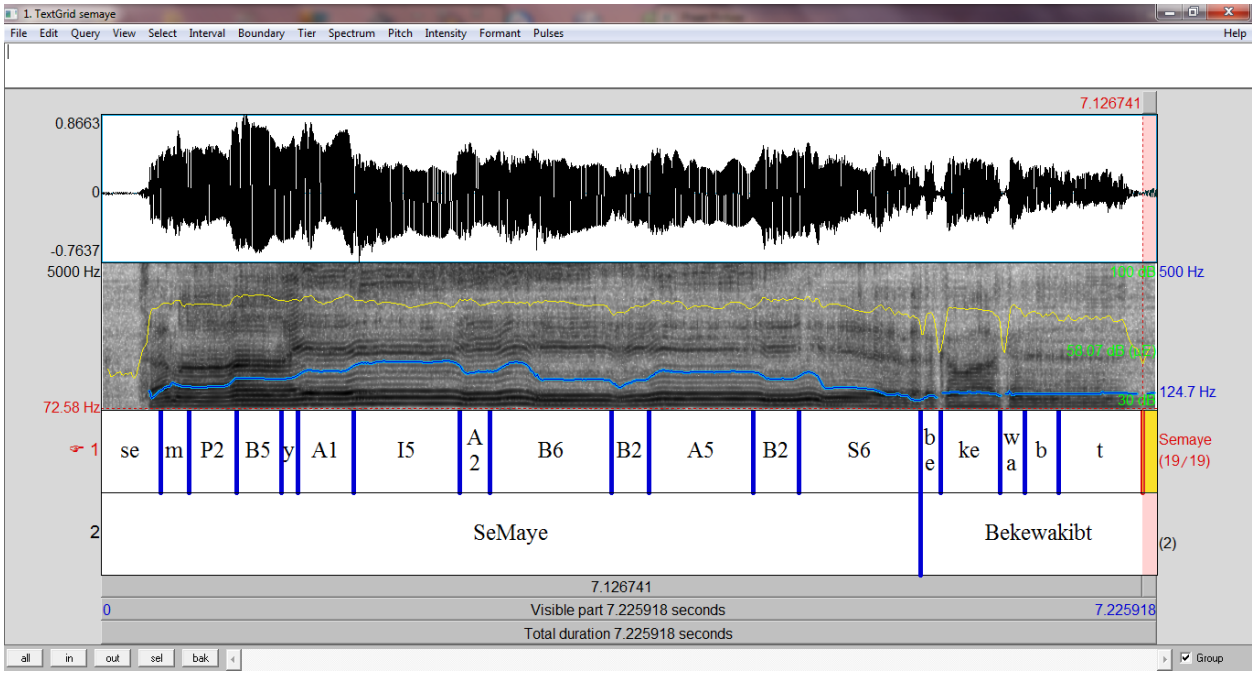


Figure 4.2 Unit labeling for the word "*semaye bekewakibt*" from the song *twiedso*

There is a possibility of having more than one note of similar even in their marker. But it will have different signal characteristics depending on the context they are found. For instance, in the word "*semaye*" stated above have a total of twelve notes of which four of them are a note of *deret* with a marker of *pier, aynu* and *bubay.* And the other three of them are under a note of

difat with different marker as *bubay, ile* and *aynu.* Two of them have also the same note *chiret* with the marker of *bubay and Selasa.* The rest note is a *yizet* note with the marker of *aynu.* Even though, the stated notes have similarity, their wave signals are stored in different file due to the different in context they occurs. Each of the unit waves are cutoff and saved in as separate file. For example the Figure 4.3 below shows the A2 note when it cut off from the "*semaye*" word time wave which has a context value of *Ile difat* from left and *Bubay chiret* from right.

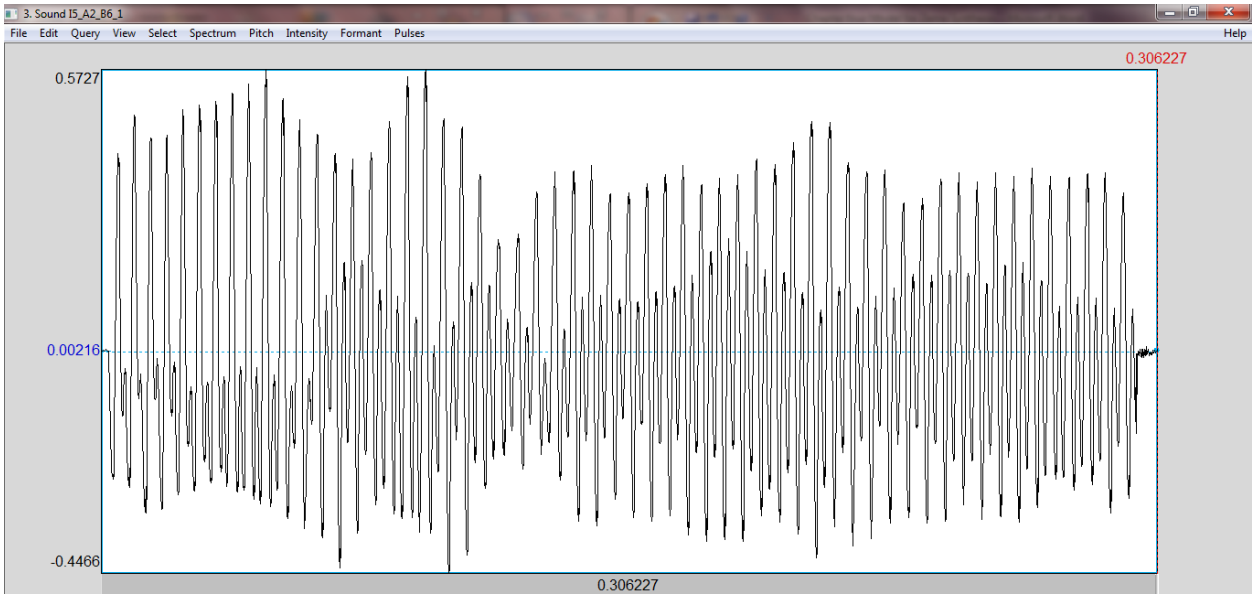


Figure 4.3 Sample wave format for A2 signal with a context of I5A2B6

4.3.4 Feature extraction

Features extracted from the unit signal waves in this work are the fundamental frequency the pitch and the duration of the unit signals. To get these parameters we use praat tools since it can show the duration of the time wave and also had a functionality to show pitch of the signal in a 10 ms interval. For the sake of this research the first pitch and the last pitch of the signal is taken manually from the pitch list and added to the transcribed dataset of the original signal in its corresponding list as shown in Table 4.2. A part of the hymn phone data set is shown on appendix B. That is for example if we take the unit A2 with the context of I5 in left and B6 at right we get a pitch list as shown in the Table 4.3 below. From that list we took the first pitch 204 Hz and the last pitch 187 Hz and store it in corresponding data list as Table 4.4 shown below.

Table 0-1 Pitch list for I5A2B6 unit

Time in s	Frequency in Hz
2.452959	*204.027802*
2.462959	*197.830776*
2.472959	*190.390999*
2.482959	*186.24098*
2.492959	*184.452316*
2.502959	*182.012553*
2.512959	*181.645853*
2.522959	*183.800239*
2.532959	*183.097083*
⋮	⋮
2.652959	*187.543531*

Table 0-2 Sample format for phone data list with duration and pitch information

Previous	Target	Next	Duration (microseconds)	First Pitch	Last pitch
Ha	B2	B2	710	154	139
I5	**A2**	**B6**	**306259**	**204**	**187**
L	P5	S6	764286	109	138

There is a case where a given phoneme is not found in the phone data list since we didn't record a song which has all combinations of notes with the whole set of Geez character sets. For the reason described in the section 4.3.1 above we set up a set of audio files for Geez alphabets. This was done by recording each letter with the voice of the same person and prepare for both audio file and phone data set. The phone data set has the following style as shown in the Table 4.5.

Once the text and song analysis are over and all the required parameters are stored in their appropriate datasets, the next step is to synthesize the notes by using the available information in the storage. The detail of the synthesis is discussed in the following section.

Table 4.5 Sample format for phone data list of Geez alphabet

phone	Duration (micro Sec)	First pitch (Hz)	Last pitch (Hz)
ba	112971	120	120
be	197687	129	132
bi	13560	119	124
bie	190635	128	130
bo	169455	110	110
b	183583	127	133
bu	134150	133	124

4.4 Hymn Synthesis Phase

The synthesis phase has three major parts. These are transcription of the input text, unit selection and wave concatenation and synthesis. Each of the phases is discussed in the following section according to the hymn synthesis with a brief description of examples.

The first step in synthesis is to know the context of the note in the input text to be synthesized. For this we have to know the transcription of the text input so that we can easily identify the notes with their context. Then the next step will be to select the candidate units from the unit inventory of the phone data set. The candidate units are further processed to select the best match unit. The units are used to find the audio files from the audio dataset. This is in turn passed to a concatenation followed by wave synthesis. The Synthesis model architecture is shown in the Figure 4.4.

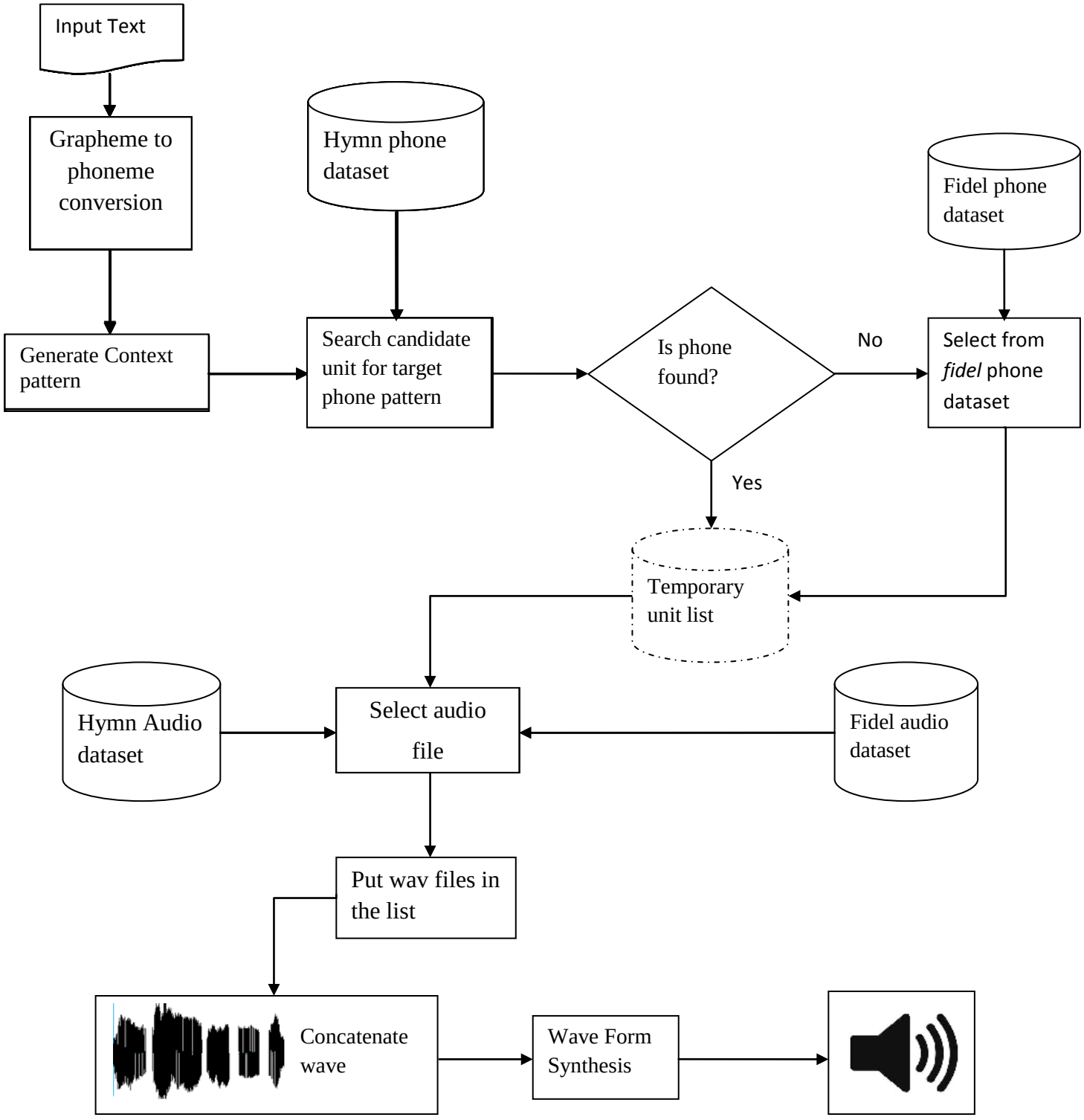


Figure 4.4 Synthesis model architecture of St Yared hymn Notation

4.4.1 Transcription of the Input

In the text analysis part of the synthesis input text is taken from the user. While inserting the text the user follows some steps as stated forward. Every note is written as note –marker combination and phonemes are all written in small letters because upper cases are used as note marker identifier in this model. For example the word ወ ልዩ is inserted as weP2B5ldA1I5A2. This preprocessing would be done manually which may take time but it brings a better choice than trying to transcribe automatically.

The next step after input text transcription is grapheme to phoneme conversion. To do this conversion the system put each character of the word in a list and parses for getting the notes following of the target character. The system takes certain parameters about the letters. The parser takes each character in a word and checks for their neighbor character. If the next character is consonant the parser keeps the character and inserts a T (silent note) between the two characters in the list to describe the phoneme has no sound. This silent symbol is often used in *akuakuam*, where the chants of Yared hymn accompanied by special dance. For example for the letter list "weP2B5ld" above the result is as shown below

- **Input** weP2B5ldA1I5A2
- **Output** weP2B5l፡ dA1I5A2

The letter l is followed by ፡ (silent symbol) according to the rule mentioned above and the list is now converted to one long string for the purpose of the next step. In hymn to produce a voice for a single phoneme it must has a corresponding note except for the case of silent consonants which we described as above. One phoneme had at least one note but in most of the cases it has more than one note for a single phoneme which is called portamento [28]. For the sake of simplifying the parsing for the next step in the model each element in a long string are separated by "_". So the above example result becomes *we_P2_B5_l፡_d_A1_I5_A2*.

After all here is the place where phonetic analysis phase of the model is come to an end. At this point one long string for the input text is formed by passing through the aforementioned steps, which describes the phonetic conversion of the given text word. The next step is unit selection which is used as a bridge between two broad divisions of text to speech synthesis these are NLP and DSP.

4.4.2 Unit Selection

In the model shown above there is a place where for each target unit in a processed list the system searches its equivalent unit phone from training phone dataset. To do that there is a few steps to follow. First selection of candidate phones from phone dataset is done and put them in a list until the training phone units are fully parsed. To do this, the system made certain preprocessing on the input phone string. As the expert on the Hymn says the phone of a single note depends mainly on the previous note and doesn't go further to the phone of previous note. For this reason we made the model to prepare a pattern for the input phone string which contains three elements. Each element in the pattern has their type of expression to handle the context as, "*Before, Target* and *after*". Target is the element under question for selecting candidate unit. "*Before"* is a left element just before the target and "*After"* is the right element just after the target element. During pattern generation there is a case where "*Before" element* becomes empty when the target is the first element and also *after* element would be empty when the target is the last element in the string. For the sake of the manipulation of the code the empty points are characterized as "#". During this search for candidate element the characters with silent note are not targeted for parsing but it is used as context phoneme. For the above example, the list of patterns for String "we_P2_B5_l_d_A1_I5_A2" becomes as follows: #_We_P2, we_P2_B5, P2_B5_l, L_D_A1, d_A1_I5, A1_I5_A2 and I5_A2_#.

After preprocessing of the input string is done well then each pattern is sent for searching for their equivalent unit in the training dataset which was previously done during text analysis. This module of the model works as in the following algorithm. The algorithm shown in Algorithm 1 can be described as: first it parses all the training dataset and if it gets contextually equivalent units it compares it with the input pattern. Comparison of the input pattern is done in three different ways these are

1. Full matched search: it looks for exact copy of the input pattern if it gets it puts it in a full matched temporary list and continues until totally parse the entire training dataset.
2. Partial matched search: it looks for data that contain the target element with either left or right context element and continues until totally parse the entire training dataset.
3. Target only search: it looks for the target element only.

1. *Take input pattern*
2. *Parse all training phone dataset*
3. *If no matching unit is found go to 11*
4. *Compare the searched unit with input pattern*
5. *If fully matched with input put in fully matched temporary list*
6. *If partially matched put in partially matched temporary list*
7. *If target only found put in a target only matched temporary list*
8. *Select the priority temporary list*
9. *Compare the candidates similarity distance using previous selected unit context(pitch)*
10. *Select a candidate with small distance with the target unit and put it in searched phone list*
11. *Parse fidel phone dataset and select the target phone unit and put it in searched phone list.*

Algorithm 1 Unit selection algorithm

The priority of the candidate selection is given to fully matched result and then for partial matched and finally for the target only matched results. The second step is to select the most appropriate phone from candidate list by comparing its context behavior. Context is as defined in the above sections, the target note or phoneme plus the left and right strings in a list. The model selects the units to smoothly concatenate the sounds based on the parameters of previous selected unit and target. We use the pitch parameters of both the target and immediate unit before target phone. Since we have pitch information for each unit during sample preparation, which includes the initial and last pitch of the unit. For the comparison of candidate units the absolute pitch difference which was calculated as the square of the difference between candidate pitch and a unit selected before the target pitch is used. And the minimum the difference of pitch used as selecting criteria .For example list 4.2 shows the result for input pattern A5_B2_B2 where 151 is the pitch of the A5 left context of the pattern, in our case *aynu difat note*, and the target B2 b*ubay deret*

Table 4.6 Sample example of unit selection search result

Level of matching	Squared difference	candidate units
full matched candidate	*289*	*A5_B2_B2_1_168*
full matched candidate	*289*	*A5_B2_B2_2_168*
full matched candidate	***256***	***A5_B2_B2_3_167***
full matched candidate	*1156*	*A5_B2_B2_5_185*
Candidate for target only	*36*	*A5_S1_P6_1_145*
partial candidate	*196*	*A5_B1_B2_1_165*
partial candidate	*441*	*A5_B1_B2_2_172*
partial candidate	*400*	*A5_B1_B2_3_171*
partial candidate	*625*	*A5_B1_B2_4_176*
Candidate for target only	*36*	*A5_P1_P6_1_145*
Candidate for target only	*16*	*A5_P1_P6_2_155*
full matched candidate	*441*	*A5_B2_B2_1_130*
Candidate for target only	*25*	*A5_P1_P6_1_146*
full matched candidate	*576*	*A5_B2_B2_4_175*

Since the phone data “A5_B2_B2_3_167” has 256 which is the least when compared to the other options in the phone data list it was selected as appropriate unit. Each selected results are put in a temporary unit list until the whole element in the long string is fully parsed to get their own unit result from training data set.

The third step in selection is to find audio file from audio data set upon each unit in temporary unit list. The result audio file is kept in audio file list in the order of units in temporary file. After all audio files are collected it passes to the next step which is concatenation of audio files. This is done by concatenating the files in an audio format to produce a single audio file. Finally this audio file is going to generate the synthesized output of the given input text. But just before that due to the perfection of the units selected the synthesized sound may not good for us to hear, it may contain many clicks of sounds. To minimize the effect off unwanted sounds produced due to the unit selection, the synthesis part has to go for further processing. The next phase which is going to take the responsibility to the aforementioned problem is the digital signal processing (DSP) phase of the model which was indicated as wave synthesis in the architecture of the synthesis model. It is discussed in the coming section.

4.5 Wave synthesis

In concatenation of the audio files, all audio files which are selected for the given input word for synthesizing is put together in their order of occurrences. Since the audios are collected from different parts of songs, it doesn't match with the neighboring phone even though special attention is considered to minimize the effect of coarticulation during unit selection. The problem occurs for varies reasons especially the amplitude, pitch and duration of the phone are among very important attributes for the synthesized sound. So in order to alleviate such problem different tasks has to be taken. This task includes pitch shifting and amplitude modification.

4.5.1 Pitch Shifting

Each note has different frequency when they appear on different phonemes. In other words the notes are assigned to different pitch according to the hymn. This produce unnecessary sound effect when different units from different word are concatenated to produce new word. Smoothing of the transition from relatively low frequency to higher or from higher to low frequency at a spot of time is difficult. To overcome these problem different types of techniques must be employed. One of these techniques is Pitch shifting.

Implementation of pitch shifting techniques is different depending on the persons will. One may work on the starting pitch and others may on the ending and others may work on both. Even though their approach of shifting varies, the method is similar. That is for instance if the current phoneme ending pitch is 120 Hz and the next phoneme starting pitch is 180Hz, this can be done by taking from 90% time of the selected unit wave length up to the end of unit's wave, the frequency is shifted up linearly to reach 180Hz. The complexity of shifting depends on the way we choose to accomplish. By shifting at both ends of the phoneme, at starting and ending, increases the process time of the synthesis while keeping the result getting better smoothening than the single point of shifting.

The technique proposed has two steps. The first step is to shift the pitch according to the neighboring context. Then space created between the slice are removed. To accomplish this task the system first evaluates the cost of each unit signal. The cost is a value given for each unit after unit selection. The cost value given for the units are three integers as follows: 1 for full match unit, 2 for partial match unit and 3 for target only unit selected. For the sake of this thesis signals

with two and three cost values are selected for analysis. In the following section how pitch shifting algorithm works is described mainly adopted from [28].

4.5.2 Pitch Shift Algorithm

The algorithm mainly lies on the principle of double the length without affecting the pitch and then plays it twice as fast. All frequencies would be doubled and therefore the pitch would be shifted and the duration would match the initial one. So now we first have to change the signal duration without changing the pitch. The scaling factor is defined as the factor used to stretch or compress the spectrum in order to adjust the frequencies such that the pitch is shifted. Once this is done, we will perform re-sampling to return to the initial duration with a shift in the pitch.

For instance, if we want to shift the pitch up by one semitone, we need a scaling factor of $2^{(1/12)}$ which equals to 1.0594. This means we first need to stretch the signal without changing the pitch such that the duration is now multiplied by 1.0594. Once this is done, we need to play the signal 1.0594 times faster as shown in the Figure 4.5.

On the other hand, if we want to shift the pitch down by one semitone, we need a scaling factor of $2^{(-1/12)}$ which equals to 0.9439. This means we first need to compress the signal without changing the pitch such that the duration is now multiplied by 0.9439. Once this is done, we need to play the signal slower at 0.9439 times the initial speed.

Superposition of frames

We will split our signals in many small frames. These frames are taken from the initial signal such that they overlap each other by 75%. These frames will then be spaced differently in order to stretch or compress in time the output signal. This is shown in Figure 4.6.

But now we have another problem. When we space the frames differently, this creates discontinuities. This is shown in Figure 4.7 (space between frames is initially of length x and becomes of length y).

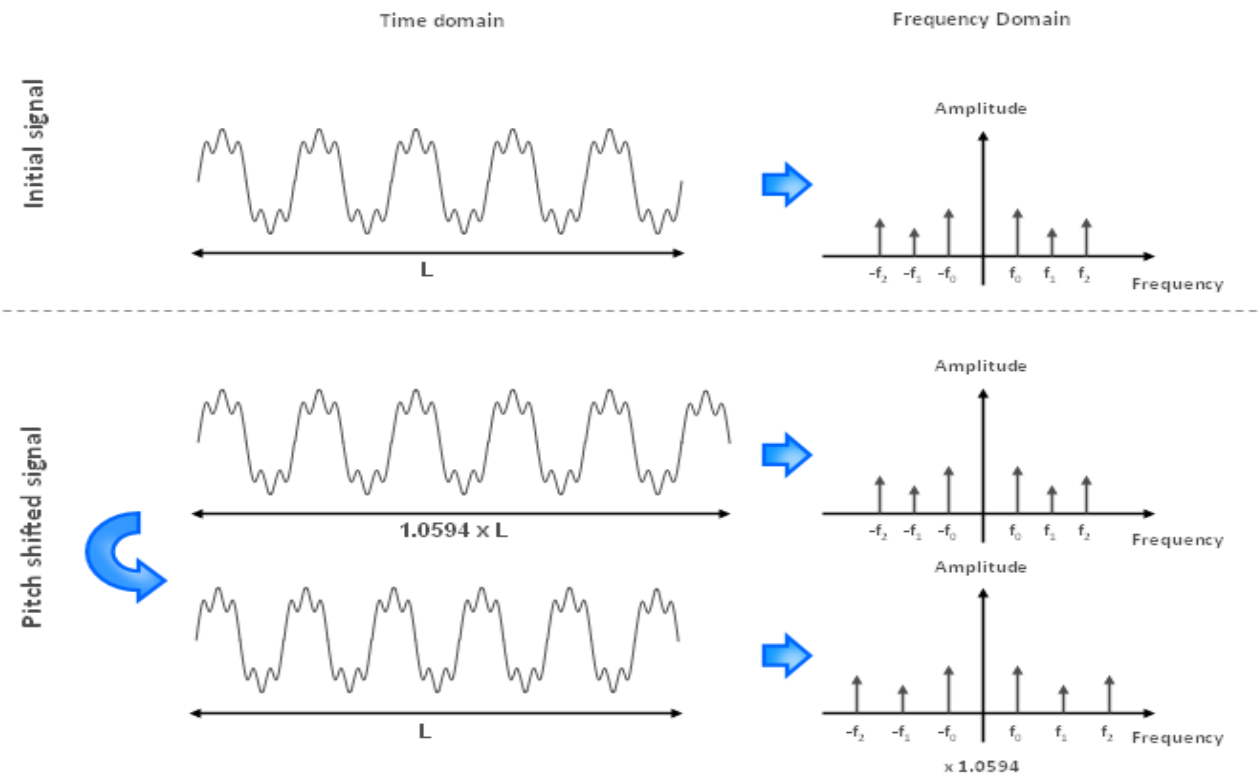


Figure 4.5 How to shift the pitch up by one semitone

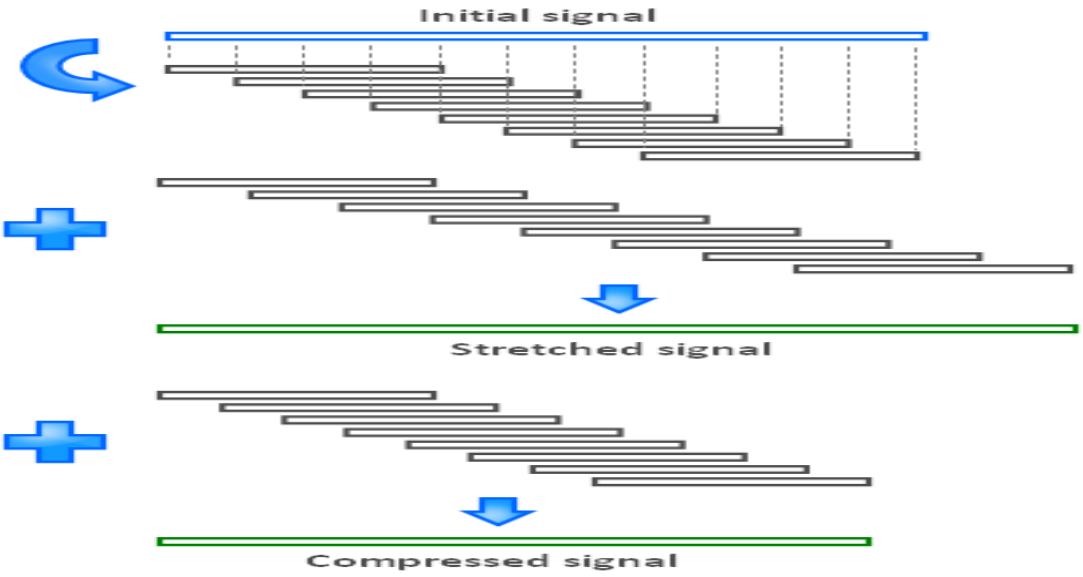


Figure 4.6 Stretching and compressing the signal frame by frame

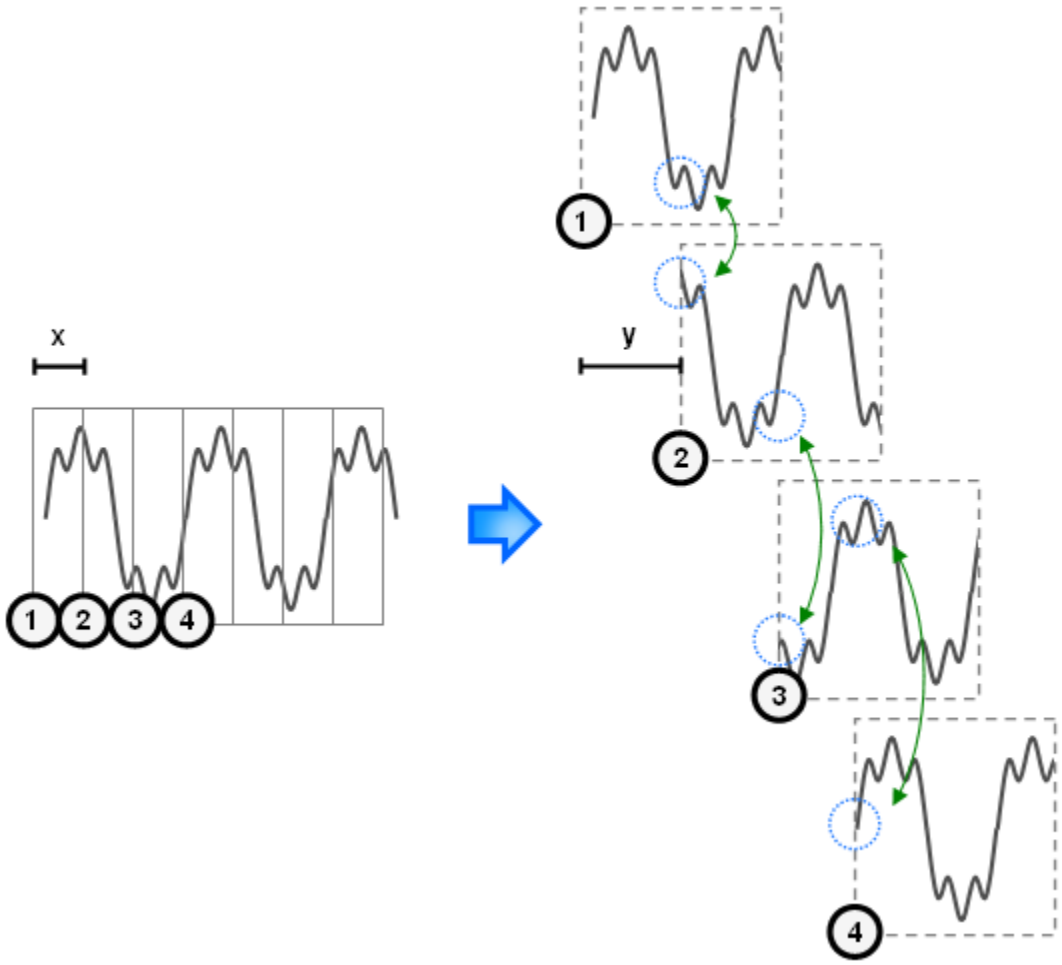


Figure 4.7 Signal discontinuities when stretching or compressing in time

This will produce glitches and will be perceived by the human ear. We need to do something to make sure there is no discontinuity. That's why we need a phase vocoder.

4.5.3 Phase Vocoder

Figure 4.8 shows how the phase vocoder algorithm works. It consists of 3 stages: analysis, processing and synthesis. We are now really needs to go through some equations to understand what the algorithm does.

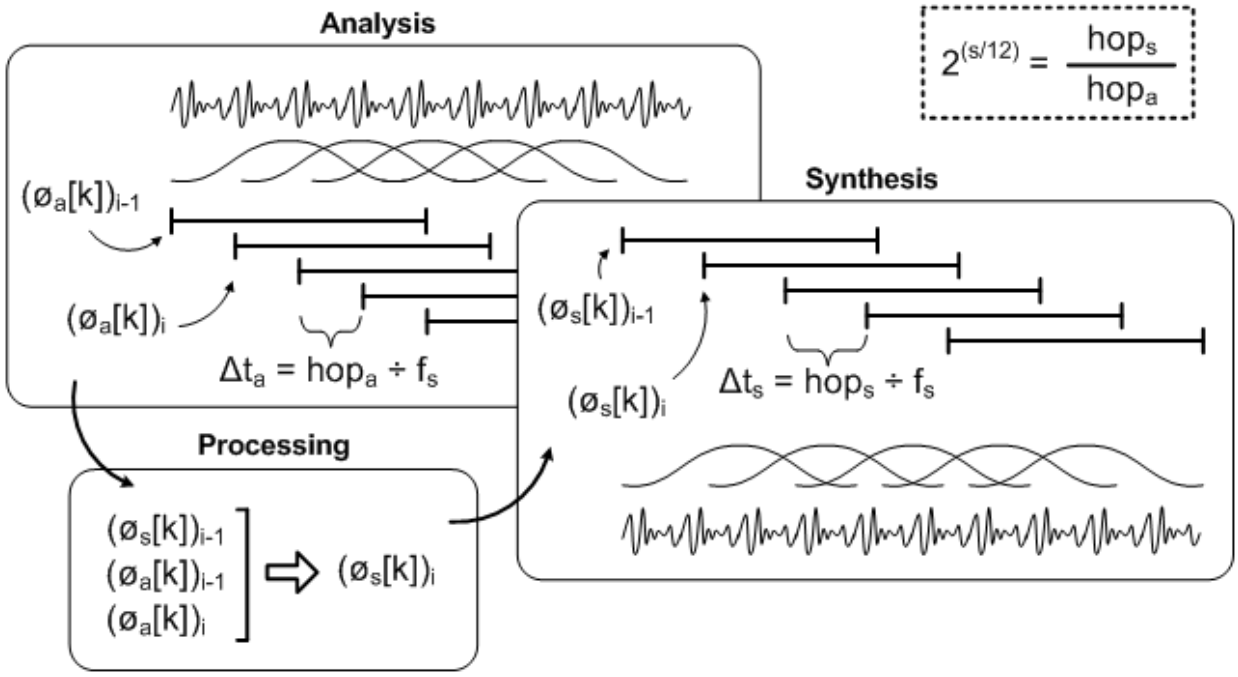


Figure 4.8 Phase vocoder overview

Analysis

Windowing refers to taking a small frame out of a large signal (you are only looking at your signal through a small window of time). The windowing process alters the spectrum of the signal and this effect should be minimized. In order to reduce the effect of windowing on the signal frequency representation, a Hanning window of size N is used.

The frame made of N samples is then transformed with a Fast Fourier Transform (FFT) as shown in equation 4.1.

$$(X_a[k])_i = \sum_{n=0}^{N-1} x[n + i \times (hop_a)]w[n]e^{-j(\frac{2\pi n}{N})} \quad \text{k= 0,1,2,...,N-1} \qquad (1)$$

In the previous equation, $x[n]$ is the sample signal, $w[n]$ represents the Hanning window and $(X_a[k])_i$ represents the discrete spectrum of the frame i. To increase the resolution of the spectrum, the windows are overlapped with a factor of 75%. The number of samples between two successive windows is referred to as the hop size (hop_a) and is equal to N/4 for an overlap of 75%.

Processing

Applying a FFT of length N results into N frequency bins starting from 0 up to *((N-1)/N)*fs* with an interval of *fs/N* where *fs* is the sampling rate frequency. A signal with a frequency that falls between two bins will be smeared and its energy will be spread over the nearby bins. The phase information is used to improve the accuracy of the frequency estimation of each bin. Figure 4.9 shows two sine waves with slightly different frequencies. These frames are divided into frames of N samples. Windows do not overlap in this case to make the explanation simpler. The first sine wave has a frequency of fs/*N* and thus falls exactly on the first bin frequency. The second wave has a frequency slightly greater than the first bin frequency. It is not centered at the first bin although its energy lies mostly in this bin.

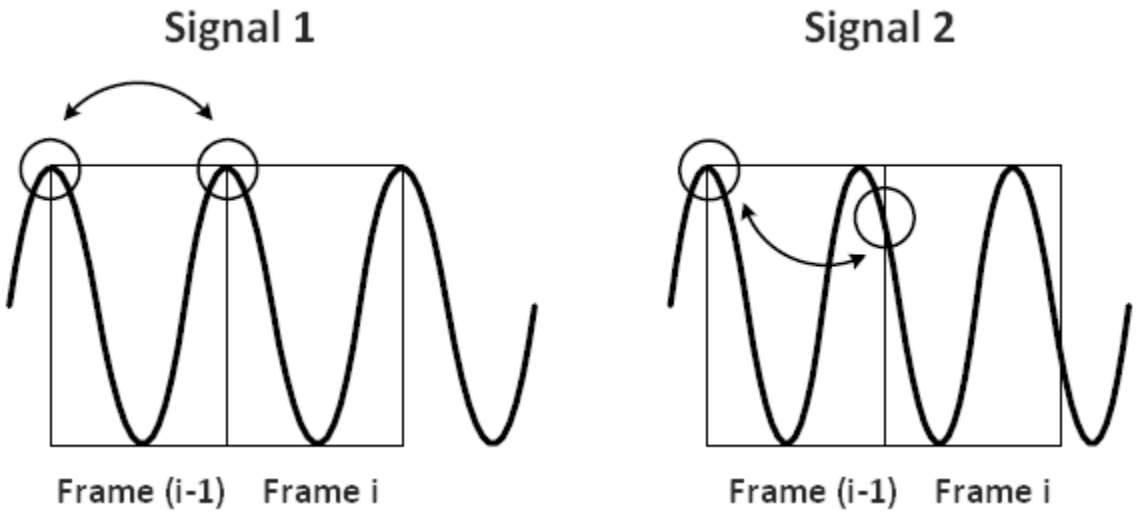


Figure 4.9 Sine waves with different frequencies and phase difference

The phase information of the two successive frames is of significant interest. For the first signal, there is no phase difference between the two frames. However, for the second signal, the phase of the first bin is greater than zero. This implies that the signal component corresponding to this bin is higher than the bin frequency. The phase difference between the two frames is referred to as the phase shift$(\emptyset_a[k])_i$. For all variables in this section, *k* and *i* stand for bin index and frame index respectively. The phase shift can be used to determine the true frequency associated with a bin. However, the phase information provided by the FFT is wrapped, which implies that $(\Delta\emptyset_a[k])i$ lies between –pi and pi. Without wrapping, the true frequency $(w_{true}[k])i$ can be

easily obtained from the phase shift $(\Delta\emptyset_a[k])i$and the time interval Δt_a between two frames as shown in Equation2. The time interval Δt_a is the hop sizehop_a divided by the sampling rate frequencyf_s.

$$(w_{true}\,[k])_i = \frac{(\Delta\emptyset_a[k])i}{\Delta t_a} \tag{2}$$

The matter is more complex since the phases are wrapped. In this case, the frequency deviation from the bin is first calculated and is then wrapped. This amount is added to the bin frequency to obtain the true frequency of the component within the frame. Equations 3, 4 and 5 illustrate this procedure. The variables $(\emptyset a[k])_i$ and $(\emptyset a[k])_{i-1}$ stand for the phase of the previous frame and the current frame respectively. Also, $w_{bin}\,[k]$ stands for the bin frequency,$(\Delta w[k])_i$ stands for the frequency deviation and $w_{wrapped}\ [k])_i$ stands for the wrapped frequency deviation.

$$(\Delta w[k])_i = \frac{(\emptyset a[k])_i - (\emptyset a[k])_{i-1}}{\Delta t_a} - w_{bin}\,[k] \tag{3}$$

$$(w_{wrapped}\ [k])_i = mod[((\Delta w[k])_i + \pi), 2\pi] - \pi \tag{4}$$

$$(w_{true}\,[k])_i = w_{bin}\,[k] + (w_{wrapped}\ [k])_i \tag{5}$$

The new phase of each bin can then be calculated by adding the phase shift required to avoid discontinuities. This is done by multiplying the true frequency with the time interval of the synthesis stage as shown in Equation6

$$(\emptyset_s[k])_i = (\emptyset_s[k])_{i-1} + \Delta t_s \times (w_{true}\,[k])_i \tag{6}$$

The phase of the previous frame for synthesis $(\emptyset_{i-1})_s$ is known since it was already calculated as the algorithm is recursive. The new spectrum is then obtained as shown in Equation7.

$$|(X_s[k])_i| = |(X_a[k])_i| < (X_s[k])_i = (\emptyset_i)_s \tag{7}$$

Synthesis

So now we managed to adjust the phase in the frequency domain for our current frame. We need to come back to the time domain. The inverse discrete Fourier transform (IDFT) is performed on

each frame spectrum. The result is then windowed with a Hanning window to obtain$q_i[n]$. Windowing is used this time to smooth the signal. This process is shown in Equation8.

$$q_i[n] = \left\{\frac{1}{N}\sum_{k=0}^{N-1}(X_s[k])_i\, e^{-j\left(\frac{2\pi kn}{N}\right)}\right\} w[n] n = 0,1,2,\dots,N-1 \quad (8)$$

Each frame is then overlap-added as shown in Equation9. The variable *L* stands for the number of frame and $u[n]$represents the unit step function.

$$y[n] = \sum_{i=0}^{L-1} q_i\,[n - i \times hop_s]\{u[n - i \times hop_s] - u[n - i \times hop_s - N]\}] \quad (9)$$

Re sampling

We have our signal that is now either stretched or compressed in time and the pitch is not changed. Now we need to resample our signal to get back to the initial duration and shift the pitch. Suppose we have a given sampling rate and we want to double the frequencies. The easiest thing to do is to pick only one sample out of two and to output the result.

Now this is easy because when we double the frequency we deal with a scaling factor that is an integer. If the scaling factor is not an integer (for instance 1.0594 when we shift the pitch up by one semitone), then we cannot use the same technique. We will use instead linear interpolation to approximate the sample that should lie at this location. Linear interpolating also behaves like a low-pass filter and thus removes most of the possible aliasing when we down sample. Figure 4.10 shows an example with a scaling factor of 1.2.

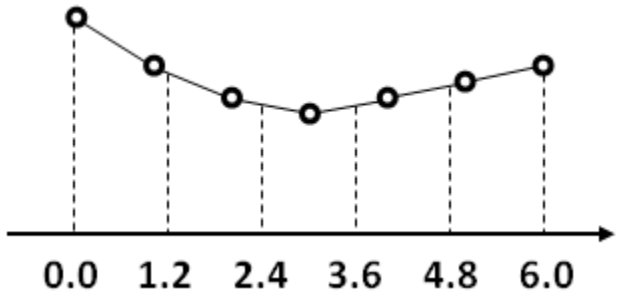


Figure 4.10 Re sampling using linear interpolation

4.5.4 Contextual Pitch Shifting

The pitch shifting discussed in the above section is for shifting the whole part of a song. This is like to act using men's voice as female by shifting with certain tone. In this section the actual pitch shifting used in this thesis work is described. For the purpose of acting a shift on the selected unit wave the pitch of the signals at starting position is taken and compared to the previous unit in the context. If the ratio of current signal pitches to the previous pitch is less than one some parts of the previous unit signal frames are down shifted. On the other side if the ratio is greater than one some parts of the previous signal is shifted up. For example Figure 4.11 and Figure 4.12 shows this phenomena for a pitch shifted up scenario of note **A2R1**. The blue line indicates the pitch contour and the oval marks shows the place where pitch shift happens.

The following equation describes the methodology used for pitch shifting in this thesis work.

$$S=\sum_{i=1}^{n} modPitch\big(si, cost(i)\big) \tag{10}$$

For each unit signal in the series of concatenation as shown in equation 10 the modified pitch is calculated with respect to their cost index. And the final signal is the sum of modified pitch signals of individual signals. Equation 11 shows how a modified pitch is evaluated with respect to signal positions at three levels; these are at introduction of unit (start) at the middle and at the end. Here the step of shifting is calculated as equation 3 and 4 for both steps at the end and start of the unit signal under evaluation respectively. The middle part of the unit signal is not shifted so that the step value for it is assigned to one.

$$modPitch(\text{s,cost})=\sum pitchShift(start, mid, end) \tag{11}$$

where
$$\begin{cases} stepEnd(i) = startPitch(i+1)/endPitch(i) \\ stepStart(i) = startPitch(i)/endPitch(i-1) \\ stepMid(i) = 1 \end{cases}$$

equation 12 shows the actual pitch shifting where a linear interpolation of the signal is calculated using the amount of pitch shifting which we call it here the step.

$$pitchShift(s) = \sum_{f=1}^{numFrame} \sum_{i=1}^{frameLen} interp1(s, step) \quad (12)$$

4.5.4 Amplitude Modification

Hearing a sound that has different intensity in a short time in a single word is not comfortable for our ear. But in the case of a simple concatenation of units from different words appear with different amplitude; this is also one of the challenges in unit selection concatenation. To overcome this problem it is suggested to smooth the amplitude of the new synthesized word.

Amplitude modification by fading in and fading out of the signal is necessary at the appropriate points. This can be achieved by amplifying the wave signal of the unit under consideration by certain amount of multiplying factor k. For a given unit signal wave first the amplitude of the signal which is peak is identified by a matlab function as max (signal). In equation 13 the procedures for the intensity modification is shown.

R=max(Sc)/max(Sp); Sc=R*Sc

(13)

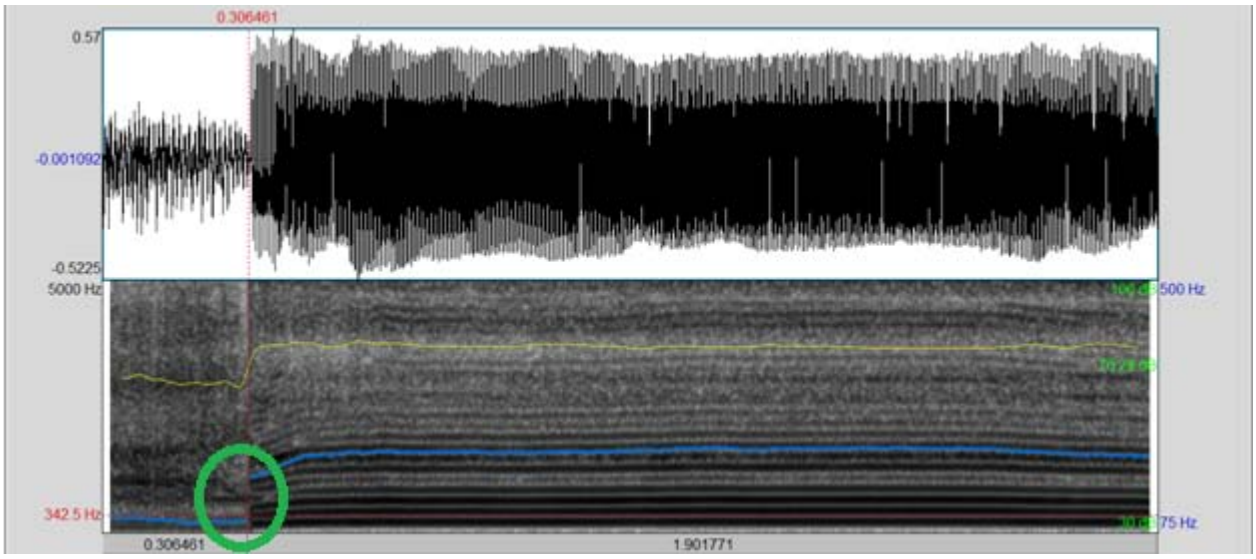


Figure 4.11 Wave signal for “A2R1” before pitch shift

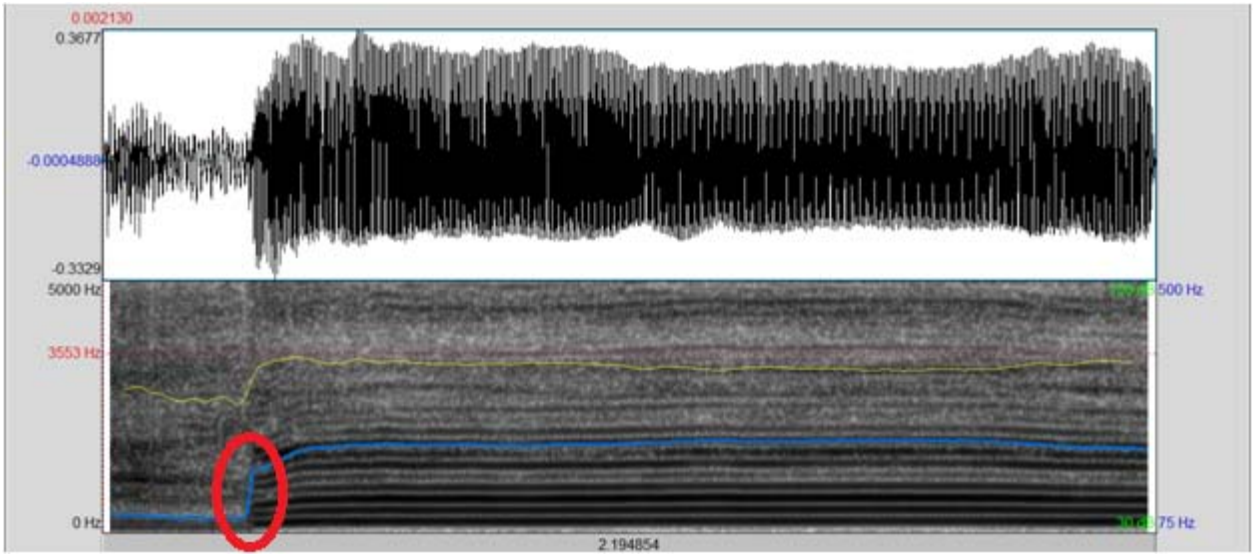


Figure 4.12 Wave signals for “A2R1” after pitch shift

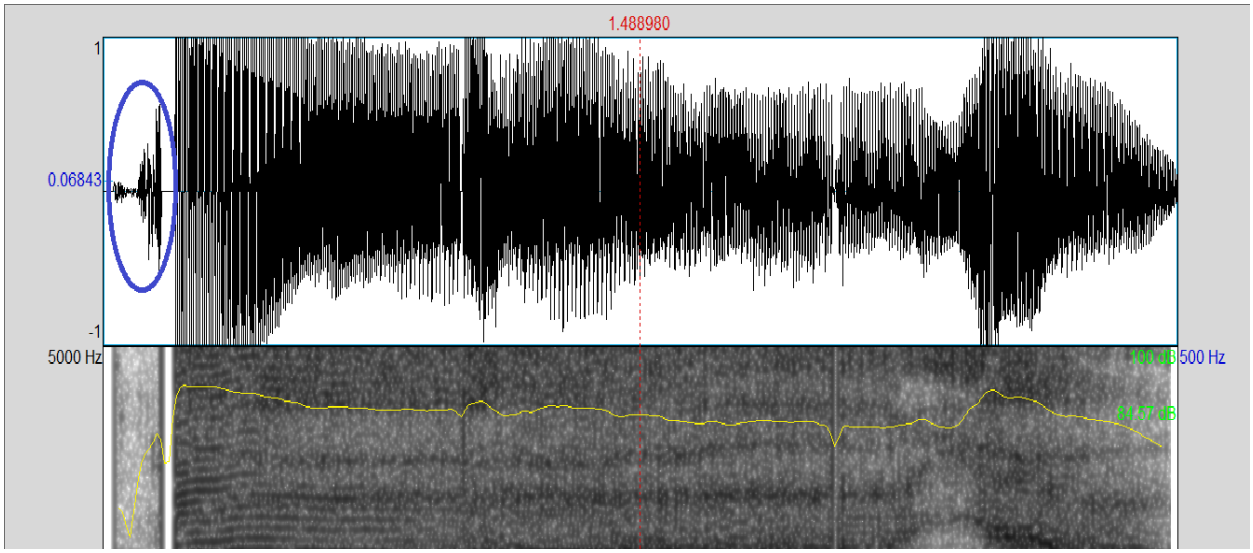


Figure 4.13 Portion of wave signal before amplified indicated by circle

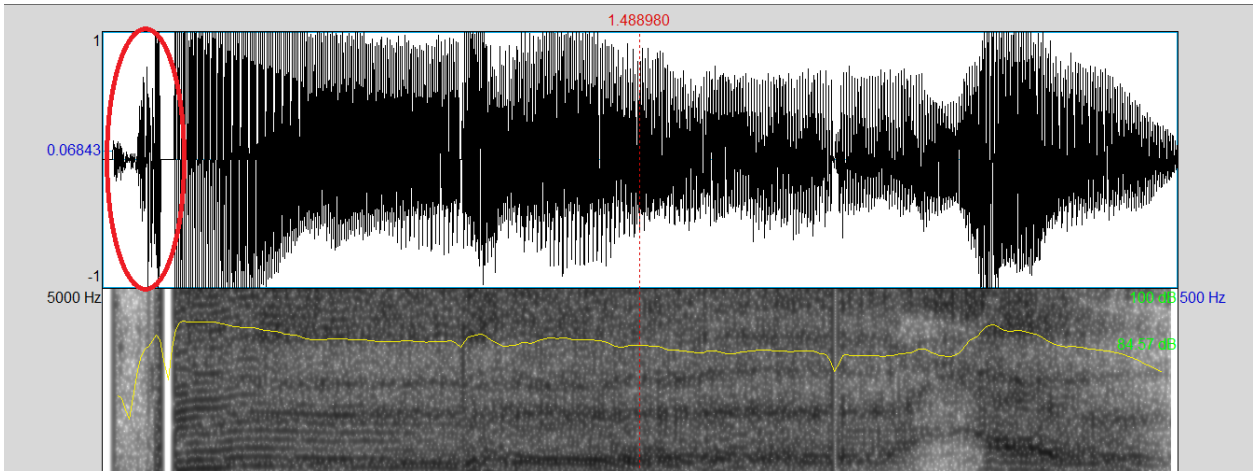


Figure 4.14 Portion of a wave signal after amplified indicated by circle

Where R is the ratio of current signal (Sc) to the previous signal (Sp)

To illustrate the above equation we can see an example signal for ***koS1A6yeI1.*** Figure 4.13 shows a signal with low amplitude for ***ko*** with a circled indication over it. Amplitude of signal for **ko** is compared to amplitude of **S1** and this ratio is multiplied with the original signal unit of **ko** to give the amplified signal as shown on the Figure 4.14.

4.6 The Prototype

To illustrate the model described in the above sections a prototype has been developed using java jdk 1.7 and matlab R2012a. The screen shot shown in Figure 4.15 is a front end processing of the user input.

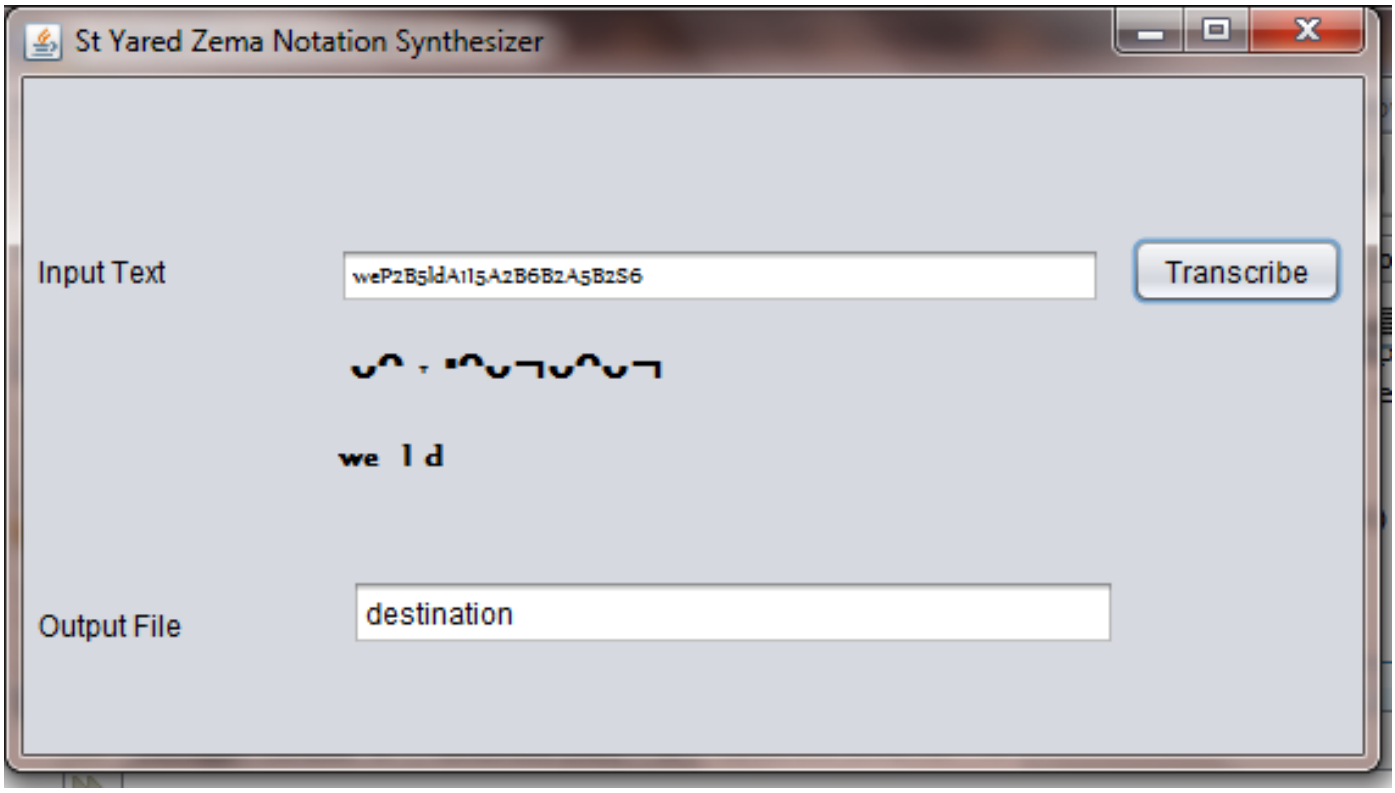


Figure 4.15 screen shot for transcription of input text

The developed prototype begins by transcribing the input text when the user press the transcribe button. The unit selected phone files are externally saved in the file specified at the output File text field. And the equivalent representation of the input text is shown on the interface as shown on the figure.

On the Figure 4.16 the output signal processed by starting with collecting data stored on the unit list file specified on the text. When the PITCHSHIFT button is pressed the system evaluate each neighboring wave files and take action on the pitch shifting according to the step needed. When the CONCTENATE button clicked the pitch shifted wave are evaluated for intensity and appropriate action depending on the amplitude difference between adjacent unit wave signals.

Finally when the Sing button is clicked the concatenated wave plotted on the next panel and the output song is voiced.

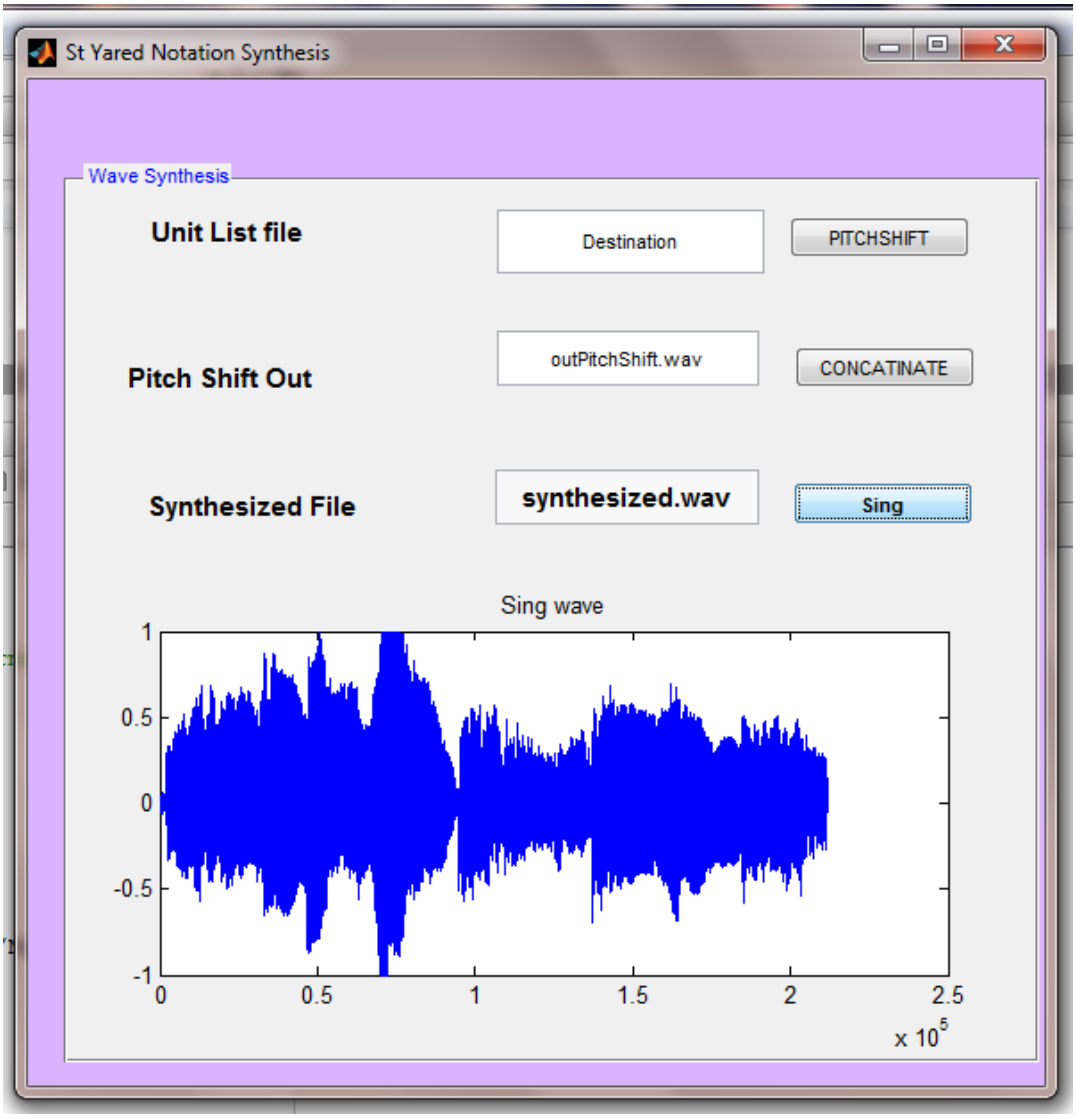


Figure 4.16 screen shot for wave synthesis

Figure 4.17 show the prototype window after aggregating the natural language processor with the digital signal processor functionalities in standalone application using MATLAB builder JA. This interface window enables us to run the system by clicking a single button SING for synthesizing the input text.

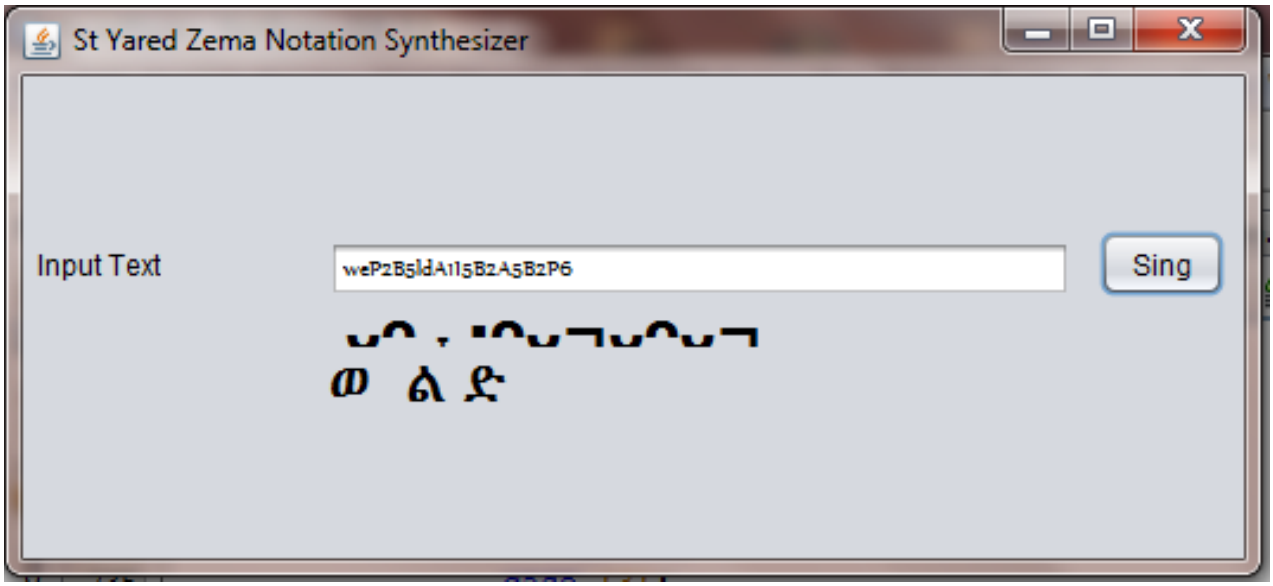


Figure 4.17 Screen shot for combined interface for Hymn Synthesizer

Chapter 5

Experimental Analysis

5.1 Introduction

Synthetic speech can be compared and evaluated with respect to intelligibility, naturalness, and suitability for used application. In some application, for example reading machines for the blind, the speech intelligibility with high speech rate usually more important feature than the naturalness. On the other hand, prosodic features and naturalness are essential when we are dealing with multimedia applications or electronic mail readers. The evaluation can also made at several levels, such as phoneme, word or sentence level, depending what kind of information is needed.

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 as input for the system are processed by the system resulting synthesized speech excerpts are collected, and subjective judgment tests are performed by human listeners for synthesized 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.).

When repeating the test procedure to the same listening group, the test results may Increase significantly by the learning effect which means that the listeners get familiar with the synthetic speech they hear and they understand it better after every listening session [27]. Concentration problems, on the other hand, may decrease the results especially in segmental methods. Therefore, the decision of using naive or pro listeners in listening tests is important.

In this research work a Mean Opinion Score (MOS) which is among the widely used technique for speech evaluation by perception of Human subjects is used. MOS is an evaluation technique where evaluators indicate their assessments on a scale of bad (1) to excellent (5). Then the average of the opinion will be taken as the performance of the system.

5.2 Experiment setup

A sample dataset were collected from different scripts which have St Yared hymn notations. On the first set up a simple concatenation was formed by using a matlab environment. For this test a note is selected without considering the note marker combination. For the phoneme we used a Geez alphabet sounds which was recorded during modeling the system. For the second scenario the words collected are inserted to prototype developed. All notes in the word are marked for its correct combination before transcribing the input text. The transcribed word is further processed to produce a set of unit files. The unit files of sounds are concatenated to generate a synthesized wave signal.

For the MOS evaluation eight users, who were naïve for synthetic sound and experts of St Yared hymn are selected. They listen to a synthesized sound for 10 sample words. The synthesized sounds ware grouped as marked and unmarked notes. On the first step they listen for unmarked note sounds and then for marked notes. The users are invited to assess the system and provide their assessments in written form by filling the form by marking the level of their acceptance by scaling from 1(bad) to excellent(5) for both intelligibility and naturalness. The questioner for assessment is shown on Appendix E. The result of the assessment is discussed as follows.

5.3 Test Result

The evaluation result for both intelligibility and naturalness is showed on Table 5.1 and Table 5.2 below for unmarked notes. As shown in the result the percentage of naturalness with average point is 37.5 percent. And for intelligibility test also shows the percentage for the level of understanding what the lyrics (word) contain. The level of hardness shows for more than 60 percent of the user it is hard and above.

Table 5.1 Naturalness test for unmarked notes

scale	Scale description	No. of Subjects	Percentage (%)
5	Very natural	0	0
4	natural	1	12.5
3	good	2	25
2	unnatural	3	37.5
1	Very unnatural	2	25

Table 5.2 Level of difficulty for unmarked notes

Scale	Scale description	No of Subjects	Percentage (%)
1	Very hard	2	25
2	Hard	3	37.5
3	Neither	3	37.5
4	Easy	1	12.5
5	Very Easy	0	0

The evaluation result for both intelligibility and naturalness for marked notes is showed on Table 5.3 and Table 5.4 below respectively. As shown in the result the percentage of naturalness with average point is 75 percent. And for intelligibility test, which shows understanding of what is the lyrics (word) it contains, for more than half of the users it is not hard to understand.

Table 5.3 Naturalness test for marked notes

scale	Scale description	No. of Subjects	Percentage (%)
5	Very natural	1	12.5
4	natural	5	62.5
3	good	2	25
2	unnatural	0	0
1	Very unnatural	0	0

Table 5.4 Level of difficulty for marked notes

Scale	Scale description	No of Subjects	Percentage (%)
1	Very hard	0	0
2	Hard	2	25
3	Neither	2	25
4	Easy	3	37.5
5	Very Easy	1	12.5

It can be summarized by putting the above result in a MOS evaluating structures as indicated in Table 5.5

Table 5.5 MOS test values for marked and unmarked notes

Level/ type	Unmarked Note	Marked Note
Naturalness	2.3	3.9
Intelligibility	2.6	3.4

We can summarize from the result obtained above that, there is high possibility of synthesizing notes if the appropriate marking and pitch parameters are taken. By selecting the best unit from the dataset, the synthesizer produced a note in a given context. Even if the experiment is conducted on a small scale, the result obtained is a promising.

The challenges we face during this research work is that having a lyrics of a song and marking them properly was the most difficult and time consuming one. And during testing period also we got a difficult for most of the users because they are new for such synthetic sounds. But, through repetition of listing and clarification of the question there was a high tendency in assessing the questions accordingly.

Chapter 6

Conclusion and recommendation

6.1 Conclusion

Different speech synthesis techniques are being developed and implemented to generate a natural and intelligible speech as speech synthesis system are becoming more applicable. Moreover, the technologies have also advanced to incorporate speech synthesis system in different computing devices. Now a day the need of synthesis sound is also becoming more crucial in music environment for different purposes. For that reason plenty of works has been done in foreign language songs.

In this thesis work, a first attempt is done to develop a notation synthesizer for St Yared hymn using unit selection concatenation. To come up with this synthesizer, first of all related works on the area of song synthesis was reviewed. In addition to that each St Yared hymn notes are studied using experts who have a deep knowledge and learnt a lot of priests in the hymn Class of the traditional school.

The work begins with the collection of hymn, which includes both text and song. The words in a song is transcribed in a contextual way for a given notes. The recorded song wave signal is also segmented in their respective transcription context. Prosodic labeling of the signal is done manually. From the segmented wave the pitch and duration parameters are stored in a data file.

The main work of the synthesis model is divided in to three basic parts namely: analysis, unit selection and wave synthesis. In the analysis phase the input text is transcribed. Every word that comes to the system should be transcribed to find the contexts of the note going to be synthesized. Temporary list is developed for each target notes until the transcribed word is parsed. The next step is unit selection for the target note depending on the context it is. This was done first searching for candidate units from the training dataset and finally selecting the best match unit signal on the parameter of pitch and degree of context match. The last step is the selected units are concatenated and smoothed to give more natural like sounds. This was done by pitch shifting and amplitude modification of the concatenated waves. A java code is written to analysis and unit selection phases and a matlab code is used for wave synthesis.

For the evaluation of the new synthesizer a formal perceptual test which we call Mean Opinion Score (MOS) is used. Participant who have a knowledge of Hymn are used to rate the utterance spoken of the notes in different contexts. The result of the test indicates synthesis of Notes is a promising field of study for Ethiopian hymn researchers.

6.2 Recommendation

In this thesis work a system is developed to generate notes from a given context, by best selection of units from candidate units. This is just the first attempt towards synthesizing local music (hymn) using common speech synthesis technique. Unit selection concatenation is used since it uses a natural speech unit for generation of the synthesized sounds. Finally what we want to say is that this is the novel work which needs much in the future to develop a fully fledged synthesizer so that end user can benefit from it. By this work we hope we open an eye for the new comers for this area. Things to be considered in the future work may contain

- In this thesis work, unit selection concatenation synthesis is used only for notes, and some consonants are tried to considered by extracting from the song and others are fed from a Geez alphabet. A promising result is found for those consonants considered from song. So to fill this gap there is a need to get all from song.
- There is no Ethiopic song synthesis is done yet. Applying what we develop from this work a lot can be achieved in the music industry of Ethiopia particularly for those which can use St Yared hymn notes.
- Till now we couldn't found duration demarcation for St Yared notes. This has its own impact on the developing of synthesizer, so this should be considered in the future work as one parameter of appropriate unit selection.

References

[1.] Elizabeth Liddy, (2001),"Natural Language Processing". *InEncyclopedia of Library and Information Science, 2nd Ed. NY. Marcel Decker, Inc*.

[2.] D. Sasirekha, (2012),"Text to Speech Simple Tutorial", *International Journal of Soft Computing and Engineering (IJSCE)* ISSN: 2231-2307, *Volume-2, Issue-1*

[3.] Sebsibe HaileMariam, S P Kishore, Alan W Black, Rohit Kumar, and Rajeev Sangal, (2005), "Unit Selection Voice for Amharic using Festivox", *5th ISCA Speech Synthesis Workshop –99 Pittsburgh, page 103-107*

[4.] Eyob B.Kasie, Yaregal Assabie, (2012), "Concattenative Speech Synthesis for Amharic Using Unit Selection Method," *The international ACM Conference on Management of Emergent Digital EcoSystems (MEDES'12)*page 27-31, *Addis Ababa, Ethiopia*

[5.] Ayele Bekerie, (2007), "St. Yared – the great Ethiopian composer",*Tadias Magazine, New York.*

[6.] Habtemariam Workneh, "Ethiopian Ancient curriculum", ጥንታዊ የኢትዮጵያ ሥርዓተ ትምህርት www.ethiopianorthodox.org. Last reviewed on May, 2013.

[7.] Sami Lemmetty, (1999), "Review of Speech Synthesis Technology" , *Master's Thesis, Helsinki University of Technology*

[8.] Thierry Dutoit, (1997), "High-quality text-to-speech synthesis: an Overview", *journal of Electrical & electronics Engineering, Australia: special issue on Speech recognition and Synthesis*, vol. 17 pp. 25-37.

[9.] Sharon Hunnicut, (1980), "Grapheme–to-phoneme rules : a review", *Speech transmission laboratory, Royal Institute of technology, Stockholm, Sweden, QPSR 2-3, pp. 38-60.*

[10.] Richard Sproat, (1998) "The Need for Increased Speech Synthesis Research", *Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis*

[11.] Dennis Klatt, (1987). " Review of TTS Conversion for English", *Journal of Acoustic society of America 82(3), 737-793.*

[12.] Robert Donovan, (1996). "Trainable Speech Synthesis", *Cambridge University, Cambridge University Engineering Department*

[13.] Daniel Jurafsky, (1999), "An Introduction to Natural language processing, Computational Linguistics and Speech recognition" .*Prentice Hall, Englewood cliffs, New jersey 09632.*

[14.] P. Zervas, I. Potamitis, N.Fakotakis, & G. Kokkinakis, (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.*

[15.] Takayoshi Yoshimura, (2002), "Simultaneous modeling of phonetic and prosodic parameters, and characteristics conversion for HMM-Based text-to-speech system", *Doctorial dissertation, Nagoya Institute of Technology.*

[16.] B. Balentine, and D. Morgan, (1999), "How to Build a Speech Recognition Application", *Enterprise Integration Group*.

[17.] Henok Lulseged, (2003), "Concatenative Text-to-Speech (TTS) synthesis for Amharic language", *Master's Thesis, Addis Ababa University*.

[18.] Jhon McNuty ,(2009) "Text to Speech to Singing Voice Synthesis", *Sonic Arts Research Centre Queen's University Belfast*

[19.] Alejandro Ramos, (2012),"Singing voice synthesis in Spanish by concatenation of syllables based on the TD-PSOLA algorithm", *Latest Advances in Acoustics and Music*.

[20.] Varvara Kyritsi (2007), "Score-To-Singing voice for Greek language*", proceeding of the inter. Computer music Conference(ICMC07),Copenhagen, Denmark*

[21.] Nadew Tademe, (2008). "Formant Based Speech Synthesis for Amharic Vowels", *Master's Thesis, Addis Ababa University, department of computer science, Addis Ababa, Ethiopia.*

[22.] Bereket Kasaye, (2008). "Dveloping Speech synthesizer for Amharic language using hidden markov model", *Master's Thesis, Addis Ababa University, department of computer science, Addis Ababa, Ethiopia.*

[23.] Mulat Shiferaw, (2012). "Sylable based text to speech synthesis for Amharic Language", *Master's Thesis, Addis Ababa University, department of computer science, Addis Ababa, Ethiopia.*

[24.] Michael Hewitt, (2008), "Music theory for Computer Musicians", *Course technology Ptr, Boston, USA.*

[25.] Jeff fessler, Music Signal Processing, http://web.eecs.umich.edu/, University of Michigan , last reviewed at July,2013

[26.] Mezmur Tsegaye, (2011) "Traditional Education of the Ethiopian Orthodox Church and Its Potential for Tourism Development (1975-present)", *Addis Ababa University*

[27.] L. Neovius and P.Raghavendra (1993). "Comprehension of KTH Text-to-Speech with Listening Speed Program". *Proceedings of Eurospeech 93 (3): 1687-1690.*

[28.] François Grondin, Guitar Pitch Shifter, www.guitarpitchshifter.com, last reviewed on Nov,2013

[29.] Ker Than, (2008), "In Search of Music's Biological Roots", *Duke magazine volume 94*

[30.] Kevin Bockelandt, (2012), Introduction to Musical Notation, http://www.br-gs.com/tutorial/csound/introduction-to-musical-notation.html, last reviewed on April, 2013

Appendix A Note Marker combination

Marker	Note	Symbol	name	description
ꕥ	▪	S1	Selasa yizet	S for Selasa and 1 for Yizet
ꕥ	ᴗ	S2	Selasa deret	
ꕥ	⁝	S3	Selasa rikrik	
ꕥ	ᵔ	S5	Selasa difat	
ꕥ	ᵕ	S6	Selasa chiret	
ꕥ	⸗	S7	Selasa kinat	
ꕥ	⸚	S8	Selasa hidet	
ꕥ	ꞏ	S9	Selasa kurt	
ꔮ	▪	P1	Pier yizet	P for Pier and 1 for Yizet
ꔮ	ᴗ	P2	Pier deret	
ꔮ	⁝	P3	Pier rikrik	
ꔮ	ᵔ	P5	Pier difat	
ꔮ	ᵕ	P6	Pier chiret	
ꔮ	⸗	P7	Pier kinat	
ꔮ	⸚	P8	Pier hidet	
ꔮ	ꞏ	P9	Pier kurt	
ꕤ	▪	B1	Bubey yizet	B for Bubey and 1 for Yizet
ꕤ	ᴗ	B2	Bubey deret	
ꕤ	⁝	B3	Bubey rikrik	
ꕤ	ᵔ	B5	Bubey difat	
ꕤ	ᵕ	B6	Bubey chiret	
ꕤ	⸗	B7	Bubey kinat	
ꕤ	⸚	B8	Bubey hidet	
ꕤ	ꞏ	B9	Bubey kurt	
ꕣ	▪	A1	Aynu yizet	A for Aynu and 1 for Yizet
ꕣ	ᴗ	A2	Aynu deret	
ꕣ	⁝	A3	Aynu rikrik	
ꕣ	ᵔ	A5	Aynu difat	

ዓይ	◞	A6	Aynu chiret	
ዓይ	⸗	A7	Aynu kinat	
ዓይ	◟	A8	Aynu hidet	
ዓይ	ኮ	A9	Aynu kurt	
ዕለ	▪	I1	Ile yizet	I for Ile and 1 for Yizet
ዕለ	◡	I2	Ile deret	
ዕለ	⁞	I3	Ile rikrik	
ዕለ	◠	I5	Ile difat	
ዕለ	◞	I6	Ile chiret	
ዕለ	⸗	I7	Ile kinat	
ዕለ	◟	I8	Ile hidet	
ዕለ	ኮ	I9	Ile kurt	
ጸጋ	▪	T1	Tsega yizet	T for Tsega and 1 for Yizet
ጸጋ	◡	T2	Tsega deret	
ጸጋ	⁞	T3	Tsega rikrik	
ጸጋ	◠	T5	Tsega difat	
ጸጋ	◞	T6	Tsega chiret	
ጸጋ	⸗	T7	Tsega kinat	
ጸጋ	◟	T8	Tsega hidet	
ጸጋ	ኮ	T9	Tsega kurt	
ቡር	▪	R1	Burkt yizet	R for Burkt and 1 for Yizet
ቡር	◡	R2	Burkt deret	
ቡር	⁞	R3	Burkt rikrik	
ቡር	◠	R5	Burkt difat	
ቡር	◞	R6	Burkt chiret	
ቡር	⸗	R7	Burkt kinat	
ቡር	◟	R8	Burkt hidet	
ቡር	ኮ	R9	Burkt kurt	

Appendix B Sample of Hymn phone dataset

Left	Target	Right	index	duration	Start Pitch	End Pitch
A	P2	B5	1	423265	140	142
A	S1	LE	1	673447	112	128
A	B1	A5	1	440454	136	164
A	A1	TA	1	507324	189	189
A	P1	QE	1	455248	180	187
A	S1	ME	1	604628	132	84
A	S8	D	1	520635	185	85
A	S8	WU	1	194694	190	136
A	T1	RE	1	1246463	75	169
A	P1	A5	1	388980	162	140
A1	P6	WIE	1	651111	162	72
A1	I5	A2	1	908730	187	214
A1	B2	A5	1	459728	166	166
A1	I5	A2	2	725918	198	208
A1	B6	ME	1	974218	276	73
A2	B6	B2	1	306009	192	17
A2	LIE	S1	1	267710	97	106
A2	B6	B2	2	831111	193	162
A2	I5	#	1	568753	136	151
A2	S5	BE	1	1728798	142	171
A5	B2	P6	1	298798	164	161
A5	P6	#	1	750658	184	142
A5	P6	#	2	571020	166	140
A5	P6	#	3	490930	164	149
A5	P6	B5	1	561905	165	71
A5	B2	S6	1	362336	93	166
A5	B2	P6	1	298798	164	161

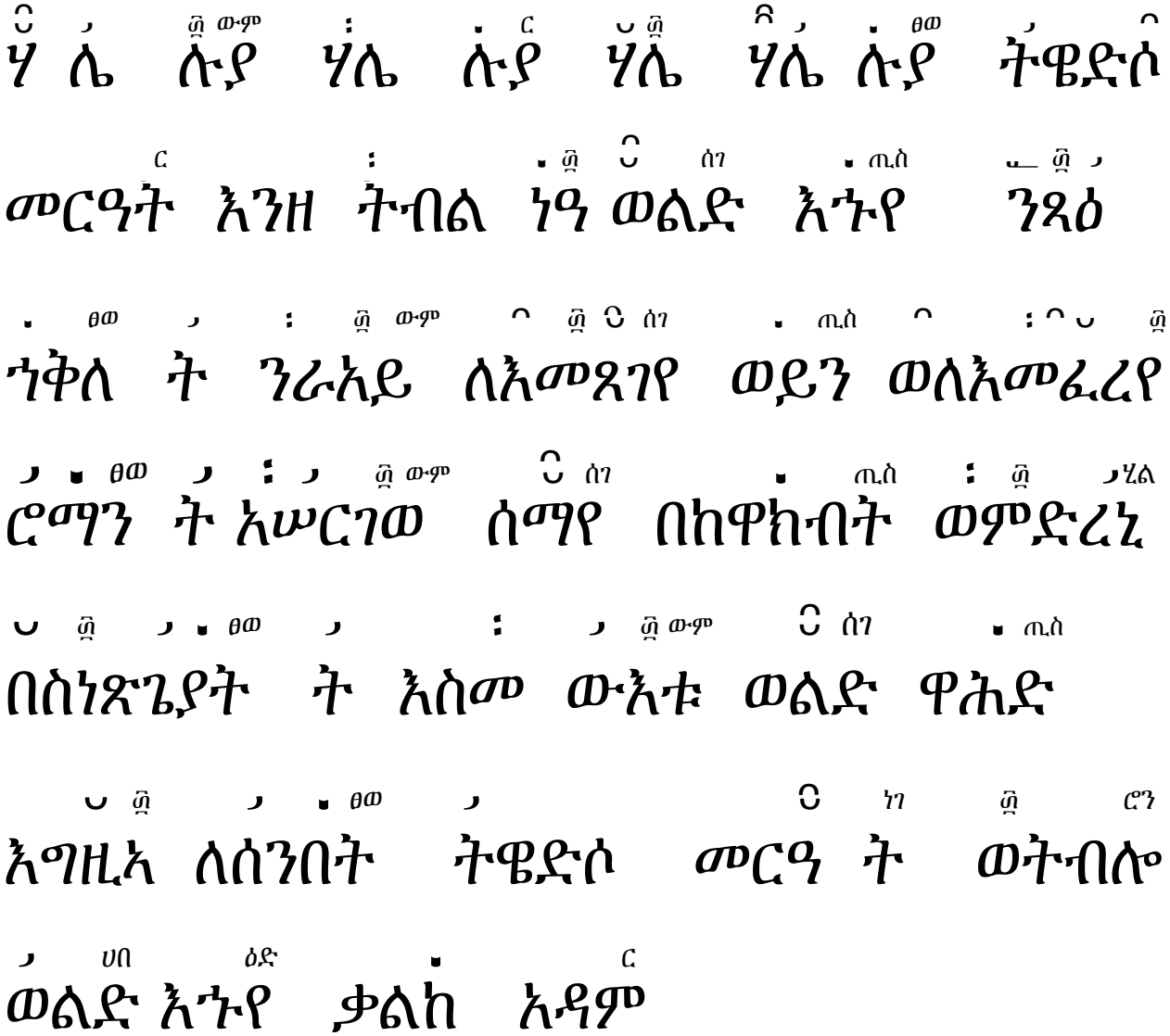
A5	B2	P6	2	257007	173	169
A5	B2	P6	3	155556	165	166
A5	B2	P6	4	223061	162	163
A5	B2	P6	5	166735	167	165
A5	B2	A5	1	474422	169	172
A5	B2	S6	2	312789	176	169
A5	B2	S6	3	259388	169	171
A5	B2	S6	4	237256	166	172
A6	P6	B5	1	732766	100	143
AY	S1	A5	1	715351	131	140
BE	A2	S	1	714	123	88
B1	A5	B2	1	880	165	183
B1	A5	B2	2	796	172	166
B1	A5	B2	3	763	171	180
B1	A5	B2	4	852	176	170
B1	A6	P6	1	770	169	93
B1	A6	#	1	577574	170	114
B1	P2	B5	1	942177	199	194
B1	P6	#	1	982018	221	186
B1	p6	P2	1	1266848	203	190
B2	A5	B2	1	1282268	168	178
B2	P6	Lie	1	1047732	170	121
B2	A5	B2	2	789796	168	183
B2	A5	B2	3	620295	167	178
B2	A5	B2	5	882290	185	183
B2	S6	#	1	337574	169	138
B2	P6	B	1	627800	183	130
B2	P6	LIE	1	1047732	170	121
B2	P6	LIE	2	852472	175	119
B2	P6	R	1	836213	174	124
B2	P6	SE	1	694898	174	123

Appendix C **Geez *fidel* data set**

#	A	#	67	67
#	AA	#	122	129
#	BA	#	120	120
#	BE	#	129	132
#	BI	#	119	124
#	BIE	#	128	130
#	BO	#	110	110
#	B	#	127	133
#	BU	#	133	124
#	DA	#	126	132
#	DE	#	128	137
#	DI	#	131	130
#	DIE	#	127	140
#	D	#	130	152
#	DO	#	119	115
#	DU	#	128	128
#	FA	#	137	129
#	FE	#	125	125
#	FI	#	128	128
#	FIE	#	128	130
#	F	#	150	155
#	FO	#	107	110
#	FU	#	132	124
#	GA	#	128	126
#	GE	#	130	133
#	GI	#	136	134
#	GIE	#	127	131
#	G	#	130	153
#	GO	#	122	115
#	GU	#	131	137

#	HA	#	122	145
#	HE	#	115	72
#	HI	#	136	139
#	HIE	#	135	140
#	H	#	162	165
#	HO	#	112	119
#	HU	#	135	139
#	I	#	127	130
#	IE	#	127	132
#	IX	#	81	78
#	QA	#	147	147
#	QE	#	488	137
#	QI	#	133	132
#	QIE	#	213	131
#	Q	#	136	163
#	QO	#	177	108
#	QU	#	135	138
#	LA	#	126	140
#	LE	#	150	137
#	LI	#	141	137
#	LIE	#	144	135
#	L	#	147	82
#	LO	#	118	113
#	LU	#	133	137
#	MA	#	142	123
#	ME	#	134	137
#	MI	#	143	139
#	MIE	#	143	134
#	M	#	167	162
#	MO	#	163	113
#	MU	#	146	138

Appendix D Halie luya song with notes



Appendix E Questionnaire

The aim of this questioner is to evaluate the performance of St Yared hymn notation synthesizer with both naturalness and intelligibility. Listen to the audio files and fill the Table with rate you judge on the space provided

2. How do you judge the naturalness of the synthesized voice?

1. Excellent 2. Very Good 3.Good 4.Poor 5.Very poor

3. Intelligibility evaluation: Was the voice easy to understand?

1. Very hard 2. Hard 3.Neither hard 4.Easy 5.Very easy

words	Unmarked notes		Marked notes	
Test	Naturalness	Intelligibility	Naturalness	Intelligibility
▪ ▪ ▪ አ ቀ መ				
ᴗ ▪ ᴖ በ ሃ				
▬ ▪ ብሲ ተ				
ᴗ ▪ ቤ ተ				
▬ ▪ ፊሪ ሆ				
⋮ ▪ ኤ ለ				
ᴗ ▪ ▪ ነ በ ረ				
ᴖ ᴐ ቃሌ				
▪ ᴐ ▪ ቆ የ				
▪ ᴖ ▪ ሰማ የ				

Declaration

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

Declared by:

Name: Girma Zemedu

Signature : _____

Date: _____

Approved by:

Name: Yaregal Assabie (PhD)

Signature: _____

Date: _____

Place and date submission: Addis Ababa, March 2014