1. Problem Statement

Text to speech system is developing in world with its full pace. With recent advancement of technology, the difficulty to collect data has been largely reduced. In this context, powerful text to speech system has been developed for popularly used languages in the world like English, Chinese etc. But we still fail to get reliable text to speech system in Nepali languages. This might be due to lack of manpower, investment, data or other. If we don't start to work immediately on it, it might backward our languages in term of use of technology.

2. Introduction

2.1. Background

Audio Signal has always been a part of interest. It is the fundamental way of communication for human and other animals. The invention of magnetic speaker and tape recorder has its own importance because these inventions made possible to produce human-like sound from non-living source. If we go even before these inventions, there has been many attempts to create human-like sound from mechanical approach which were able to produce sounds of some vowels and consonants. Artificial Speech Synthesis is another link in the chain of human curiosity over audio signal. It has been an important part of Machine Learning and some amazing progress has been recorded in this field in first two decades of twenty-first century.

But speech is not the efficient form of communication in some context. For example, speech is difficult and costly to store (even impossible before a century). So, people tried to express the information of speech communication with the help of symbols which finally became the foundation of text communication.

Now, with the progress in computer technology, one of the important effort of today's generation is to fill the gap between these two forms of communication.

2.2. Description

The title of the project we completed as our minor project is: "**Nepali Audio Signal Synthesizer**". In this project, we have used audio signal generating algorithms and models to generate human-like voice for Nepali Language (Nepali Language written in Devanagari script). The input to the system is the meaningful sentences in text format and output the corresponding audio signal. This system has a frontend part made in HTML where the text to be converted can be input. The text is then analyzed, processed and then converted into human-like voice in backend program. The system has really simple and easy to use user interface and the conversion is almost in real-time.

2.3. Motivation

We live in the era of rapid development of communication all over the world. Thus, in this situation, there may not be a single motivational source, rather a whole aura of technological development acts as motivation. However, we are highly motivated by the work of different researchers of Nepal on Nepali Natural Language Processing like Dr. Bal Krisha Bal, Kathmandu University, Dr. Basanta Joshi, I.O.E, Tribhuvan University and many more. Virtual Assistance Technology like Siri, Cortana has always amazed us. Humanoid Robot "Sophia" made by Hanson Robotics Ltd. is another product that fascinated us. Food serving robot "Ginger" made by Paaila Technology ignited fire inside us to start this project.

2.4. Objectives

The major objectives for doing this project are as follow:

- To learn more in the field of Signal Processing and Artificial Intelligence with "learning by doing" approach.
- To contribute in the field of Nepali Natural Language Processing.
- To learn practical knowledge on team work, project management, knowledge implementation and presentation.
- To implement the knowledge that we have learned till now to make some meaningful product.
- To fulfill the course requirement of doing minor project in third year second part of Electronics and Communication Engineering.

2.5. Application

The project we have done can have large area of application. Some of them are as follow:

- It can be used as assistance technology for verbally impaired people.
- It can be used for giving humanoid ability to robots.
- It can be used for making voice bots which can serve people in many field like hospitals, government offices etc.
- It may change the way we produce audiobooks, radio programs and many more.

3. Literature Review

As it has already been mentioned, the widely used languages of world like English and Mandarin Chinese have powerful text to speech system. Many technological company are providing these services. There has been experiments with many techniques for it including concatenative model, parametric model, generative model etc. Google has launched WaveNet technology in 2017 which is currently considered as the most advance state of art technique for speech generation.

But if we consider particularly to Nepali languages, we are way behind. At present, there has been some researches in this field by various Nepali as well as foreign universities. Yet no any meaningful product has been launched till date. We can find some project in internet with partial success in this field. One of the noticeable service is google translate developed by Google which provides medium to synthesize Nepali audio signals from text. Still, the amount to research done in this field, particularly in Nepali language, is in nascent state and thus the available products have non-humanly sound.

Some of the tech companies of Nepal have started to do research in this field. But we still don't have satisfactory output released in public domain. It can be expected that such system might be released in near future for public use, but still we are not in the state to completely rely on them.

4. System Details

4.1. Theoretical Background

Fundamental knowledge of Speech communication

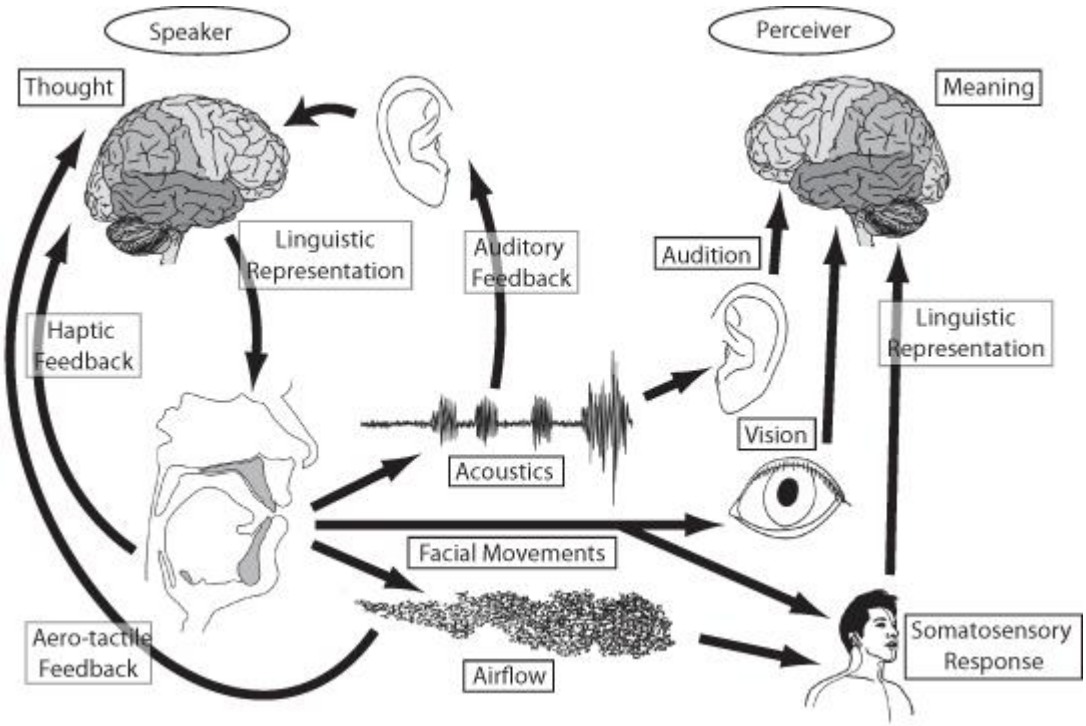

Figure 1: Basic flow diagram of speech communication

Whenever we are having speech communication, what we basically doing is creating different pressure patterns which travels in air and hit the eardrum of listener. Then this pressure fluctuation is understood by listener's brain.

Fundamental knowledge of Text communication

In text communication, each word in speech is represented by unique (exception exists) pattern of symbols. These symbols are known as alphabets. Each meaningful word in text has corresponding speech signal.

Types of currently available Text to speech techniques:

- Concatenative model
- Parametric model
- Generative model

Because Parametric and Generative model need a large set of data to produce satisfactory result, so we chose concatenative model for our project.

Concatenative synthesis is a technique for synthesizing sounds by concatenating short samples of recorded sound (called units). The duration of the units is not strictly defined and may vary according to the implementation, roughly in the range of 10 milliseconds up to 1 second. It is used in speech synthesis and music sound synthesis to generate user-specified sequences of sound from a database built from recordings of other sequences.

4.2.Methodology

Project is done in following sequence:

1. Collection of Data

   For the concatenative model we have used, we need a collection of waveform of all possible phonetic sounds. This was done manually by recording each phonetic sound and removing the unwanted part of it.

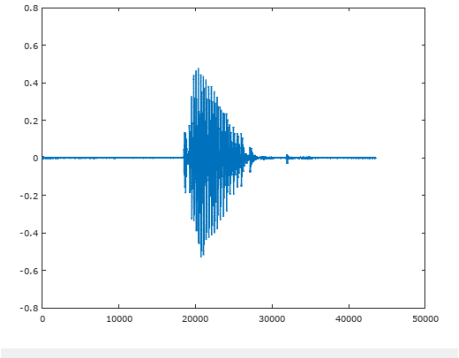

Figure 2: Waveform while recording 'क'

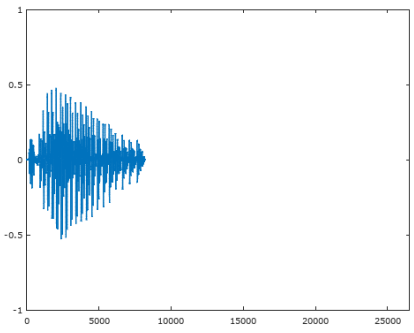

Figure 3: Waveform after removing silent part

The specification of recorded audio is:

- Bits Per Sample = 16
- Number Of Channels = 1 (Mono)
- Sample Rate = 44100 Hz

2. Developing of Model

Our model first takes text input.

For example: पुल्चोक $ *

Then it removes unknown character comparing with its vocabulary.

Then we get the output as: पुल्चोक

After that, it divides the text into phonetic symbols:

पुल्चोक = पु + ल् + चो +क

This part is easier than we thought, since devnagari alphabate is itself in phonetic in nature. So, our job is just to find the combination of character making a single phonetic sound.

Then we concatenated waveform for each phonetic sound and applied filter on it. Then finally we played it.

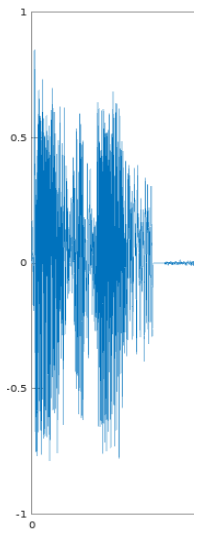

Figure 4: output waveform

The tools we used are as follow:

- Audacity software for sound recording
- Octave for silent part clipping
- HTML and Django server to take text input and pass it to main system.
- Python version 3.6 for text analysis and audio concatenation
    - Numpy
    - Scipy
    - Sounddevice
    - Matplotlib

4.3.Block Diagram

In the simple level, our system works in following steps:

1. Take text input
2. Checks if it has unknown character. If yes, remove it.
3. Then divide text into respective phonetic encoding.
4. Then load corresponding phonetic waveform in the required order.
5. Concatenate them.
6. Apply filter to make them smooth.
7. Play the audio.

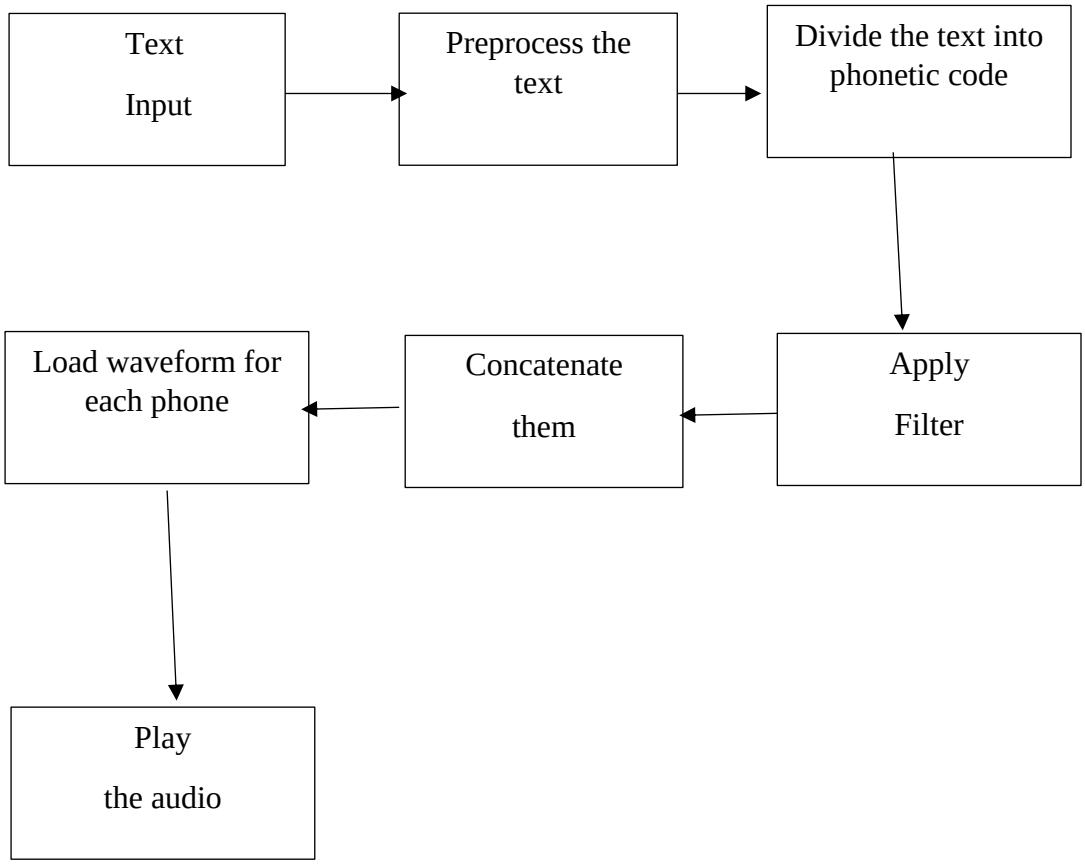

Figure 5: Block Diagram of the system

5. Output

We gave Nepali Text as input to the system. The system played corresponding speech signal.

The input interface looks like below:

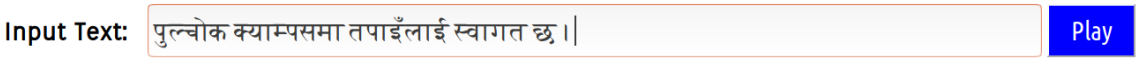

Figure 6: Input Interface

Input Text: पुल्चोक क्याम्पसमा तपाइँलाई स्वागत छ।

Output waveform:

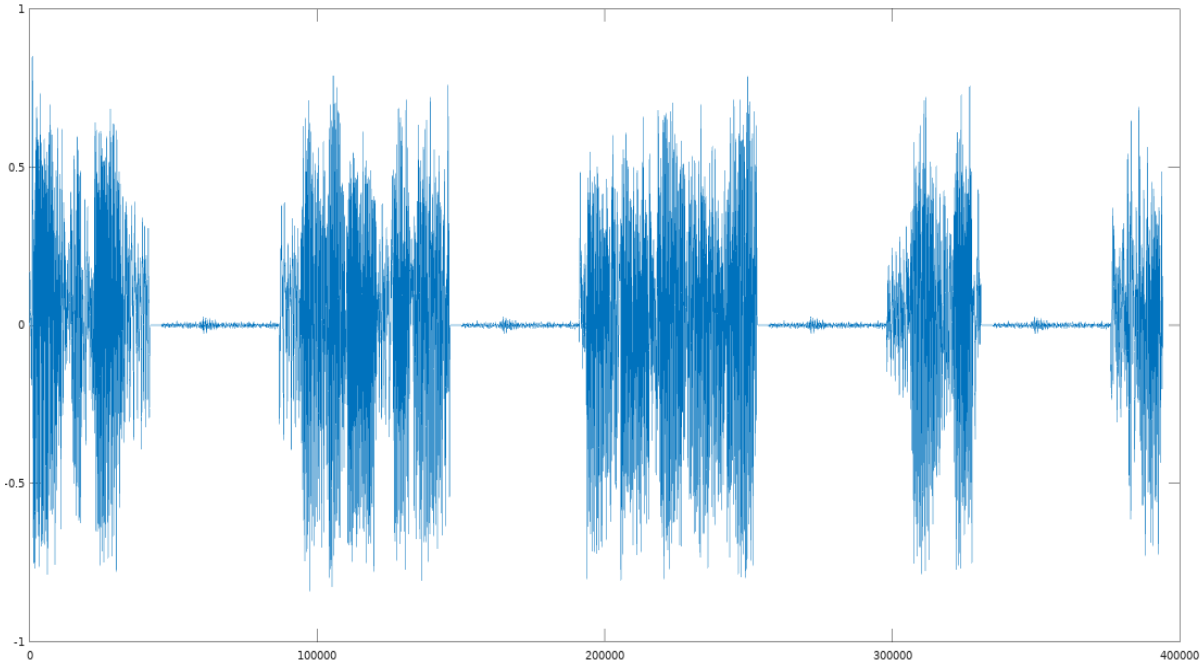

Figure 7: Output waveform for पुल्चोक क्याम्पसमा तपाइँलाई स्वागत छ।

Input Text: फ्रान्स विश्वकप बिजेता हो।

Output Waveform:

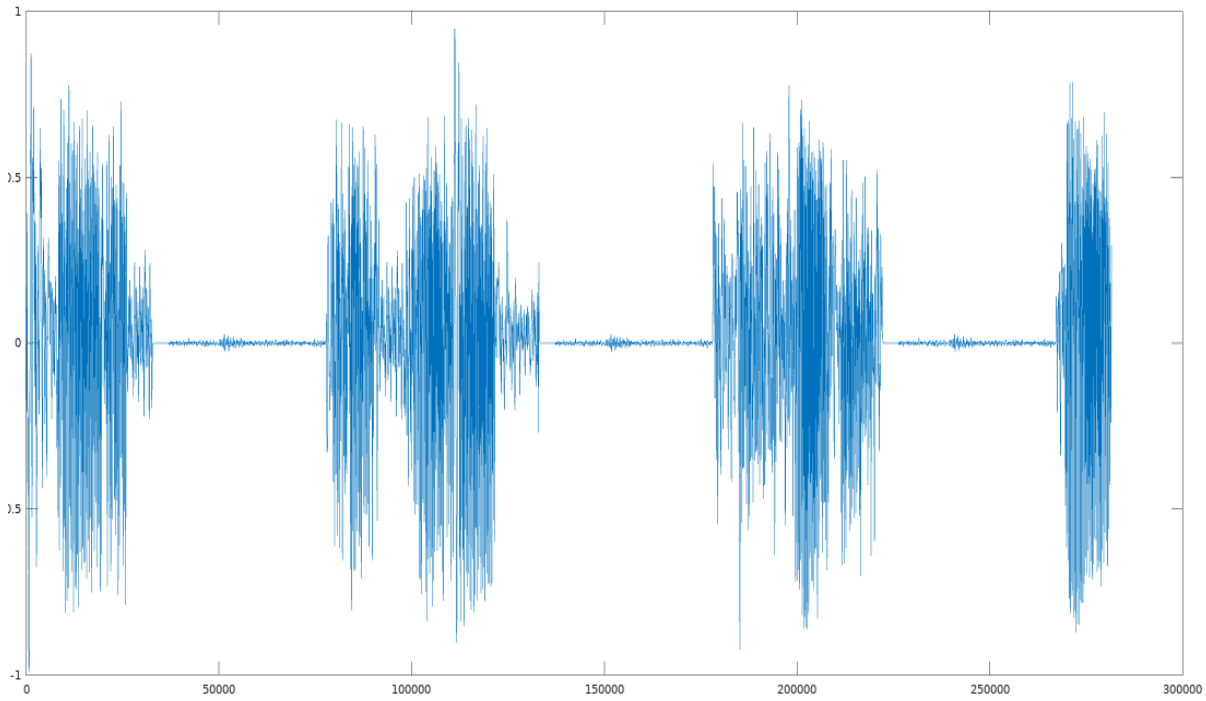

Figure 8: Output waveform for फ्रान्स विश्वकप बिजेता हो।

6. Budget Analysis

The approximate expenditure of project till date is as follow:

Table 1: Expenditure Analysis

S.N.	Expenditure	Amount (in R.S)
1.	Microphone	1,000
2.	Research and Development	1,000
3.	Publication	500
4.	Miscellaneous	500
	Total	3000

The total expenditure is divided equally to each member of project.

The expenditure is highly reduced due to following reason:

- Use of Open source tools and library like Octave, Audacity, Scipy, Numpy etc.
- Maximum use of already available resources like speakers, laptops etc.
- Use of social media to reach seniors and ask for help.

7. SWOT Analysis

Strength:

- The system can produce satisfactory speech.
- It can handle many unexpected conditions.

Weakness:

- The output is still insufficient to meet present standard to state of art text to speech system.
- System still cannot produce sounds corresponding to Nepali symbols like chandrabindu (चन्द्रबिन्दु) and sirbindu (शिरबिन्दु).

Opportunity:

- The project has a lot of way open for further development.
- With enough data available, the current system can also do better.

Threats:

- The model we have adapted is outdated in present world.

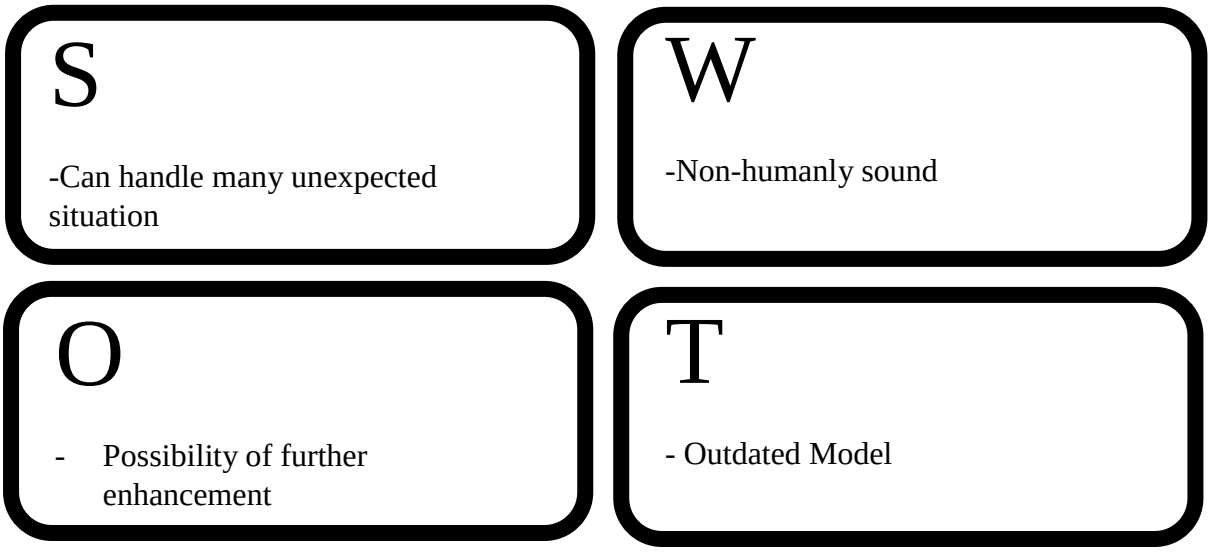

Figure 9: SWOT Analysis

8. Conclusion

Hence, in this way, we made the Nepali Audio Signal Synthesizer. The output of the system is satisfactory. There are still many doors open for further enhancement of it. We hope our contribution in this project will have meaningful effect in the field of Nepali Text to Speech and Natural Language Processing.

9. Reference

- Taylor, P. (2007), Text-to-speech-synthesis
- CH EN, H. W., 2018, Text to speech Synthesis
- Andrew J. Hunt et. al. (2016), Unit selection in a concatenative speech synthesis system using a large speech database
- Ng., A. (n.d.). *Machine Learning.* Retrieved 03 01, 2018, from Coursera: https://www.coursera.org/learn/machine-learning?authMode=login&errorCode=invalidCredential
- https://en.wikipedia.org/wiki/Speech_synthesis Retrieved 03 01, 2018