Project Report on Building Natural Sounding Nepali Speech Synthesis System

Roop Shree Ratna Bajracharya

2018-04-15

Contents

Abstract

Text-to-Speech (TTS) synthesis has come far from its primitive synthetic monotone voices to more natural and intelligible sounding voices. Festival is one of such systems that uses a concatenative speech synthesis method to produce natural sounding voice. This project aims to develop Nepali voice using Festival system with tools included in Festvox as a part of the Nepali-TTS project currently being conducted in Information and Language Processing Research Lab (ILPRL) in Kathmandu University. This project includes studying and improving different steps and procedures involved in the speech synthesis process and will aim to produce more natural sounding Nepali voice.

Keywords: Text-To-Speech, Festival Speech Synthesis, Concatenative Speech Synthesis

Chapter 1

Introduction

1.1 Text to Speech Synthesis

Text-To-Speech (TTS) system converts the given text to a spoken waveform through text processing and speech generation processes. These processes are connected to linguistic theory, models of speech production, and acoustic-phonetic characterization of language [1]. There are three approaches to TTS system: 1) articulatory-model 2) parameter-based and 3) concatenation of stored speech [1]. This project deals with the third approach to concatenate stored speech segments (units). This approach includes storing and concatenation of speech from huge number of recorded sentences of a language to produce a speech waveform.

1.2 Concatenative Speech Synthesis

A concatenative speech synthesis system produces sound by combining recorded sound clips. The system can speak by using available recordings of sound. The recorded speech can be further broken down into smaller components, so that it can be used to make any spoken word. There are three main types of concatenative synthesis [2]:

1. Unit selection synthesis that uses large databases of recorded speech and creates database from recorded utterance.

2. Diphone synthesis that uses a minimal speech database containing all the diphones (sound-to-sound transitions) occurring in a language.

3. Domain-specific synthesis which concatenates prerecorded words and phrases to create complete utterances. This approach is useful for limited domain applications like talking clocks, station announcement systems which can create large variations of speech by using only small number of unique recorded speech.

This project makes use of the unit selection method.

1.3 Festival Speech Synthesis System

Festival is a free software multi-lingual speech synthesis multi-platform workbench developed by Alan W. Black at Centre for Speech Technology Research (CSTR) at the University of Edinburgh [3]. It designed as a component of large speech technology systems. It provides an environment for investigating and developing speech synthesis techniques. It is a concatenative speech synthesis system. The system is written in C++ with a Scheme-like command interpreter for general customization and extension (CSTR) [3]. This allows the user to add or edit its components in Scheme without touching the underlying C++ code.

1.4 Festvox for Building Voices

The Festvox provides a systematic and a well documented way to build new synthetic voices. It consists of documentation and scripts that explains the steps and theory behind building new voices for speech synthesis using Festival speech synthesis system. It also provides full examples for building synthetic voices using limited domains. It supports data-driven synthesis algorithm known as unit selection algorithm [4].

Festvox can be used with Edinburg Speech Tools (EST). The Edinburgh Speech Tools Library contains a range of tools for general speech processing, written at the Centre for Speech Technology Research at the University of Edinburgh [5].

Chapter 2

Objective

This project aims to build natural sounding Nepali voices using Festival system which will be later used in the free Windows screen reader application called Non-Visual Desktop Access (NVDA). The voice build from this system can be used as a plugin in NVDA which will allow the system to produce natural sounding screen reader voices. So this project will look into the different factors associated with the Nepali voice synthesis including the language related as well as the technical aspects of the Festival system using Festvox tools while improving the synthesized voice.

A secondary aim of this project is to learn about the processes involved Festival in taking text as input and producing audio output and make improvements to the process to produce better quality sound.

Chapter 3

Related Work

3.1 Current Speech Synthesis System for NVDA

NVDA supports a number of speech synthesis including eSpeak, Acapela TTS, Eloquence, Nuance, Festival, Flite. etc. [6]. Microsoft also has its own speech synthesis system called Microsoft Speech API (SAPI) which is supported by NVDA. But out of these only Festival is open-source and free voice. Festival Speech Synthesis System and FestVox was created in *Center for Speech Technology Research (CSTR)* located at the University of Endinburgh which uses the concatenative approach of speech synthesis. Festival is widely used among researchers in the field of speech synthesis research. [7]

Chapter 4

Voice Synthesis Process

The voice synthesis in festival includes a number of step ranging from designing of input prompts to finally testing the input text for its synthesized voice. Figure 1 shows the major components of this process.

The steps for voice synthesis are as follows [2]:

1. Data Collection: Data to be used for recorded are obtained by crawling new websites for most frequent sentences.
2. Cleaning up of Data: The collected data is analysed and cleaned to cover most relevant sentences. Then the final set of selected sentences are compiled.
3. Recording: The selected sentences are recorded in a noise-free studio for maximum clarity.
4. Text Preprocessing: The selected text which is in Devnagarik unicode is transliterated to corresponding English version. Then these transliterated text are syllabified so that it can be supplied to create lexicon files need for the Festival system.
5. Labeling: The wavefiles are then auto labeled using the EHMM (Hidden Markov Model) labeler to get accurate syllable boundaries. The labeling step includes building utterance structures, building pitch marks and cepstrum parameter files.
6. Training Phase: Using the wavefiles and their transcriptions the nepali language unit selection voice is built.
7. Testing Phase: Using the voice built, Festival system is used to test the input text.
8. Miscellaneous Processes: Apart from the above main steps the project also includes writing scripts to automate repetitive tasks and also to automate the voice building process.

Chapter 5

Unit Selection Databases

The process of unit selection includes the selection of a unit of speech which may range from syllable to diphone (in this case phone) from the array of available units [8]. The database has a range of the same unit of which an appropriate one is selected by applying proper mechanism of selection during runtime. The mechanism may include clustering acoustically similar units based on features like phonetic context, prosodic features, stress and accents together [8].

Chapter 6

Conclusion

This project aims to develop Nepali voice in Festival Speech Synthesis System using Festvox tools under the unit selection synthesis approach of concatenative speech synthesis technique. Through this project various theory related to TTS and Festival system will be studied and analysed with the object of improving the output and process of synthesizing the Nepali voice using the given resources.

Bibliography

[1] D. H. Klatt, "Review of texttospeech conversion for english," 1987.

[2] R. S. S.P. Kishore and M. Srinivas, "Building hindi and telugu voices using festvox," 2002.

[3] T. U. o. E. The Center for Speech Technology Research, "Edinburgh speech tools." `http://www.cstr.ed.ac.uk/projects/festival/`. Accessed on 2018-04-06.

[4] K. A. L. Alan W. Black, "Optimal data selection for unit selection synthesis," 2001.

[5] T. U. o. E. The Center for Speech Technology Research, "The festival speech synthesis system." `http://festvox.org/docs/speech_tools-2.4.0/estlicence.html`. Accessed on 2018-04-06.

[6] M. Curran, "Extra voices for nvda." `https://github.com/nvaccess/nvda/wiki/ExtraVoices`, Feburary 2018. Accessed on 2018-04-07.

[7] M. Hood, "Creating a voice for festival speech synthesis system," 2004.

[8] K. A. L. Alan W. Black, "Building synthetic voices," 2014.