# **ACKNOWLEGDEMENT**

We would like to extend our gratitude towards Department of Computer Engineering, Nepal Engineering College and Pokhara University for providing a platform, opportunity to undertake an exclusive Major project.

We would like to thank all the members in Department of Computer Engineering and Project Coordinator along with all the teachers for providing help in the every step during the development of the project.

Furthermore, we would like to give special thanks to our Project Supervisor Mr. Krishna B. Shah for providing knowledge, support, help and encouragement in accomplishing the task that are needed for the development of our Major project.

Thank you, everyone.

Ashmita Ghimire (011-302)

Kiran Kumar Chaudhary (011-311)

***nec***

April, 2016

# **ABSTRACT**

This project report confer the tools and methodology used in developing a Nepali Text to Speech Synthesis System using FreeTTS and is written entirely in the Java programming language using FreeTTS synthesizer. Vocalized form of human communication is Speech. Here the Nepali Language is Synthetized based on formant approach and the use of most popular generic frameworks FreeTTS that is available in public domain for the development of a TTS synthesizer. The Text To Speech Architecture putting more emphasis on a components, namely Natural Language Processing (NLP) rather than Digital Signal Processing (DSP) component. Nepali language being mostly used language in Nepal and some parts of India and abroad, a text-to-speech (TTS) synthesizer for this language will prove to be a useful Information and communication technology (ICT) based system to aid to those majorities of people who are illiterate and also to those who are physical impairments like visually handicapped and vocally disabled physically handicapped. This ability to convert text to voice may reduce the dependency, frustration, and sense of helplessness of these people. The system can be extended to include more features such as more emotions, improved tokenization, interactive options and the use of minimal database.

Keywords: TTS, Prosody, Di-phone, Phoneme, Concatenation, Speech Synthesizer, Nepali vowel and consonants, Speech.

# **List of Abbreviations**

|  |  |
| --- | --- |
| **Abbreviation** | **Explanation** |
| TTS | Text To Speech |
| OCR | Optical Character Recognition |
| NLP | Natural Language Processing |
| POS | Part of Speech |
| NeLRaLEC | Nepali Language Resources and Localized for Education and communication |
| ICT | Information and Communication Technology |
| WAV | Waveform |
| ISO | International Organization for Standardization |
| TD-PSOLA | Time Domain- Pitch Synchronous Overlap Address |
| MOS | Mean Opinion Score |

# 

# **List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No.** |
| Fig 1.1 | Schematic Representation of text to speech (TTS) system | 2 |
| Fig 1.2 | A simple functional diagram of TTS system | 3 |
| Fig 3.1 | System Overview of TTS | 17 |
| Fig 3.2 | Basic Flow Diagram of Nepali TTS synthesis System | 20 |
| Fig 4.1 | Flowchart of Nepali TTS | 24 |
| Fig 4.2 | DFD-level 0 | 25 |
| Fig 4.3 | Use Case Diagram | 26 |
| Fig 4.4 | UML Class Diagram | 27 |
| Fig 5.1 | Evaluation Chart of Vowel, Consonant, and Number | 32 |
| Fig 5.2 | Pie Chart Representation of Accuracy of Pronunciation | 33 |
| Fig 5.3 | Evaluation Chart of Words | 34 |
|  |  |  |

# **List of Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Title** | **Page No.** |
| Table 4.1 | List of phone set | 22 |
| Table 5.1 | List of Numbers, Vowel, and Consonant used in test cases | 30 |
| Table 5.2 | Evaluation Table for Numeric Keywords | 31 |
| Table 5.3 | Evaluation Table for Vowel Keywords | 31 |
| Table 5.4 | Evaluation Table for Consonant Keywords | 32 |
| Table 5.5 | Evaluation Table for Individual Words | 33 |
| Table 5.6 | Weight of Words in each Categories | 34 |
| Table 5.7 | Evaluation of Total Words for Accuracy | 35 |

# **List of terminology**

**Speech synthesis terminology:**

Before going on to detail to explain technology of speech synthesis systems general terms used in the domain are briefly explained here.

* **Utterance:**  A spoken word, statement, or a vocal sound or the action of saying or expression something aloud i.e. How a sentence is spoken.
* **Prosody:** The pattern of rhythm and sound. In phonetics, the use of pitch, loudness, tempo, and rhythm in speech to convey information about structure and meaning of an utterances i.e. Way things are spoken usually a word.
* **Words:** It consist of one or more syllables.
* **Units:** It can be phones or phonemes or syllable which is treated as concatenating unit by the speech synthesis system.
* **Phones:** Lingual phone gets generated from the flow of air through the vocal cord or those sounds which get generated form the vocal cord due to the flow of air through it are called lingual phones. Lingual phones are those meaningful sounds that are generated by various vocal organ of the human. These sounds cannot be broken into further smaller units.
* **Phonemes:** It is the smallest unit of the spoken language .Every language has their own smallest units. One or more phones make the phoneme. Phoneme has no meaning in itself but determines the meaning of the word i.e. group of phoneme form the word. In some languages like Nepali, English, Hindi phonemes are divided into two major groups i.e. vowels and consonants. Vowels are generated when air flows smoothly through vocal track without any constriction or disturbance in the vocal track, whereas consonants are generated by disturbance to the air flowing through the vocal track.
* **Syllables:** It is a unit of organization for sequences of speech sounds or it is a sound of vowel (a, e, i, o, u) that is created when pronouncing the letter. No of times you hear the sound of a vowel is no. of syllable in word. Group of phonemes form syllable. For Nepali language Syllable is called (अकर: e.g. क, ख, अ, आ, कक etc.). Group of syllable determines the prosody of the word and prosody of all word combined determines the utterance structure of the sentence.
* **Syllabication:** The syllabication algorithm breaks a word such that there are minimum numbers of breaks in the word, as minimum number of joins will have fewer artifacts. The algorithm dynamically looks for polysyllable units making up the word, cross checks the database for availability of units, and then breaks the word accordingly. If polysyllable units are not available, then algorithm naturally picks up smaller units. This mean, if database is populated with all available phones of language along with syllable units, algorithm falls back on phones if bigger units are not available. A basic syllable types are: V, VC, CV, VCC, CVC, CCVC and CVCC etc. where V and C represent vowel and consonant respectively that are used for languages. There are twelve vowel found in Devanagari language, the 12 Devanagari vowels. Devanagari script has about 36 consonants.
* **Di-phones:** It is formed by two or more phones, it is similar to syllables.
* **Phonetics:** In most languages the written text does not correspond to its pronunciation so that in order to describe correct pronunciation some kind of symbolic presentation is needed. Every language has a different phonetic alphabet and a different set of possible phonemes and their combinations. The number of phonetic symbols is between 20 and 60 in each language. A set of phonemes can be defined as the minimum number of symbols needed to describe every possible word in a language. In English there are about 40 phonemes .In Nepali language there are about 50 phonemes. Due to complexity and different kind of definitions, the number of phonemes in Nepali and most of the other languages cannot be defined exactly. Phonemes are abstract units and their pronunciation depends on contextual effects, speaker's characteristics, and emotions. During continuous speech, the articulatory movements depend on the preceding and the following phonemes. The articulators are in different position depending on the preceding one and they are preparing to the following phoneme in advance this cause smooth transition between phonemes which is hard to achieve in any kind of synthesis especially in the concatenative synthesis. This also causes some variations on how the individual phoneme is pronounced. These variations are called allophones which are the subset of phonemes and the effect is known as co-articulation. For example, a word रल contains a light र and सर contains a distinct र. These र are the same phoneme but different allophones and have different vocal tract configurations. Another reason why the phonetic representation is not perfect, is that the speech signal is always continuous and phonetic notation is always discrete. Different emotions and speaker characteristics are also impossible to describe with phonemes so the unit called phone is usually defined as an acoustic realization of a phoneme. The phonetic alphabet is usually divided in two main categories, vowels and consonants
* **Intonation:** It is the variation of spoken pitch that is not used to distinguish words i.e. it is used for a range of functions such as indicating the attitudes and emotions of the speaker, signaling the difference between statements and questions, focusing attention on important elements of the spoken message and also helping to regulate conversational interaction. It consist of rising intonation (yes/no questions), rising falling intonation, low or falling intonation (who, what, when, why, and how), fall-rising tone (expresses uncertainty).
* **Rhythm:** a strong, regular repeated pattern of movement or sound. In phonetic, the sense of movement in speech, marked by the stress, timing, quality of syllable (the beat in poem).
* **Stress:** In phonetics, the degree of emphasis given to a speech i.e. for sound or syllable in speech.
* **Vowel:** They are the sound in a spoken language, pronounced with an open vocal tract so that there is no build-up of air pressure at any point above the glottis. Or Vowels are always voiced sounds and they are produced with the vocal cords in vibration. Vowels have considerably higher amplitude than consonants and they are also more stable and easier to analyze and describe acoustically.
* **Consonants:** Consonants may be either voiced or unvoiced and they are more difficult to synthesize properly because consonants involve very rapid changes.
* **Velar:** Pronounced with the back of the tongue near the soft palate for e.g. ka, kha, ga, gha.
* **Retroflex:** Pronounced with the tip of the tongue curled up towards the hard palate for e.g. ta, tha, da, dha.
* **Alveolar:** Pronounced with the tip of the tongue on a near this ridge for e.g. na, cha, cha, ja, jhha.
* **Bilabial:** sound formed by closure or near closure of the lips for e.g. pa, pha, ba, bha.
* **Rhotic:** r is pronounced before a constant and at the ends of words for e.g. ra.
* **Glottal:** Pronounced by the glottis for e.g. ha, ह.