TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

CENTRAL CAMPUS PULCHOWK

THESIS NO.:069MSCS660

Nepali Text to Speech using Time Domain Pitch Synchronous Overlap Add Method

By

Pratistha Malla

A THESIS

SUBMITTED TO THE DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SYSTEM AND KNOWLEDGE ENGINEERING

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

FEBRUARY 2015

Nepali Text to Speech using Time Domain Pitch Synchronous Overlap Add Method

By

Pratistha Malla

(069/MSCS/660)

Thesis supervisor:

Dr. Basanta Joshi

Lecturer Electronics and Computer Engineering

Institute of Engineering, Nepal

A thesis submitted in partial fulfillment of the requirements for the

Degree of Master of Science in Computer System and Knowledge Engineering

Department of Electronics and Computer Engineering

Institute of Engineering, Pulchowk Campus

Tribhuvan University

Lalitpur, Nepal

February, 2015

COPYRIGHT

APPROVAL PAGE

TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

PULCHOWK CAMPUS

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

The undersigned certify that it has been read and recommended to the Department of Electronics and Computer Engineering, for acceptance, a thesis entitled "Nepali Text to Speech using Time Domain Pitch Synchronous Overlap Add Method", submitted by "Ms. Pratistha Malla" in partial fulfillment of the requirement for the award of the degree of "Master of Science in Computer System and Knowledge Engineering".

_____

Supervisor, Dr. Basanta Joshi
Lecturer, Electronics and Computer Engineering

_____

External Examiner,
Xxxx
xxxxx

_____

Date

ACKNOWLEDGEMENT

I would like to express my deepest appreciation to all those who provided me the possibility to complete this report. I am thankful to Prof. Dr. Shashidhar Ram Joshi, Asst. Prof. Dr. Sanjeeb Prasad Panday, Dr. Aman Shakya, Prof. Dr. Subarna Shakya, Dr. Arun Timilsina, Asst. Prof. Dr. Dibakar Raj Pant and Mr. Baburam Dawadi for granting me the permission to do this thesis. I have to appreciate the guidance given by the panels especially in our thesis presentation that has improved mine presentations skills. I deeply appreciate the continuous guidance, ideas and supervision provided by my supervisor Dr. Basanta Joshi. I am very thankful to him for showing the right direction for the progression of thesis.

Also I would like to express my gratitude to Mr. Pradhyumna Kumar Shrestha for all the possible help and guidance and support for the completion of my thesis and without whom I would have been lost.

Last but not least, sincere thanks to all my friends and family for continuous direct/indirect help, support and motivation without whom I could not have carried on with this thesis

Pratistha Malla

(069MSCS660)

ABSTRACT

Nepali text to speech has range of application. Nepali Text to speech synthesizer provides help in reading for differently-able and illiterate people. Text to speech is the system of converting Nepali written text to speech through text analysis and speech synthesis. Since Nepali is phonetically rich language, the system uses concatenative approach. The Time Domain Pitch Synchronous Overlap Add (TD-PSOLA) concatenative synthesis is simple and efficient concatenative method. It is based on pre-recorded samples, cut this up into small pieces and then recombines these to form new speech. The raw Nepali text undergoes the letter to sound conversion. The letter to sound conversion is based on diphones concatenation. The diphone dictionary is created from the pre-recorded speech signals. The speech database consists of the diphones, start position, end position of speech segment extracted from pre-recorded speech. For each diphone signal the pitch is determined using autocorrelation method. Hanning window and Hamming window has been used for windowing of speech signal. TD-PSOLA method implies the signals are overlapped and added to generate the synthetic speech signal. Hence TD-PSOLA concatenative method has been implemented to generate synthetic speech signal for Nepali text.

Key words: Diphone, Text to speech (TTS), speech synthesis, TD-PSOLA

TABLE OF CONTENTS

TITLE **PAGE**

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURE

FIGURE **PAGE**

LIST OF ABBREVIATIONS

CODE	FULL-FORM
ASCII	American Standard Code for Information Interchange
DSP	Digital Signal Processing
ESNOLA	Epoch Synchronous Non Overlap Add
FD-PSOLA	Frequency Domain Pitch Synchronous Overlap Add
FFT	Fast Fourier Transform
IFFT	Inverse Fast Fourier Transform
MARY	Modular Architecture for Research on speech sYnthesis
MBROLA	Multi Band Resynthesis Overlap Add
NLP	Natural Language Processing
PDA	Pitch Detection Algorithm
PSOLA	Pitch Synchronous Overlap Add
SIOD	Scheme In One Defun
TD-PSOLA	Time Domain Pitch Synchronous Overlap Add
TTS	Text To Speech

CHAPTER ONE: INTRODUCTION

1.1 Background

Mostly the documents are in written text form. It is difficult for the illiterate and handicap people to read or gain knowledge from the written form. The Text To Speech (TTS) system has made their accessibility to written documents easy. Text to speech synthesis is the automated transformation of a text to speech that sounds like native speaker reading the text. TTS system takes the input text in standard format. A full text to speech system should be able to handle the numbers, date and abbreviation in the input text, with reasonable tolerance. Vaguely TTS systems can be of two types: limited domain TTS and generic TTS. Limited domain TTS is built to serve a specific purpose which has limited application. Such system uses limited words and sentences for speech synthesis. Generic TTS system is capable of reading anything from the document. The main properties of the synthetic speech signal are naturalness and intelligibility. TTS conversion should generate intelligible speech with human like voice for naturalness from the text.

The input text is an ASCII characters which are broken into characters for processing. Each sentence is broken into words and words are broken into characters. Each character undergoes text analysis and normalized for speech synthesis. There are many approaches for the generation of synthetic speech. Most popular and easy to implement and that gives more naturalness synthetic speech is concatenative method. Conatenative method generates the speech signal by the concatenation of the pre-recorded speech segments which produces natural speech. Most popular concatenative approach is Time Domain Pitch Synchronous Overlap Add method. The speech segments can be demi-syllables, diphones or phonemes or word etc. A system that stores phones or diphones provides large output range. For English, ASCII is used whereas for Nepali Unicode is preferred. Nepali is the phonetically rich language. Nepali has 13 phonologically distinct vowels and 33 consonants. The phoneme does not have meaning on its own. The phonemes combine to form a word. A group of phoneme forms a syllable. Syllables determined the prosody (way the word is spoken) of the word. Syllables are those sounds that are produced by vocal cord. Diphone is a stable part of speech between two phones. It is an adjacent pair of phones. If the number of phones in a language is P then the possible

numbers of diphones are P^2. Diphones are useful in speech synthesis as the pre-recorded diphones are joined to create the synthesize speech. Diphonic concatenation produces much more natural speech.

1.2 Problem in Speech Synthesis

There are many problems in speech synthesis. As human communicate not only with speech but also with gesture and facial expression. The Text to speech system does not corporate these features. The synthetic speech may not have the expression and emotion of the written text as that of actual human speech. The concatenation method requires large database for the pre-recorded words. It is difficult to develop the pre-recorded database in noiseless environment. The letter to sound conversion is also the difficult part of TTS system. In concatenation method, required speech portion are extracted from the pre-recorded speech signal and are added to generate the synthesized signal, so there may be discontinuities in the wave. The correct pronunciation of the words is also a major problem. The pronunciation of the words may differ then the natural sounding. There are many problems in processing of numbers, abbreviations and acronyms. The analysis of quality of the synthesized speech is difficult job. It needs proper method to determine the synthesized speech. The problems in text to speech can be explained as

1.2.1 Letter to sound conversion

It is a difficult task to convert the input text to the corresponding speech signal. A large set of rules are required to produce the correct synthetic speech signal. For the proper letter to sound conversion, it depends on the text preprocessing and analysis. Numbers are the most difficult to convert correctly into the speech form. If the input is dates, numbers, acronyms etc the digits and numbers should be properly expanded. It is difficult and complex task if input includes fraction like 2/5 which may represent as two-fifth or the second of May if date. The first three ordinals must be expanded differently than the others like 1st as first, 2nd as second and 3rd as third. Also there will be contextual problem Roman numerals. For the roman representation character ‘I’ can be pronounced

as pronoun ‘I’ or as a number. Similarly for Roman numeral ‘III’ in Chapter III can be pronounced as Chapter three whereas in Henry III Roman numeral III is pronounced as Henry the third. Abbreviations should be represented in expanded form. But there may contextual problems for text like Dr. which can be Doctor or drive. Special characters like $, /, % also causes problems in conversion. For example $100 million has to be read as one hundred million dollar not as one hundred dollars million.

1.2.2 Pronunciation

Even though Nepali language is phonetically rich language, it is pronounced as it is in written text. It is difficult task for correct pronunciation in speech synthesis field for the texts which are spelled in same way but meanings are different. For example word ताल has different meanings as below and are spelled same but are pronounced differently.

ताल:पोखरि

ताल :बसाइ, हिडाइको ढङ्ग

ताल :चोटि मौका

The pronunciation of such words may be different due to contextual effects.

1.2.3 Prosody

Prosody deals with the correct intonation, stress and duration from the written text. It is the most difficult step in Text to Speech system. The prosody of continuous speech depends on many separate aspects such as the meaning of the sentence and the speaker characteristics and emotions. The prosodic dependencies are shown in figure 1.1.

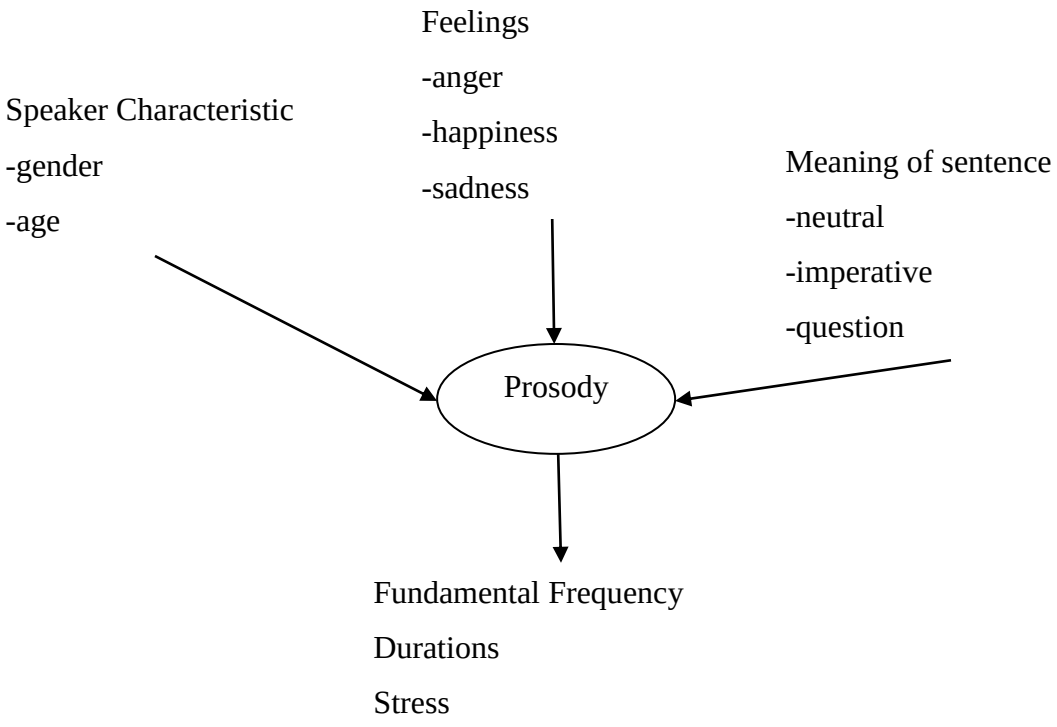

Figure 1.1 Prosodic Dependencies

1.2.4 Problem in Low Level Synthesis

There are many methods to generate the synthetic speech signals, each with its own pros and cons. The main problem in concatenative method is the creation of large speech database. The use of large database may cause problems related to memory and system requirements. The sampling and labeling of each pre-recorded sound signal is time consuming. While concatenating the speech signal there may cause distortion to the speech signal.

1.3 Objective

The main objective of this thesis is:

- To synthesize speech from Nepali text using Time Domain Pitch Synchronous Overlap Add method of concatenation.

1.4 Application

The text to speech system is highly researched field. Speech synthesis has many applications. A TTS is a system that reads text aloud automatically. So it can be used for educational purposes. Since it is difficult for the visually impaired, illiterate and blind to read and understand the content. The TTS system provides the means to help them understand the content whether the input is introduced by input stream or a scanned input. It can act as accessibility tool for vocally impaired, instead of using the sign language the vocally impaired can use TTS system to communicate with others. It can also be used to teach pronunciation and spelling of language. It can also be used for mobile application like voice enabled email, talking clock. TTS system can also be used as notifier for emails, time in mobile phones. TTS is counterpart of speech recognition so it can be further implemented to develop speech recognition.

1.5 Organization of report

The report is organized in chapters. The report begins with the brief introduction to the topic and gives definition of the text to speech system. The chapter 1 also gives introduction about the TTS system. Chapter 1 also states the problems in speech synthesis and problems in low level synthesis. The objective of the as well as the application of the text to speech system is mention in chapter 1.

Chapter 2 is the literature review about this thesis. It describes about the different previously done work in text to speech and speech synthesis. It also gives brief overview about the existing TTS systems like MARY and Festival speech synthesis. It includes the general approach for the development of the text to speech system.

Chapter 3 gives the brief explanation about the speech synthesis methods like articulatory synthesis, formant synthesis and concatenative synthesis. It also gives brief introduction to PSOLA concatenative method and TDPSOLA method.

Chapter 4 is the methodology and algorithm implemented in this thesis.

Chapter 5 is result and discussion which depicts the output generated in this thesis. Also the output has been analyzed and evaluated.

Report ends in chapter 6 with the conclusion and recommendation.

CHAPTER TWO: LITERATURE REVIEW

2.1 Existing Systems

2.1.1 MARY Text to Speech

MARY (Modular Architecture for Research on speech sYnthesis) is a system for Germany text to speech synthesis. The system's main features, namely a modular design and an XML-based system-internal data representation, are pointed out, and the properties of individual modules are presented. MARY system uses an XML-based data representation, it not only displays the intermediate processing results but also allow their modification by user [1]. MARY system used MBROLA for synthesizing the utterance. MBROLA was selected because of the comparatively low degree of distortion introduced into the signal during digital signal processing.

Four parts of the TTS system can be distinguished [1]:

a. the preprocessing or text normalization;
b. the natural language processing, doing linguistic analysis and annotation
c. the calculation of acoustic parameters, which translate the linguistically annotated symbolic structure into a table containing only physically relevant parameters;
d. and the synthesis, transforming the parameter table into an audio file

2.1.2 Festival Speech Synthesis System

Festival speech synthesis offers a general framework for building speech synthesis systems as well as including examples of various modules. As a whole it offers full text to speech through a number APIs: from shell level, though a Scheme command interpreter, as a C++ library, from Java, and an Emacs interface [2]. Festival is multi-lingual (currently English (British and American), and Spanish) though English is the most advanced. Other groups release new languages for the system. The system is written in C++ and uses the Edinburgh Speech Tools Library for low level architecture and has a Scheme (SIOD) based command interpreter for control. Festival is free software.

2.2 Related Work

A text to speech system should be able to convert any text to speech. TTS is an automated system that transforms text to speech whether it be the computer input stream or the scanned text. Nepali Text to speech synthesis system which is based on concatenative approach employing ESNOLA, which uses dictionary having raw sound signal representing parts of phonemes as a speech database. ESNOLA method provides the complete control on implementation of intonation and prosody [3]. This system develops an un-intonated (flat) signal where the pitch of pre-recorded speech signal remains the same throughout [3]. The basic database is created consisting of the pre-recorded natural speech in Nepali. Text analysis is the front end of the TTS. The text is first normalized and it transforms the input sentences into manageable lists of word-like units and stores them in the internal data structure [3]. While for synthesis the ESNOLA method has been used in [3]; in which the phonemes obtained from text analyzer are concatenated to get the raw output signal. There are many other speech synthesis methods that uses the pre-recorded speech signals to generate the raw signal. Mostly the concatenation method is widely used dues to its simplicity and efficiency. The text analyzer converts the text to phonemes through grapheme to phoneme conversion. There are many methods for grapheme to phoneme conversion like Support vector machine, hidden Markov Model. A rule based grapheme to phoneme mapping has been mostly used for languages like Hindi [4]. A set of rules has been developed for phoneme mapping. The text to speech synthesis has been developed for languages like English, French, Finnish, and Germany. Mostly the concatenative method has been used for speech synthesis. The framework for multilingual text to speech system has been developed in [5]. It discusses the necessary steps required to build the synthetic voice from the scratch in a new language. It concerns the building of a new voice without recording any new acoustic data and the restrictions that imposes. The main goals of Text to Speech is to describe specific tasks concentrated during Text to Speech conversion namely preprocessing & text detection, text normalization, prosodic phrasing, acoustic processing [6]. Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems that store the phones or diphone provide the

largest output but may lack clarity [7]. The major operation of the high level synthesis is the text analysis and grapheme to phoneme conversion. For grapheme to phoneme conversion mainly rule based solutions has been used. The rule based grapheme to phoneme is simple to implement and requires low memory consumption and efficient in computation. In rule based solution, pronunciation rules are generated from the phonological knowledge of dictionaries [7]. PSOLA is a speech processing technique used in speech synthesis applications. It allows pitch and/or duration modifications of speech. The speech modifications are performed either in the frequency domain (FD-PSOLA), TD-PSOLA depending on the length of the window used in the synthesis process [8]. While TD-PSOLA is capable of modifying both duration and pitch scales, FD-PSOLA can only modify pitch scale [8]. It is difficult to create database of number of phonemes, syllables, words for a particular language. And then concatenating these syllables, phonemes to produce an effective speech as output from TTS engine [9]. The text to speech conversion may be done in different steps: text preprocessing, text analysis, text phoneziation, prosody generation and then the speech synthesis using various algorithms [9, 10]. The steps to convert the text to speech are shown in figure 2.1.

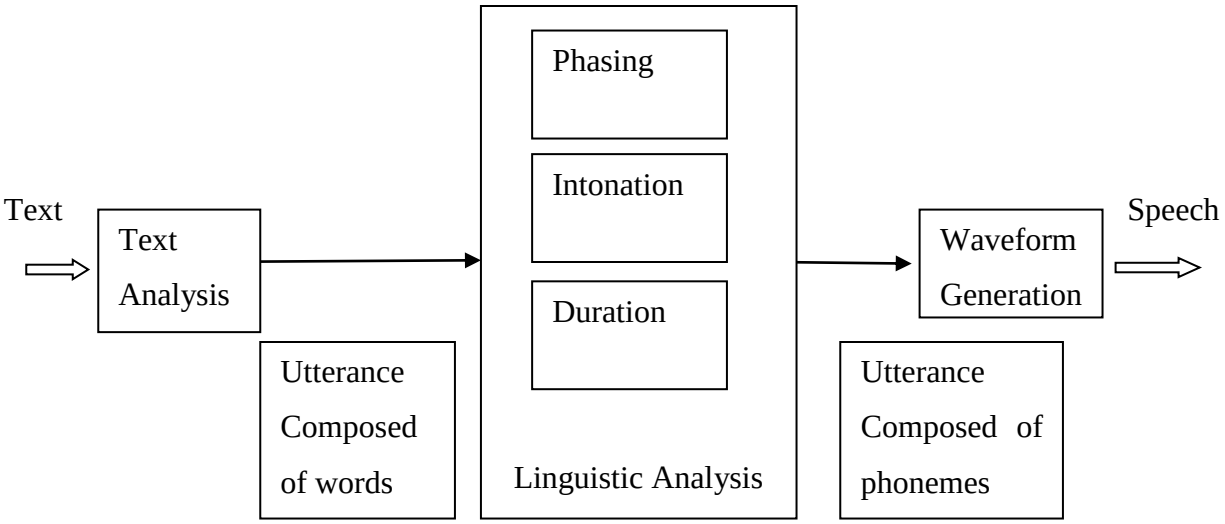

Figure 2.1 Steps in TTS System

Similar sounding words are recorded and saved in .wav format. The recorded sounds are sampled and then the sampled values are taken and separated into their constituent phonetics. The separated syllables are then concatenated to reconstruct the desired words [11]. The input text is first converted into phonetic transcription using Letter-to-Sound rule. For synthesis of new speech TTS system selects the recorded phoneme units from database file and modifies the duration according to the rule based on spelling using TD-PSOLA. The modified phoneme units are concatenated by synchronizing the pitch-periods [12].

CHAPTER THREE:SPEECH SYNTHESIS METHODS

There are generally three popular methods for generation of synthetic speech signal namely

a. Articulatory Synthesis
b. Formant Synthesis
c. Concatenative Synthesis

Formant and concatenative synthesis are mostly use in speech synthesis.

3.1 Articulatory Synthesis

Articulatory synthesis method is based on model of human vocal tract. This method tries to produce the high quality synthetic speech signal. But it is the most difficult method to implement. It involves models of human articulators like tongue, jaw and lips and vocal cord. This method deals with the generation of synthetic speech by modifying position of the speech articulators, such as the tongue, jaw, and lips which is represented by devices like tubes, bellows and pipe. Hence this synthesis tries to mimic the vocal tract using sources and filters. The first articulatory model was based on a table of vocal tract area functions from larynx to lips for each phonetic segment (Klatt 1987). Different sounds are produced by changing the shape of the vocal tract. For articulartory synthesis the correct articulator parameters are difficult to determine. These parameters should be extracted from X-ray photography, MRI or EMS imaging. Also, the movements of tongue are so complicated that it is almost impossible to model them precisely. Because of this difficulty the articulatory synthesis is not best technique for speech synthesis.

3.2 Formant Synthesis

The vocal tract has certain major resonant frequencies. These frequencies change as the configuration of the vocal tract changes. The vowels are characterized by combination of frequencies in different relationships to each other; these characteristics frequency is called formants. It is because of this formant human ear is able to differentiate one speech from other. Formants are measured using a spectrogram or a spectrum analyzer. This composition of speech sounds led to formant synthesis which is often called synthesis by

rule. Formant synthesizer makes use of the acoustic tube model. A basic block diagram of formant synthesizer is shown in figure 3.1 below [13].

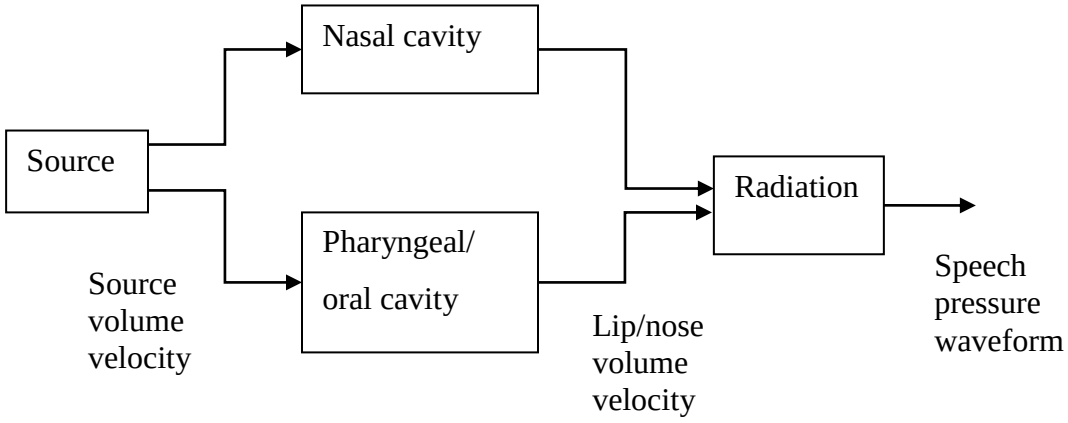

Figure 3.1 Block Diagram of Basic Formant Synthesizer

The sound is generated from source which is periodic for voiced sounds. The source signal is fed into the vocal tract model. In formant synthesizer the nasal cavity and oral cavity is modeled parallel. So the signal passes into these components parallel. The output from these components are combined and passed through radiation component which simulates the load and propagation characteristics of the lips and nose. Formant synthesizer could only produce lowest two formants. Hence as formant synthesis requires spectrogram to determine the formant position of the speech sound it is not a feasible method for synthesis.

3.3 Concatenative Synthesis

The most popular synthesis method is concatenative synthesis method although it is usually limited to single speaker. Concatenative synthesis method is simple and easy to implement. It is an efficient method to generate synthetic speech signal. Although concatenative method is time consuming it is mostly implemented as the synthetic speech signal generated from concatenative synthesis is much natural and understandable to human ear. This method produces the speech signal which depends on the pre-recorded

speech signal. The only pre-requisite for concatenative method is the pre-recorded speech signal. So unlike formant synthesis which tries to create synthetic speech signal from its fundamental frequency, concatenative synthesis concatenate the speech segments like phonemes, syllables, diphones etc to generate synthetic speech signal. Concatenative synthesis depends on the length of samples of recorded sound. There are different concatenative synthesis methods like TDPSOLA, MBROLA, PSOLA, FDPSOLA, ESNOLA. The Nepali text to speech system described here is based on TDPSOLA.

PSOLA is the most widely used second generation signal processing method. It modifies the pitch and timing of the speech signal. It uses the short term signal from the pre-recorded speech that is used for the concatenation. The short term signals are usually two pitch period long windowing signal frame. PSOLA works by dividing the speech waveform in small overlapping segments. To change the pitch of the signal, the segments are moved further apart to decrease the pitch or closer together to increase the pitch. To change the duration of the signal, the segments are either repeated multiple times to increase the duration or eliminated to decrease the duration. The segments are then combined to produce synthetic signal.

3.3.1 Time Domain Pitch Synchronous Overlap And Add (TDPSOLA)

TDPSOLA is the popular PSOLA method for overall pitch and timing modification of the speech signal. It generates the synthetic speech signal from the pre-recorded speech signal by concatenating the frames of speech signal. It works pitch synchronously with one analysis frame per pitch period. Pitch period is the duration in which the signal repeats itself. The important step in TDPSOLA is the determination of the pitch of the speech signal. There are many pitch detection algorithm like cepstrum, linear prediction analysis, autocorrelation method. The easy to implement to determine the pitch of signal is autocorrelation method. The autocorrelation of the signal is defined as in equation i.

$$R[n] = \sum_{m=-\infty}^{\infty} y[m]y[n-m] \qquad i$$

Since the speech signal is separated into frames using Hanning window so the range used will be $0 \leq n < N$. Each frame is centered at the pitch of the speech signal. These window frames can then be recombined to generate the final synthetic speech. The basic operation of the PSOLA is shown in figures below extracted from [13]. In figure 3.2 the sample of voice waveform with its pitch position indicate is shown [13].

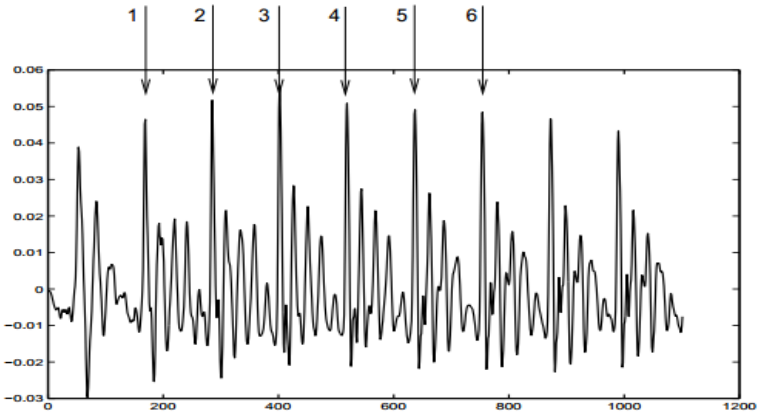

Figure 3.2 Sample of Voice Waveform with Pitch Position

A frame is created at each pitch of the waveform which is shown in figure 3.3 as below [13].

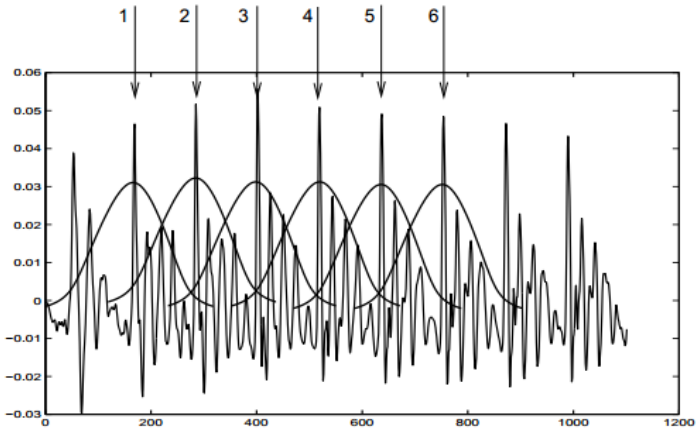

Figure 3.3 Frame Created at Each Pitch Position

Separate windowed signals are formed from Hanning window. Sequence of the separate windowed signal is shown in figure 3.4 below [13].

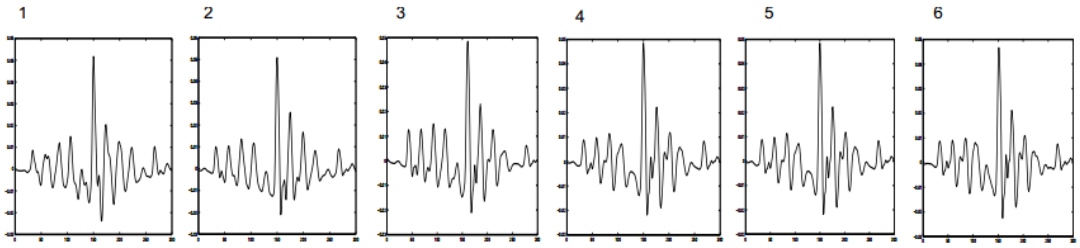

Figure 3.4 Sequence of Windowed Signal

Such separate frames are recombined by overlap add to generate the synthetic speech signal which is shown in figure 3.5 [13].

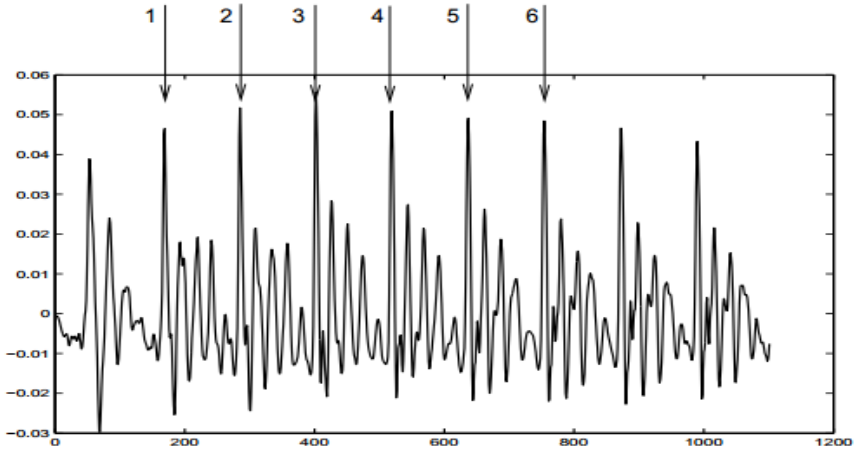

Figure 3.5 Separate Frames Recombined by Overlapping Adding Separate Frames

Since PSOLA technique modifies both pitch and timing of the speech signal. Time scale modification can be obtained by duplicating the speech frame or by eliminating the speech frame. The figure 3.6 shows the time scale modification where the pitch is kept same but the speech signal is made longer by duplication the frame of speech signal [13].

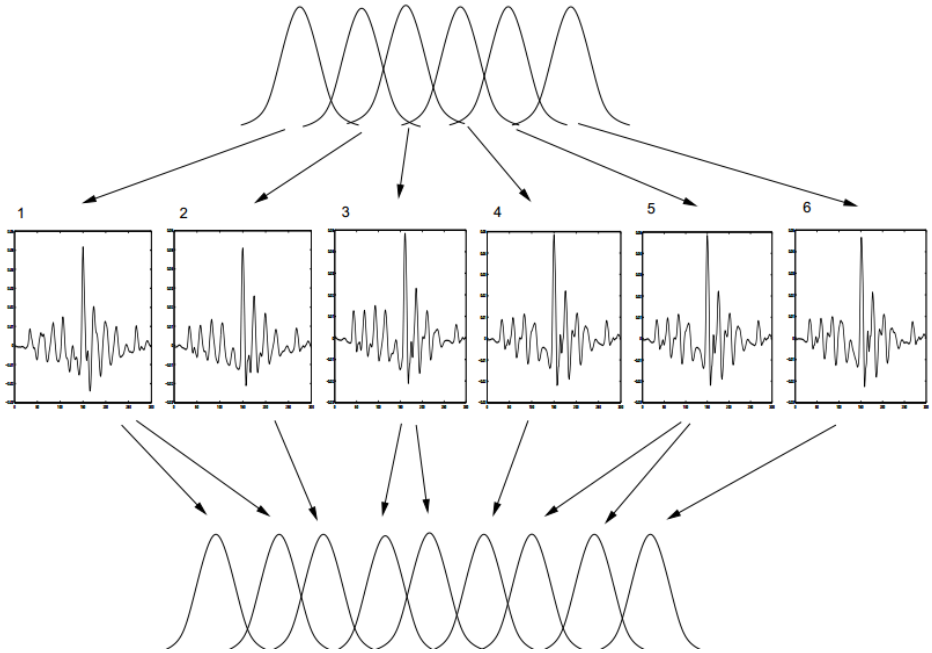

Figure 3.6 Time Scale Modification

When the frames are duplicated and are recombined by overlap add, the signal lengthens but will create the signal which will be identical to the original speech waveform. Similarly if the speech frame is eliminated the length of the signal will be shorter but it will be negligible to the listener of the speech signal.

Similarly pitch scale modification can be obtained by recombining the speech frames different epochs/frequency. This will result in the pitch synchronous speech frames. If the pitch is lowered the speech frames will be apart from each other whereas the high pitch speech signal will be generated if the speech frames are recombined closer to each other. This can be seen in figure 3.7 [13].

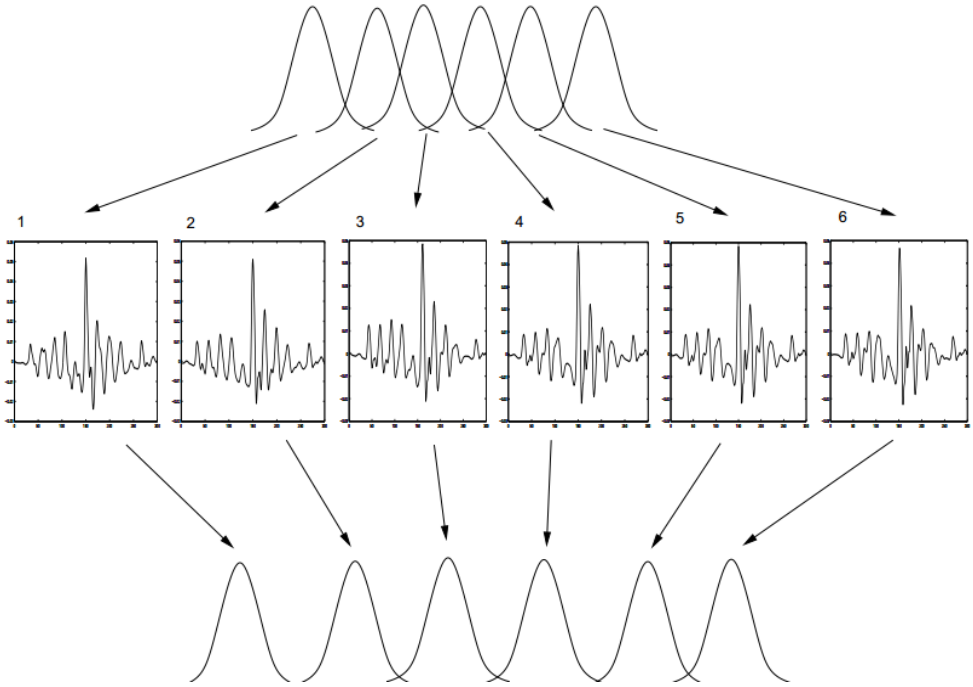

Figure 3.7 Pitch Scale Modification

A TDPSOLA is simple and efficient concatenative method than other approaches it is most popular method to generate the synthetic speech signal.

CHAPTER FOUR: RESEARCH METHODOLOGY

The TTS system mainly consist of two steps namely Natural Language Processing (NLP) or High level synthesis and Low level synthesis or Digital Signal Processing (DSP). The block diagram representation is shown in figure 4.1.

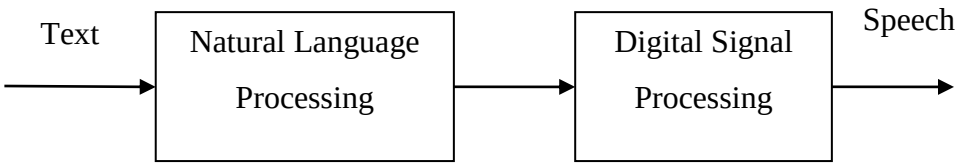

Figure 4.1 Block Diagram Representation of TTS

The high level synthesis mainly deals with the text analysis and low level synthesis deal with the creation, concatenation of the synthesize sound.

For the development of TTS system the main components are the text analyzer and speech synthesizer.

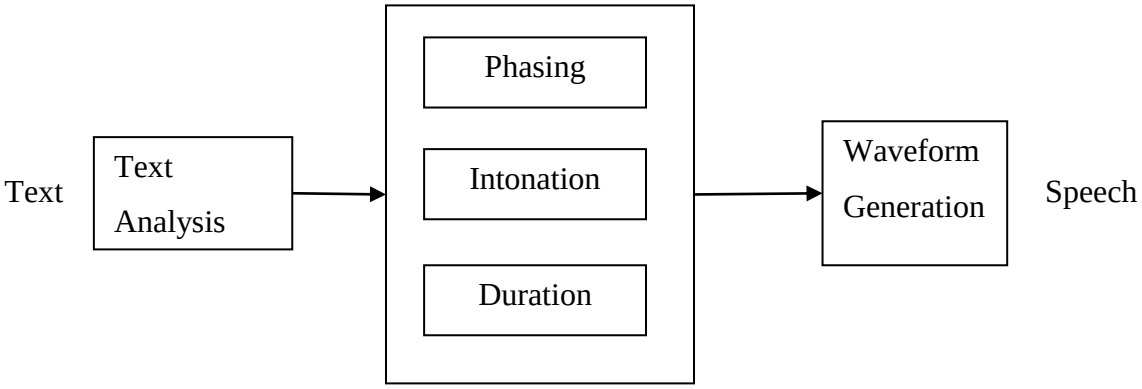

Figure 4.2 Typical TTS System

The typical component of the TTS system is shown in figure 4.2. It consists of text analysis, linguistic analysis and speech synthesis. The text first passes through the text analysis. The basic components of the current TTS system are as below and its architecture is shown in figure 4.3.

a. Text preprocessing
b. Text normalization
c. Letter to sound conversion

d. Speech database

e. Waveform synthesis

f. Synthesized speech signal

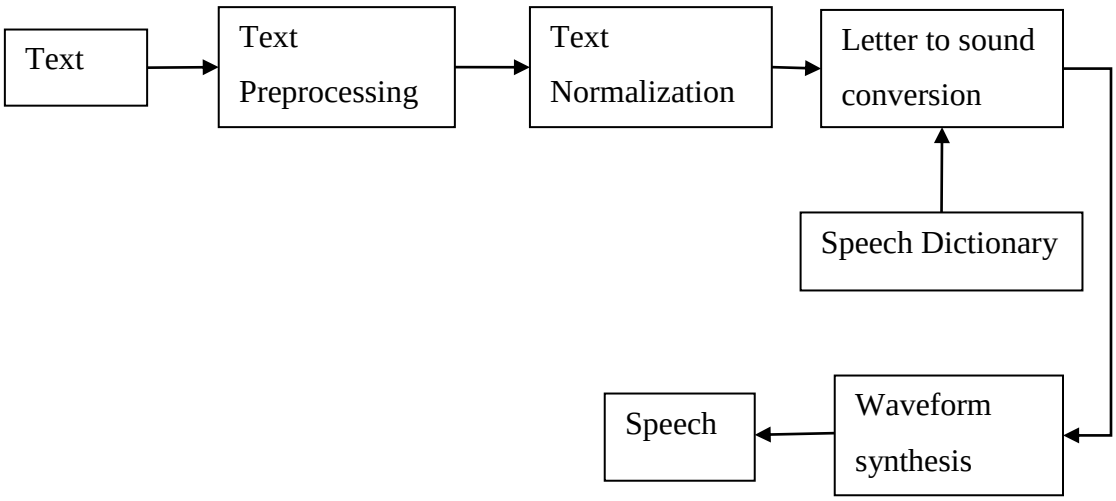

Figure 4.3 Block Representation of TTS

The components in TTS system can be vaguely further separated as

1. High level synthesis
    a. Text preprocessing
    b. Text normalization
    c. Phoneme representation
2. Low level synthesis
    a. Letter to sound conversion
    b. Waveform generation
    c. Synthetic speech signal generation

4.1 High Level Synthesis

This is mainly natural language processing which deals with the text analysis and processing which are explained below.

4.1.1 Text preprocessing and normalization

The Nepali text is the input to the front end to TTS. The input is written in UNICODE format. The text preprocessing converts input data to proper format for synthesizer. So each letter is fed as input using its corresponding representation in English. The text undergoes the preprocessing in which the character encoding issues and possible multilingual issues are identifies. The input text may contain numbers, abbreviations, years, dates etc. The abbreviations is distinguished by the '.' Present after the character. These characters need to be represented in standard form in order for their correct pronunciation. Hence the text normalization needs to be done. The abbreviation is converted to its full form; the numerical data may be either spelled out. The text normalization may be done using lookup technique.

4.1.2 Phoneme Representation

The texts are separated in single characters. Each letter is mapped to corresponding phonemes representation. The table 1 below shows the phonetic representation for Nepali characters. Similarly in table 2 the phonetic representations for Nepali vyanjans are shown. Such representation is used for mapping Nepali characters to corresponding phonemes.

Table 4.1 Phonetic Representation of Nepali Vowels

SVARA		REPRESENTATION
आ	ा	a
इ	ि	i
ई	ी	ii
उ	ु	u
ऊ	ू	uu
ए	े	e
ऐ	ै	ai

ओ	ो	o
औ	ौ	ou
अं	ं	am
अः	ः	aha

Table 4.2 Phonetic Representation of Nepali Vyanjans

VYANJANA	REPRESENTATION
क	k
ख	kh
ग	g
घ	gh
ङ	ng
च	ch
छ	chh
ज	j
झ	jh
ञ	yn
ट	tt
ठ	tth
ड	dd
ढ	ddh
ण	nd
त	t
थ	th
द	d

VYANJANA	REPRESENTATION
ध	dh
न	n
प	p
फ	ph
ब	b
भ	bh
म	m
य	y
र	r
ल	l
व	w
स	s
ष	skh
श	sh
ह	h
क्ष	khs
त्र	tr
ज्ञ	gy

The normalized text undergoes the letter to sound conversion. Letter to sound conversion is done by mapping processing.

4.2 Low Level Synthesis

Low level synthesis deals with the generation of synthetic speech signal. This is mainly digital signal processing. This is important part of the TTS system as it generates the synthesized sound. The steps involved are explained.

4.2.1 Letter to sound conversion

a. Creating Diphone Dictionary

The main requirement for letter to sound of normalized text is the pre-recorded speech files. The pre-recorded files contain the basic speech segments of Nepali that may be word or sentences. The pre-recorded speech has to be manually prepared. The speech database has to be created manually from pre-recorded files. The possible number of diaphones depends on number of phonemes as:

No. of diaphones = (No. of Phones)^2

In Nepali language, number of phones = 36 consonants + 12 vowels which makes

No. of Phonemes = 48^2

= 2304

Diphone is a stable part of speech between two phones. Diphone between ‘m’ and ‘a’ is shown in illustration 4.4 below.

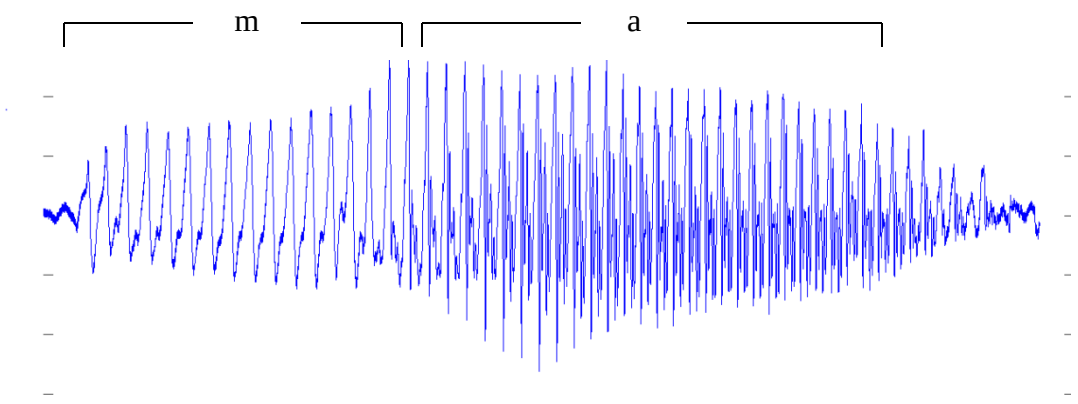

Figure 4.4 Diphone Between m-a

The diphone units with average pitch and duration values should be picked as it minimize the average amount of change.

b.Obtaining Diphone from Dictionary

The suitably recorded and analyzed diphone set is prepared which is used to concatenate. The speech database consists of the diphones extracted from the pre-recorded sound files. The sound database consist of a diphone name of the form P1-P2, pre-recorded sound filename, start time, mid time and end time of the diphone. The speech database contains the following information

		Positions in millisecond		
phones	filename	start time	mid time	end time
m	T0001.wav	34908	35038	36627

The start time, mid time and end time has been manually marked from the pre-recorded sound file and all the required information are saved in the database. After the recognition of the diphones, the grapheme to phoneme conversion needs to be done. Since Nepali is the phonetically rich language the written words are exactly pronounced as they are written. A basic dictionary based grapheme to phoneme has been implemented to retrieve the correct sound.

4.2.2 Waveform Generation

The corresponding diphone sounds are extracted from the speech database in letter to sound conversion. The signals are concatenated to generate the synthetic speech signal. The synthesized speech is generated based on diphonic concatenation in which the speech data are extracted for the pre-recorded sound file depending on the start and end marks of the diphone. The synthesized diphonic signal needs to be smoothening for the generated signal to be understandable hence the TD-PSOLA approach has been implemented. TD-PSOLA approach works with pitch synchronously with one window per pitch period. Pitch is correlated with dominant frequency of the speech signal called fundamental frequency. Hence it is very important to determine the pitch of each speech signal. To determine the pitch of the sound signal there are many pitch detection algorithm. There are many pitch detection algorithm (PDA) like cepstrum, linear prediction analysis, autocorrelation method. Autocorrelation method is the easiest method to determine the fundamental frequency of speech.

a. Pitch Detection Algorithm

Pitch detection is very important for speech processing and for prosodic variation in TTS. Pitch detection algorithm is a method to determine the fundamental frequency of the speech signal. The fundamental frequency mostly represented as f_0 is the main cue of pitch. There are many pitch detection methods namely zero-crossing, autocorrelation, cepstrum etc. Mostly autocorrelation is used for pitch detection as it is simple and efficient at mid to low frequencies. Autocorrelation can be used to determine the fundamental frequency in a signal. Mathematically autocorrelation can be represented as below in equation ii,

$$y(n) = \sum_{K=1}^{M} u(k)u(k+n) \qquad \text{ii}$$

The autocorrelation is the cross-correlation of a signal with itself. The autocorrelation of the signal can be seen in figure 4.5.

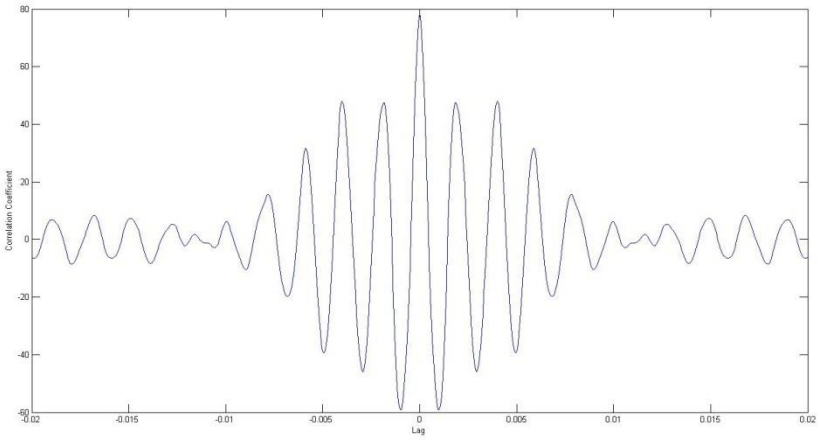

Figure 4.5 Autocorrelation of Signal

b.Windowing

The TD-PSOLA works in windowed data so the Hanning window has been applied centered at the pitch of the signal with 50% overlapping. Mathematically Hanning window can represented as in equation iii,

$$W(n) = 0.5 - 0.5\mathrm{Cos}\left(\frac{2\pi n}{N-1}\right) \qquad \text{iii}$$

The Hanning window can be seen in figure 4.6 below.

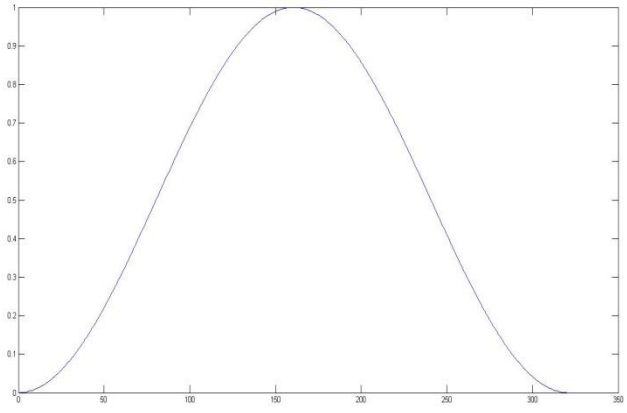

Figure 4.6 Hanning Window

Speech signal is separated into frames of speech signal using Hanning window. If the short term signal be xm(n) which is obtained from digital speech waveform x(n) by multiplying the signal by a sequence of pitch synchronous analysis Hanning window hm(n), then it can be represented as in in equation iv below,

$$x_m(n)=h_m(t_m-n)\, x(n) \qquad \text{iv}$$

The windowed sample signal for frame length of 320 can be seen in figure 4.7 below.

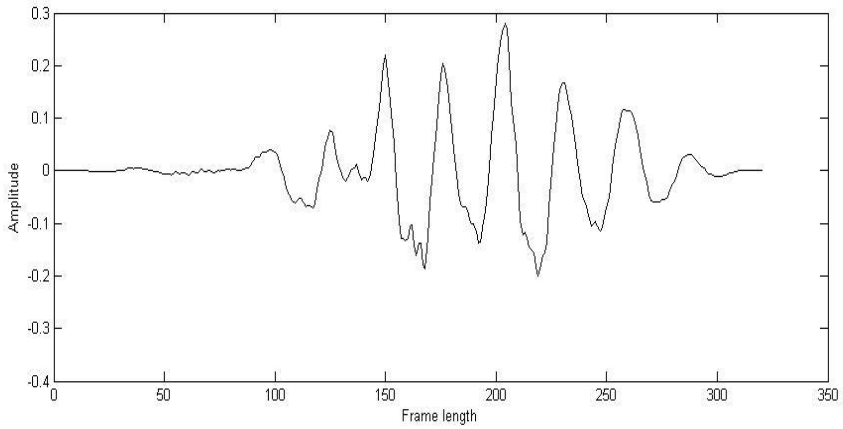

Figure 4.7 Windowed Sample Signal

c. Time Domain Pitch Synchronous Overlap Add Method

As TD-PSOLA name suggest it involves overlapping and adding windowed speech signals to modify frequency and timing of the speech signal. The windowed speech data has been overlapped and added to generate the synthesize speech signal and with a required pause in between to produce synthetic speech. It is windowed by a window length 2T i.e. it uses 50% overlap to achieve original duration of speech. The windowed signals has only been overlap add thus lengthening the signal which can be seen in figure 4.8. The use of greater overlap provides the increase of the pitch and the length of the signal decrease whereas to decrease pitch the overlap is small and length of the signal increases. However if the signals are omitted there may be loss of speech information.

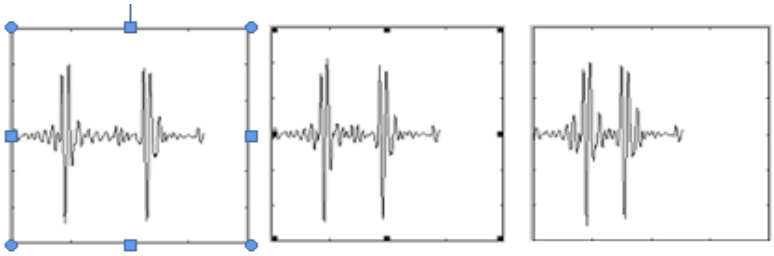

Figure 4.8 Sample of Overlapped Signal

Similarly the time scale modification can be achieved by either eliminating or add the frame of the speech signal. This brings no change in the frequency but the timing of speech signal is changed. If the duration is elonged there may be echo effect in the speech signal.

CHAPTER FIVE: RESULT AND DISCUSSION

Since the concatenation method for speech synthesis is the simple and easy to implement hence to implement this approach the pre-recorded database is required. The pre-recorded speech signal in male voice has been created. The recorded files are in .wav format. Currently only five pre-recorded wave files has been used namely T0001.wav, T0002.wav, T0003.wav, T0004.wav and T0005.wav. These files have been recorded in with sample rate of 16000 with 16 bits per sample. The files contain few sentences. From these pre-recorded files the diphones has been extracted. The speech dictionary has been prepared from these pre-recorded files. The speech database contains the following information

		Positions in millisecond		
phones	filename	start time	mid time	end time
m	T0001.wav	34908	35038	36627

The text in UNICODE characters is fed as input to the current system which has been shown in figure 5.1.

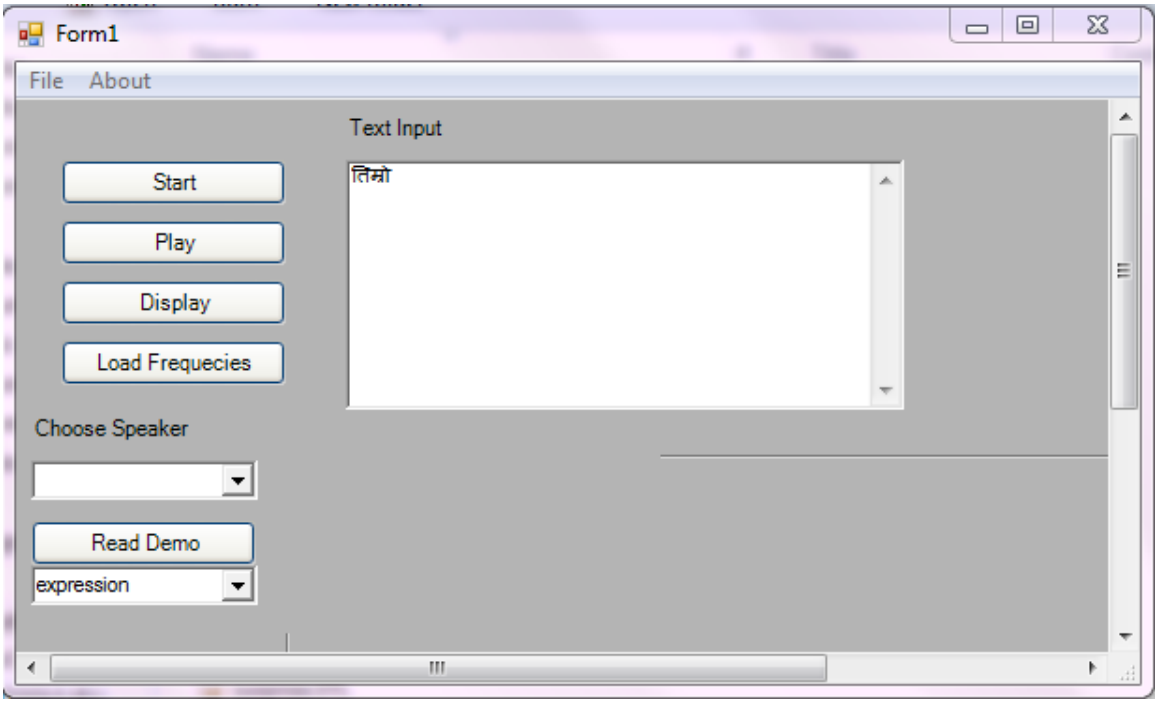

Figure 5.1 Application Form

When the input is fed into the system the text undergoes the text pre-processing, text normalization, letter to sound conversion and synthetic speech signal is formed. Each input text is separated into characters. If the text contains non standard texts like date, numbers, abbreviations such text should be represented in standard form i.e. it should be expanded. For each character its corresponding phonetic representation is done by mapping from table 1 and table 2. For each phonetic representation the corresponding speech segment is extracted from the pre-recorded speech files using the speech dictionary. The pitch of the speech signal has been determined using autocorrelation. Since autocorrelation has multiple multiplications the simple way to determine the autocorrelation is using FFT and IFFT. The FFT of the speech signal is determined of the power spectrum is determined and then the IFFT is done to determine the autocorrelation of the speech signal. The speech frame is generated by multiplying the speech signal by Hanning window. In pitch synchronous method it is important to determine the pitch of the signal. The pitch is the fundamental frequency of the speech signal. These extracted frame segments are recombined overlapped and added to generate the synthetic speech signal. The generated synthetic speech signal is then saved as .wav file for analysis. The generated synthetic speech signal is identical to the original speech signal.

5.1 Output Analysis

The system has been test with different word and sentences. Few examples are shown below.

a. For single character

For character का the English representation is ka whereas the required diphone set is

(-,k)+(k,a)+(a,-)

Hence for the required number of diphone set are 3. The character '-' at the beginning and at the end represents the blank space. The sample of windowed signal of frame length

320 obtained by multiplying the speech segment by Hanning window is shown in figure 5.2 below.

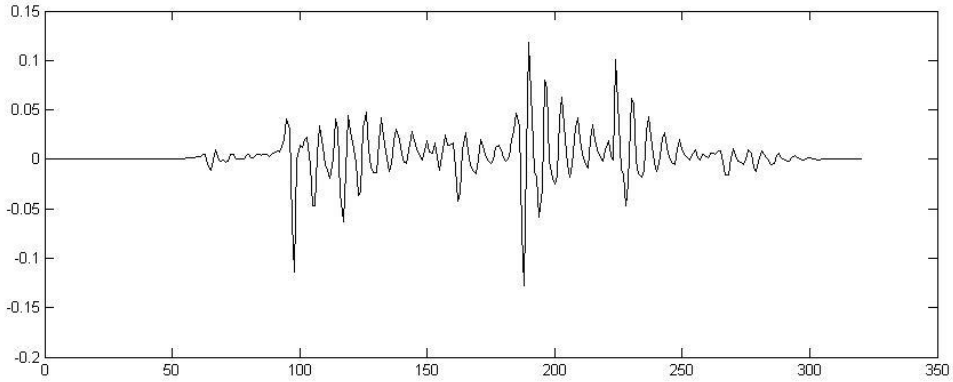

Figure 5.2 Extracted Sample of Windowed Signal of a Character

The synthetic speech signal for the character का is shown in figure 5.2 below.

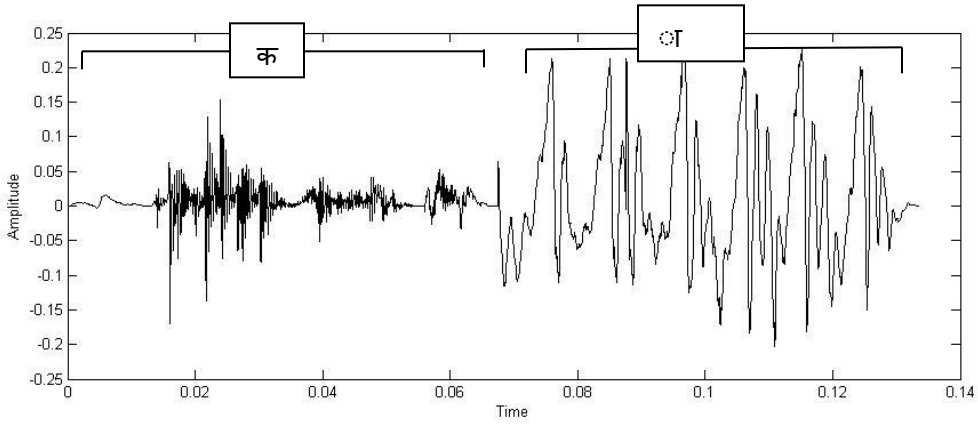

Figure 5.3 Waveform of Synthetic Speech Signal of का

b. For Words

Similarly for input word तिम्रो the word is broken down as below

ि – त – म् - र–ो

The phonetic representation of the word is

t-i-m-r-o

But, since diphonic concatenation has been implemented the diphone sets are

(-t) + (ti) + (im) + (m r) + (ro) + (o-)

where '-' represent the blank or space. Depending on these diphone set, the diphone segments are extracted from speech database which are processed for speech synthesis. So synthesizing the word तिम्रो it requires 6 diphones. For each speech segments the pitch detection is an important step. The figure 5.4 shows the pitch-mark plot.

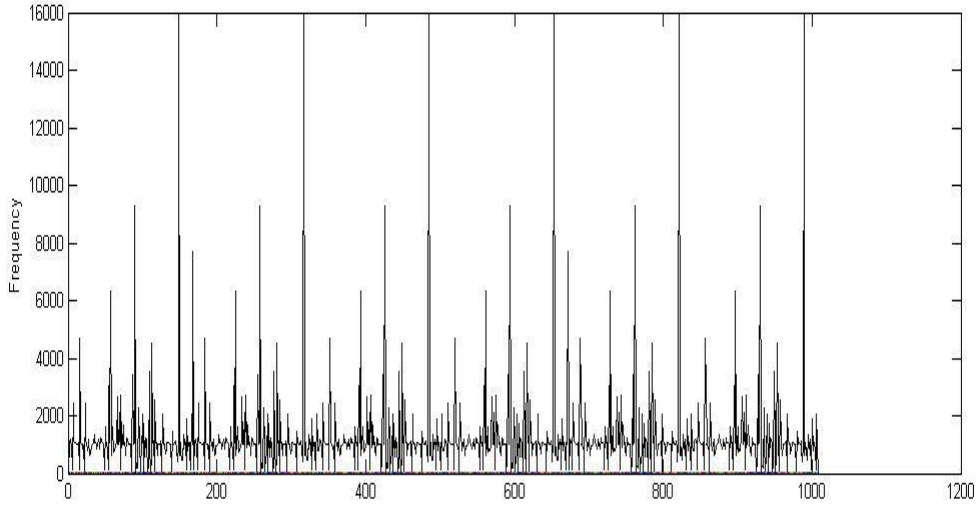

Figure 5.4 Pitch-mark Plot

Similarly the sample of windowed signal of frame length 160 is shown in figure 5.5.

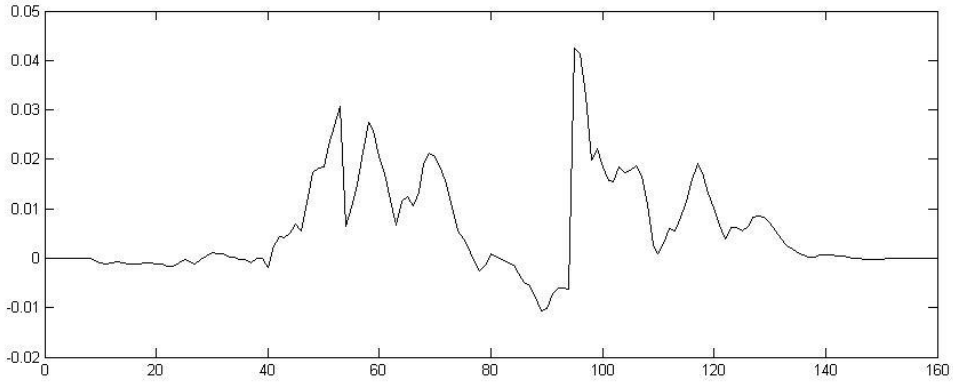

Figure 5.5 Extracted Sample of Windowed Signal of Word

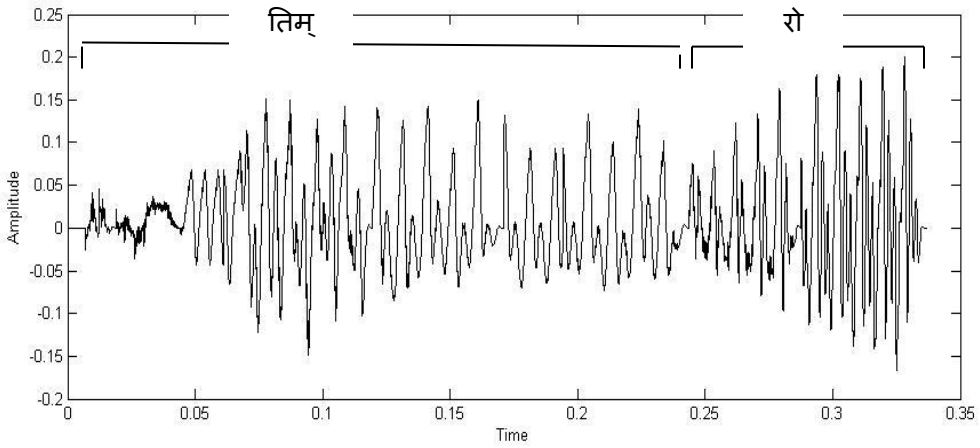

Figure 5.6 Waveform of Synthetic Speech Signal of तिम्रो

The waveform of synthetic speech signal of the word तिम्रो is shown in figure 5.6.

Similarly for input word कमल the English representation is k-m-l whereas the required diphone set is,

(-k) + (kae) + (aem) + (mae) + (ael) + (lae)

The number of required diphone set is 6. Each Nepali vyanjan ends with the अ sound so 'ae' has been added. Considering the frame duration of 0.02 the frame length will be 320 hence the windowed signal can be seen in figure 5.7

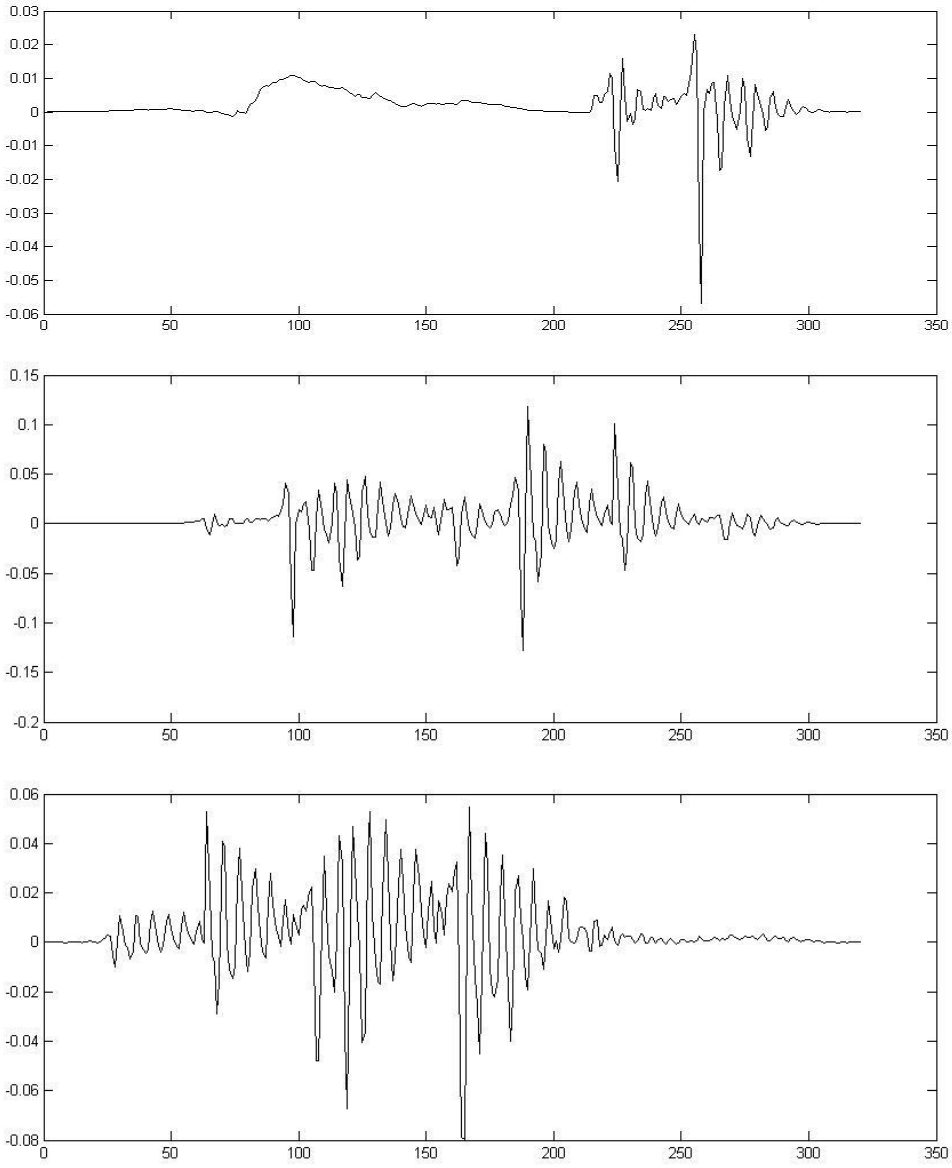

Figure 5.7 Sample of Windowed Signal

Such windowed signals centered at the pitch of the speech signal are recombined overlapped and added to generate the synthetic speech signal. The synthetic speech signal is shown in figure 5.8.

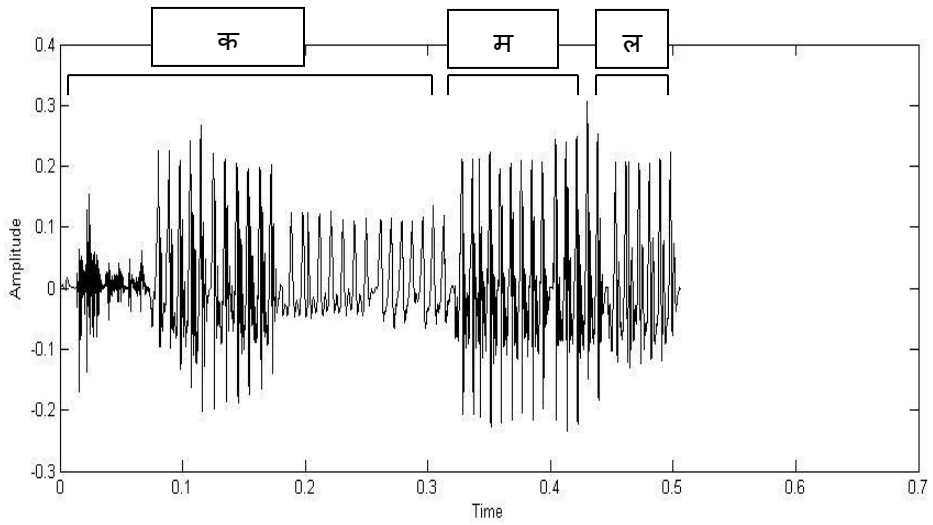

Figure 5.8 Waveform of Synthetic Speech Signal of कमल

c. For sentences

For Sentence I,

मेरो नाम सिखा हो।

Character wise representation,

म े र ो न ा म ि स ख ा ह ो

Required diphone sets,

(-m) + (me) + (er) + (ro) + (o-) + (-n) + (na) + (am) + (mae-) + (-s) + (si) + (ikh) + (kha) + (aa) + (-h) + (ho) + (o-)

So for the given sentence 17 set of diphones are required. '-' in front and at the last of characters represent the space for example here (o-)+(-n) it represents a space in a sentence. The synthetic speech signal for the sentence is shown in figure 5.8.

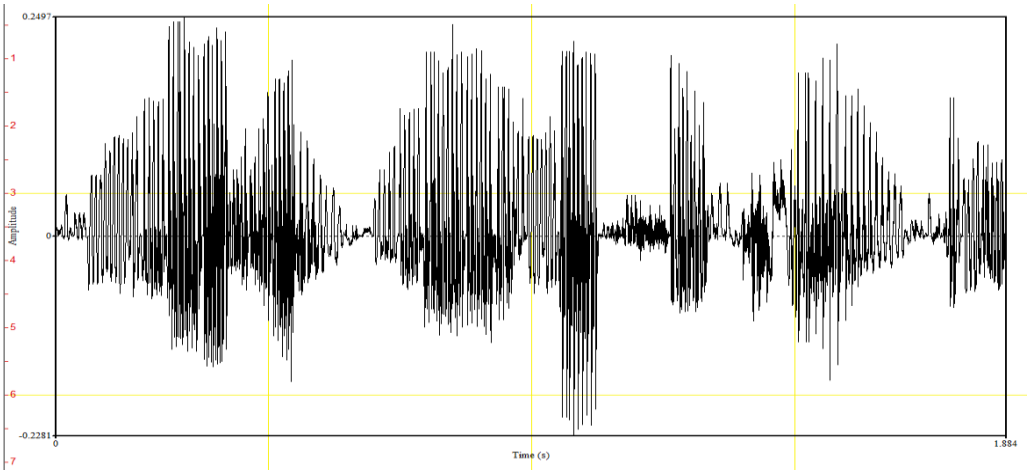

Figure 5.9 Waveform of Synthetic Speech Signal for Sentence I

Similarly for Sentence II,

चर्को घाममा हातमा हलो बोकी किसान खेत तिर गयो।

The character wise representation is

च र् क ो घ ा म म ा ह ा तम ा ह ल ो ब ो क ी क ि स ा न ख े त त ि र ग य ो ।

The required diphone sets are,

(-ch) + (chae) + (aer) + (rk) + (ko) + (o-) + (-gh) + (gha) + (am) + (mae) + (aem) + (ma) + (a-) + (-h) + (ha) + (at) + (tae) + (aem) + (ma) + (a-) + (-h) + (hae) + (ael) + (lo) + (o-) + (-b) + (bo) + (ok) + (kii) + (ii-) + (-k) + (ki) + (is) + (sa) + (an) + (nae) + (-kh) + (khe) + (et) + (tae) + (-t) + (ti) + (ir) + (rae) + (-g) + (gae) + (aey) + (yo) + (o-)

The synthetic speech waveform is shown in figure 5.9.

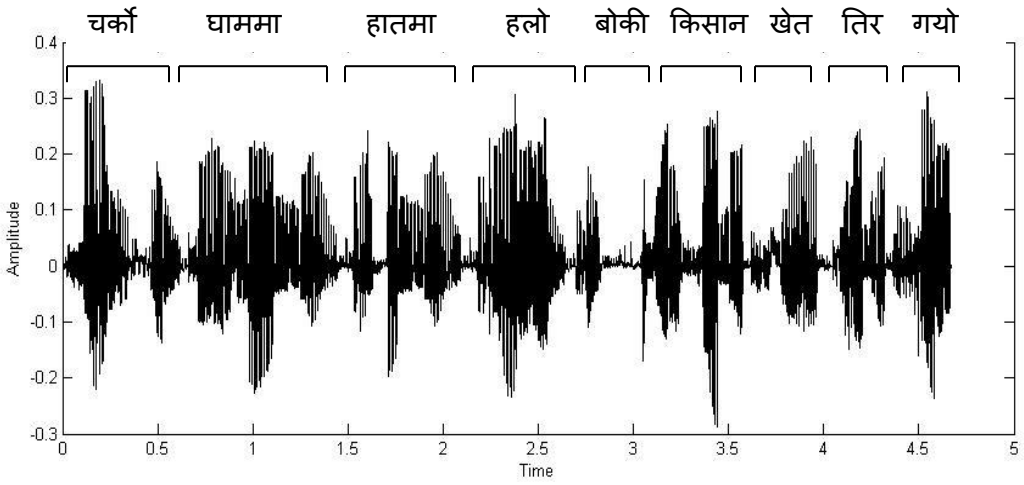

Figure 5.10 Waveform of Synthetic Speech signal for Sentence II

d. For number and abbreviation

For number inputs, for example ७ the input is matched with UNICODE character and the on the basis of mapping the standard text is determined and is pronounced in expanded form. The numerals are expanded to सात form for which the required diphone sets are

(-s) + (sa) + (at) + (tae)

Hence the synthetic speech signal can be seen in figure 5.11 below.

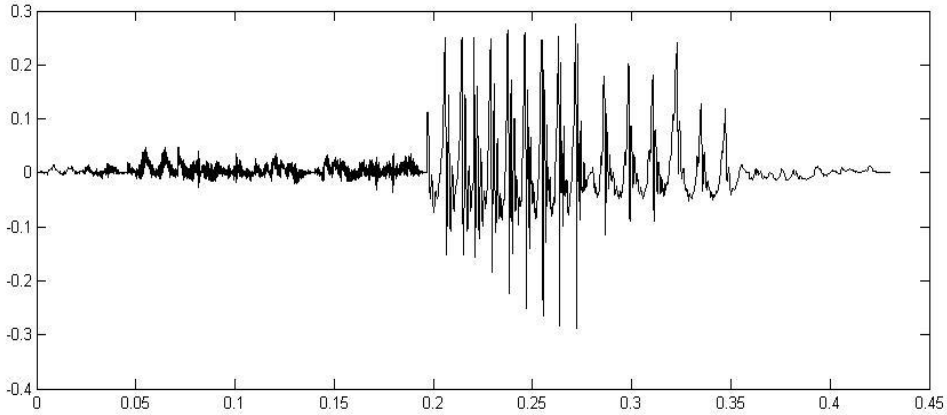

Figure 5.11 Waveform for numeral ७

When input is fed as an abbreviation like बि.सं., the input is distinguished as abbreviation by the '.' present at the end of the character. For abbreviation the direct text बिक्रम संबत is determined by direct mapping. For the word बिक्रम संबत the diphone set is determined as (-b) + (bi) + (ik) + (kr) + (rae) + (aem) + (mae) + (-s) + (sam) + (amb) + (bae) + (aet) + (tae)

The synthetic speech signal is generated which is unclear for the later part of the word. The synthetic speech signal waveform is shown in figure 5.12.

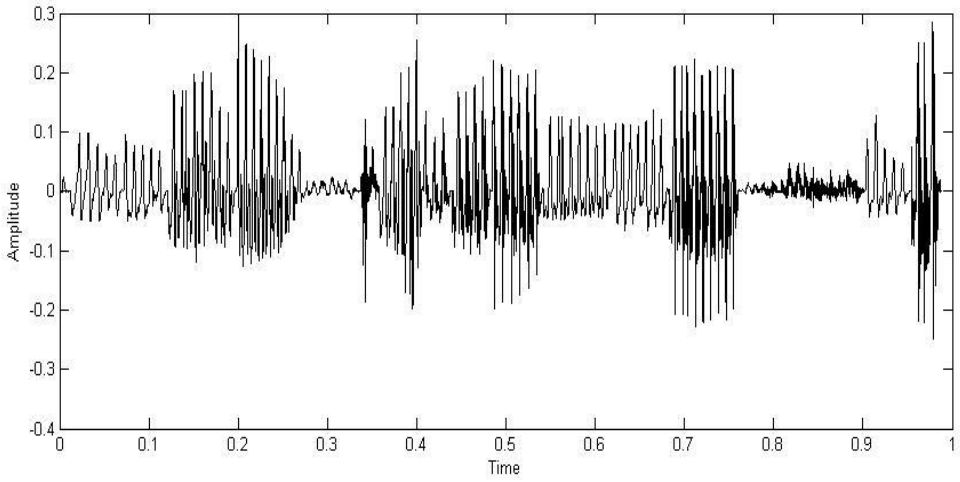

Figure 5.12 Synthetic Speech Signal Waveform of बि.सं.

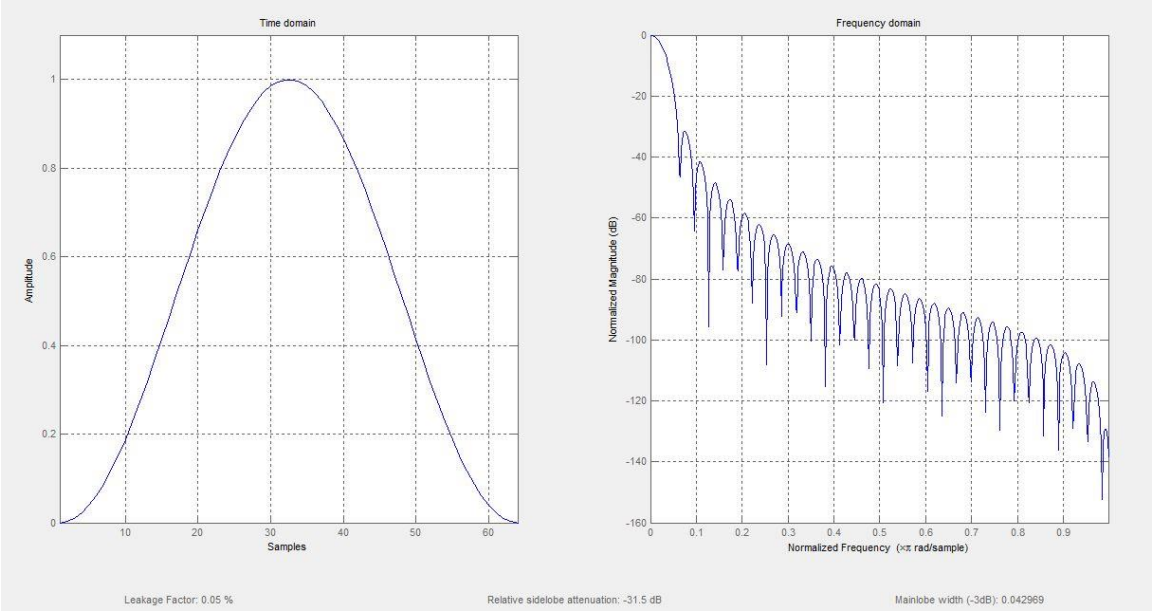

Figure 5.13 Hanning Window

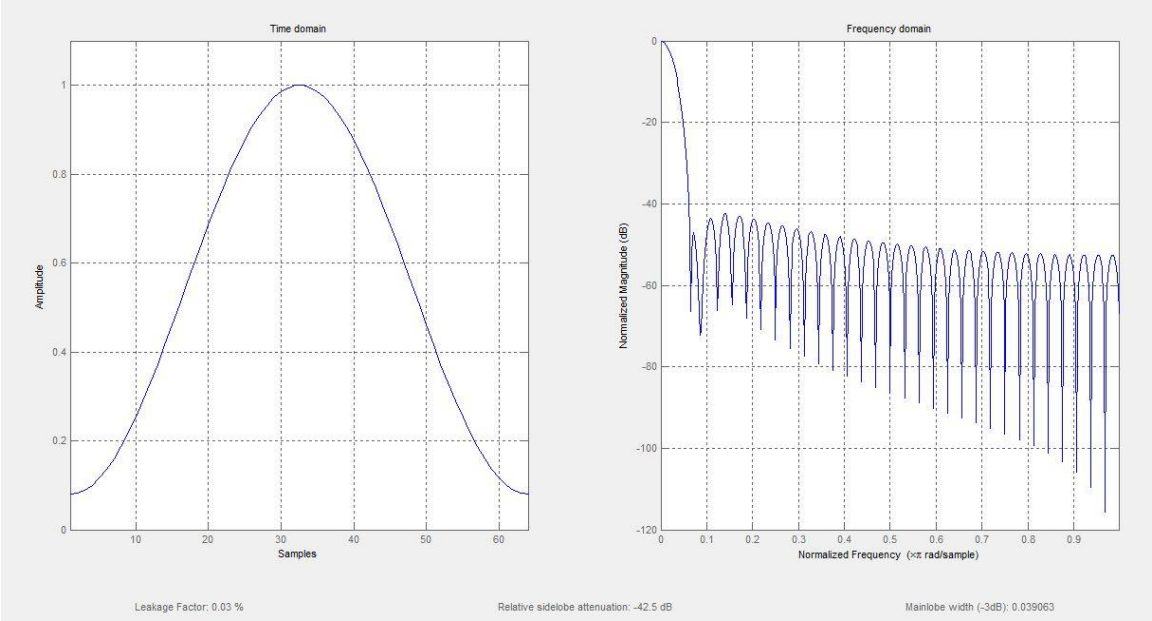

Figure 5.14 Hamming Window

5.2 Analysis on Hanning and Hamming Window

The Hanning and Hamming window both belongs to same family known as raised cosine. For windowing Hanning window has been used. Similar result has been obtained when

Hamming window was applied. To generate the frame of speech signal both the Hanning and Hamming window has been used. Hanning window is a smooth amplitude weight of a time signal that is zero at the beginning and the end of the time record which can be seen in figure 5.13 which has frame length of 160. Hamming window does not get as close to zero near the edges as does the Hanning window which can be seen in figure 5.14. The general mathematical representation is shown in equation v,

$$w(n) = \alpha - \beta \cos\left(\frac{2\pi n}{N-1}\right) \qquad v$$

Here for Hanning window, $\beta = \alpha = 0.5$ whereas for Hamming window $\beta = 0.46$ $\alpha = 0.54$.

There was no difference in synthetic speech signal generated using both Hanning and Hamming window in time domain. The frequency domain graph for both Hanning and Hamming window is shown in figure 5.13 and 5.14. In Hamming window the first side lobes -42 dB, whereas the Hanning window's first side lobes are only -32 dB. Thus, the Hamming window has better selectivity for large signals, but it has the disadvantage of high side lobes due to slow roll-off rate. The first side lobe of the Hamming is than the first side lobe of the Hanning, but the distant side lobes of the Hanning are lower than the Hamming thus the Hanning is better. The Hanning window has frequency response which sharpens progressively unlikely in Hamming window which as sharp central band. The spectrum of the Hanning Windowed and Hamming Windowed signal has been shown in figure 5.15 and 5.16 respectively. From the spectrum also it can be seen that there is no significance difference in synthetic speech signal when Hanning and Hamming window applied to generate the frames of speech signal.

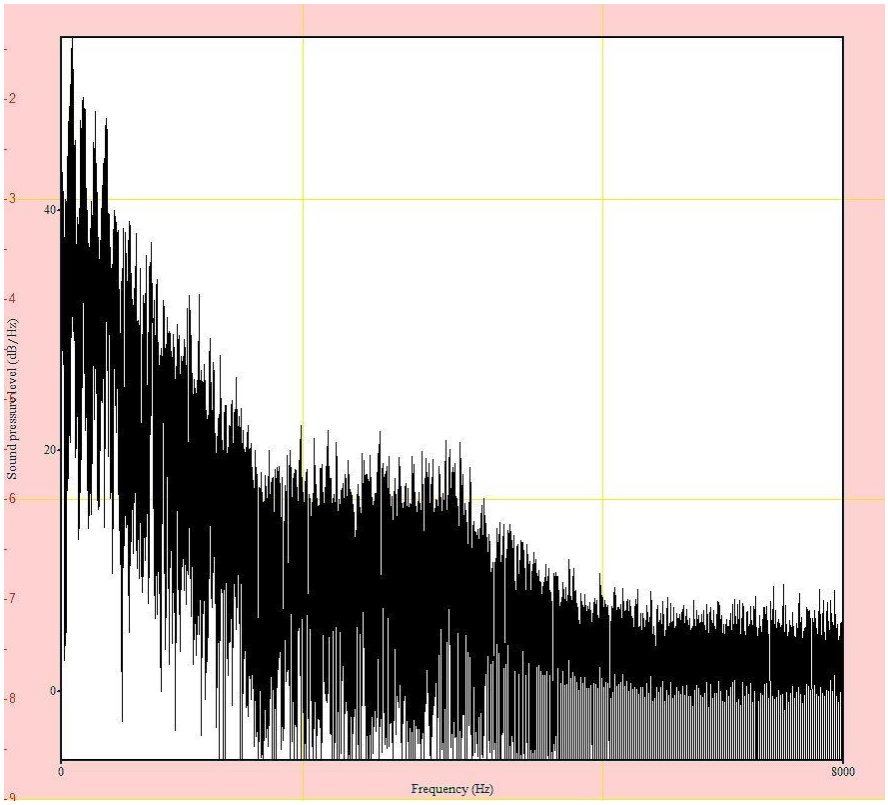

Figure 5.15 Spectrum of Hanning Windowed Signal

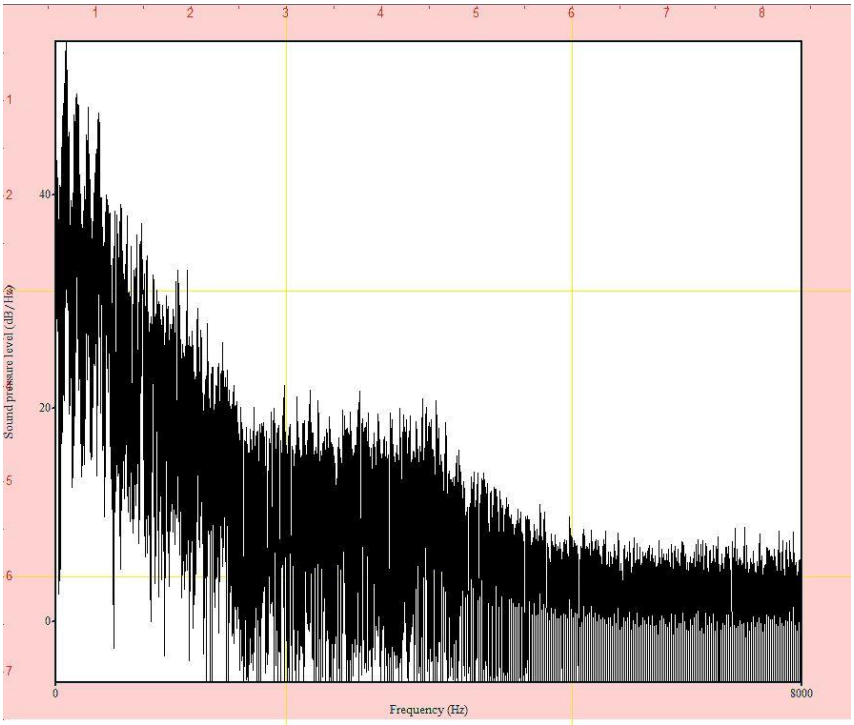

Figure 5.16 Spectrum of Hamming Windowed Signal

5.3 Comparison with Previous Work

B. Chettri, K. B. Shah [3] uses the ESNOLA method for Nepali text to speech system. In such method the speech signal dictionary was created by extracting from the pre-recorded speech signals such that the speech segment begins and ends all at the positive zero crossing so that there will be less distortion during concatenation. The system uses partneme as the unit of concatenation which makes the size of speech dictionary quite small. The nonsensical word of the form CVCVCV has been recorded from which the partneme is extracted as shown in figure 5.17.

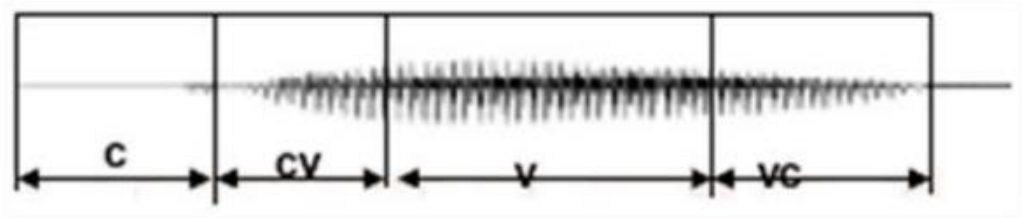

Figure 5.17 Speech Segment

Creating speech database from the nonsensical words is a difficult task as each transition from consonant to vowel should be properly distinguished and saved. Since this method is non-overlapping method spectral smoothing needs to be done at the concatenation point to eliminate disturbances. In TDPSOLA as name suggest it is the overlap adds method which involves the frame of the speech segments with the 50% of overlap for the natural and intelligible speech signal. Since in TTS using ESNOLA [3] there might be some disturbance in the concatenation point of the synthetic speech signal as the windowing is done in output signal. This method involves the tokens for distinguishing between consonant and vowels to extract the speech segments from the speech dictionary which is a tiresome task. It may not be possible to involve all the consonants vowel combination. In Nepali TTS using ESNOLA the Nepali text is first converted to some phonetic representation through a grapheme to phoneme conversion upon which the phonological rule base which contains linguistics rules are applied which can be seen in second column of table 5.1. The table shows the intermediate output from both the

system. In Nepali TTS using ESNOLA the token system has been implemented for the phonemes generated. Based on the tokens assigned the pre-recorded speech segments are extracted from the speech dictionary and are concatenated to get the synthetic speech signal. In Nepali TTS using TDPSOLA the input text are first separated into characters and the corresponding English representation is done. Based on the English representation the diphones sets are determined which can be seen in third column of table 5.1. Based on these diphone sets the pre-recorded speech segments are extracted and frames of speech signal is generated using Hanning window centered at the pitch of the speech signal. The grapheme to phoneme rule has not been completed implemented. Such frames of speech segments undergo concatenation to generate the synthetic speech signal. The use of frames of speech signal makes the synthetic speech much understandable and natural. Since frames are used the duplication/elimination or change in frequency of the frames to generate the synthetic speech signal is easy.

Table 5.1 Input Text Processing Comparison between TTS using ESNOLA and TTS using TDPSOLA

Sentences	TTS Using ESNOLA [3]	TTS Using TDPSOLA
मेरो नाम कमल हो	Grapheme to phoneme conversion, ME3RO3 NA4M KML HO3/ After applying phonological rules, #MA2RO NA4M KAMAL HO/	Diphone sets, (-m)(me)(er)(ro)(o-)(-n)(na)(am)(mae)(-k)(kae)(aem)(mae)(ael)(lae)(-h)(ho)(o-)
मेरोमा १० रुपया छ	Grapheme to phoneme conversion, #ME3R03MA4 DS RB3PYA4 C1/ After applying phonological rules, #MA2ROMA4 DAS RUPYA4 C1A/	Diphone sets, (-m)(me)(er)(ro)(om)(ma)(a-)(-d)(dae)(aesh)(shae)(-r)(ru)(up)(pae)(aey)(ya)(a-)(-chh)(chhae)

Similarly in Bhasasanchar [14] has implemented the concatenation method to develop Nepali TTS system. This system is based on festival system.

Table 5.2 Speech Database Comparison

	Bhasasanchar TTS System[14]	TTS using TDPSOLA
Total Words	13904	66
Most Occurring Vowel	अ	अ
File Format	.wav	.wav
Total Diphones	1247	350

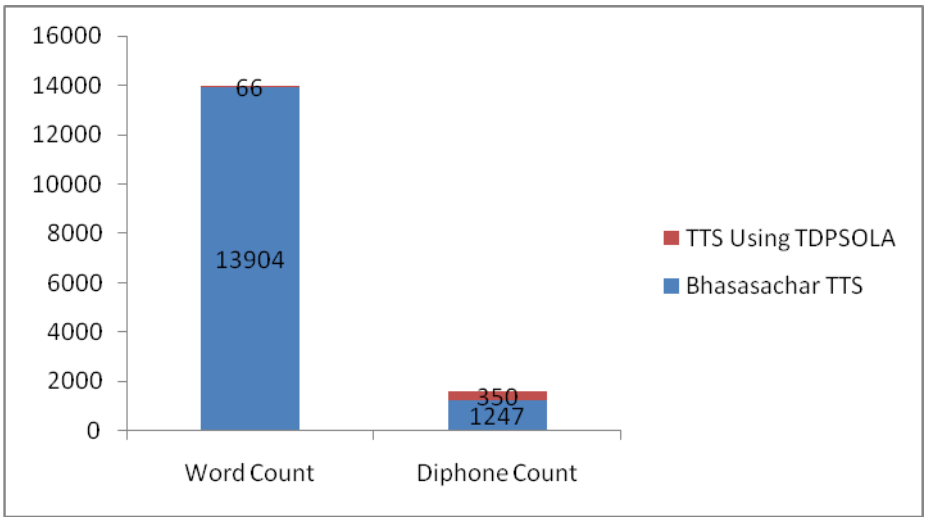

Figure 5.18 Graphical Representation of Table 5.2

The table 5.2 shows the speech database for the Bhasasanchar TTS system [14] and the current system with TDPSOLA. Even though the current speech database has only 66 words it is capable of converting 313 words from text to speech successfully. Depending on these words the diphones are extracted and concatenated to generate the synthetic speech signal. The Bhasasanchar TTS system [14] has the phones of 45045 and diphones of 1247 whereas the TTS System using TDPSOLA has the speech dictionary containing

only 350 diphones. These data has been represented graphically in figure 5.18. The successful conversion of word to speech for both systems is shown in table 5.3 below.

Table 5.3 Comparison of Successful Speech Conversion of Words

	Success	Accuracy Percentage
Bhasasanchar TTS System[14]	507	82.84
TTS System Using TDPSOLA	313	83.023

Even with limited number of diphones the current system gives 83% of accuracy, if the numbers of diphones are increased the accuracy would have been more.

CHAPTER SIX: CONCLUSION AND RECOMMENDATION

6.1 Conclusion

Many researches have been done on Nepali TTS system. Nepali language being a class of Sanskrit language it is difficult task to develop the Nepali text to speech system. For other language like English, Finnish etc. already many research have been done to develop the speech dictionary. Still much analysis is ongoing in Nepali language to develop morphological analyzer, stemmer and other systems. The efficient and easy method to develop the TTS system is concatenative method.

The proposed method of concatenation modifies both in pitch and time domain of the speech signal which produces the natural and understandable synthetic speech. The Hanning window used to separate the speech signal into frames is more efficient windowing method than Hamming window as it has sharp progressing frequency response. Both the windowing signal were used to generate the frames of speech signal and there was no certain difference in synthetic speech signal. 313 test words has been tested in current system giving 83% of accuracy.

6.2 Recommendation

Currently the thesis involves single male voice pre-recorded speech files. So as future enhancement the female voice pre-recorded speech files can be included so that the both the male and female synthetic speech signal can be generated. The speech dictionary does not consist of all the diphones so all the Nepali text is not converted into speech signal. The speech dictionary can be expanded by recording more words to include all the diphones. The text to speech for letters like र, ल, न are not clear whereas for letters like फ there is no diphone present in the speech database. Since the pre-recording sound signals are limited the characters like दु, गो etc cannot be read and converted to sound. Also the characters are not read as written if the character ends with अ sound so only the half sound of such character is read. The errors and limitation in TTS system is due to the limited sound library or the faulty recording.

The synthetic speech signal does not have the prosodic and intonation characteristics. The prosodic and intonation characteristics could be integrated in the TTS system. The prosodic feature could be added by manipulating the frequency of the speech signal depending on the expression and meaning of the sentence. But that would be a whole new topic to involve the prosodic feature in TTS system.

REFERENCE

1. M. Schroder, J. Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching", Institute of Phonetics, University of the Saarland, Saarbrucken, Germany
2. The Festival Speech Synthesis System, System Documentation Edition 1.4, June 1999
3. B. Chettri, K. B. Shah, "Nepali Text To Speech Synthesis System Using ESNOLA Method Of Concatenation", International Journal of Computer Application January 2013
4. M. Choudhury, "Rule Based Grapheme To Phoneme Mapping For Hindi Speech Synthesis", Department of Computer Science and Engineering, IIT
5. A. W. Black, K. A. Lenzo, "Multilingual Text To Speech Synthesis", Language Technologies Institute

6.D. Sasirekha, E. Chandra, "Text To Speech: A Simple Tutorial", IJSCE , March 2012

7.I. Lsewon, J. Oyelade, O. Oladipupo, "Design and Implementation Of Text To Speech Conversion For Visually Impaired People", IJAIS, April 2014

8.F. Ykhelf, L. Bendaouia, "Pitch Making Using The Fundamental Signal For Speech Modification Via TD-PSOLA", IEEE International Symposium On Multimedia, 2013

9.S. Padda, N. Nidhi, R. Kaur, "A Step Towards Making an Effective Text to Speech Conversion System", IJERA, March-April 2012

10. S. K. Thakur, K. J. Satao, "Study of Various Kinds of Speech Synthesizer Technologies and Expression for Expressive Text to Speech Conversion System"
11. T. K. Patra, B. Patra, P. Mohapatra, " Text to Speech Conversion with Phonematic Concatenation", IJECCT, September 2012
12. M. Ahmed, S. Nisar, "Text-to-Speech Synthesis using Phoneme Concatenation", International Journal of Scientific Engineering and Technology", February 2014
13. P. Taylor, "Text To Speech Synthesis"
14. Nepali TTS Bhasasansar Manual, March 2008

BIBLIOGRAPHY

1. Y. Wang, R. T. Tsai, “Rule-base Korean Grapheme To Phoneme Conversion Using Sound Patterns”
2. R.Sproat, J. Olive, “Text To Speech Synthesis”
3. A. Chauhan. V. Chauhan, S. P. Singh, A. K. Tomar, H. Chauhan, “A Text To Speech System For Hindi Using English Language”, IJCST, September 2011
4. The Festival Speech Synthesis System, System Documentation Edition 1.4, June 1999
5. S. Lemmetty, “Review of Speech Synthesis Technology”, March 1999