GSoC 2016 PROPOSAL
Language model and Acoustic model for Malayalam language for speech recognition system in CMU Sphinx

Personal Information:

Name:	Sreenadh T C
e-mail:	kesav.tc8@gmail.com
Blog url:	http://www.teczoz.com
Freenode IRC nick:	SREENADH
University and Current Education:	College of Engineering Trikaripur, affiliated to CUSAT

I enjoy volunteering and supporting open source development. I came across the works of SMC and straight away realized the importance of developing and/or supporting in any manner I can towards Malayalam. I then came to know about Indic project which supports development and research towards Indian languages. I find this opportunity to work with Indic project as a golden one and the one that can help me grow better in the developer community and at the same time contribute more towards open source community.
This is my first time with GSoC but wished to do last year with SMC.
I am ready and happy to work full time during the summer.
I consider GSoC just another huge platform to contribute and support projects of Indic Project. So I am always happy and willing to contribute even after GSoC 2016 in areas related to Malayalam language support.
The amount of projects during my course has helped me build myself as a better programmer. I have always stayed curious about new ways of solving issues and always enjoy while doing so.
I have already completed the very same project for my S8 Major Project in which I was successful in creating Language Model as well as Acoustic model for approximately 100 words. (Only due to project completion limits) with acceptable accuracy. So am confident that this idea can be taken forward by me and complete a full-fledged LM and AM for Malayalam.

Proposal Description:

Speech is a complex phenomenon. People rarely understand how is it produced and perceived. The naive perception is often that speech is built with words, and each word consists of phones. The reality is unfortunately very different. Speech is a dynamic process without clearly distinguished parts. It's always useful to get a sound editor and look into the recording of the speech and listen to it.

The aim of the project is to develop a large vocabulary ***Statistical Language model for Malayalam Language*** (or in other words, its a probabilistic distribution over a sequence of Malayalam words to understand and efficiently tune the occurrences of Malayalam words.)

The technique that is planned to be used to develop the Language model is : *N-Gram model and variants technique*.

Also, the project aims at developing a corresponding ***Acoustic model for Malayalam*** by collecting various speech samples and training it using CMU SphinxTrain libraries. These two along with a Phonetic Dictionary for Malayalam can then later be used to develop applications that seeks the help of Malayalam Speech Recognition using the CMU Sphinx libraries.

All modern descriptions of speech are to some degree probabilistic. That means that there are no certain boundaries between units, or between words. Speech to text translation and other applications of speech are never 100% correct. That idea is rather unusual for software developers, who usually work with deterministic systems. And it creates a lot of issues specific only to speech technology.

This idea was already tried for a word count of ~100 as a part of my course project. Even though this is not in any manner perfect, I am now well aware of the process and the way to do it in a better way.
Link to the current work I have done: https://github.com/sreecodeslayer/ml-am-lm-cmusphinx

The process I followed for my initial trial of the idea during my Major project:

1. Use The Datuk corpus as the source for Malayalam corpus (http://olam.in/open/datuk/)
2. Build a Phonetic dictionary using this corpus
3. Develop unigram, digram, and trigram Statistical language model using CMU-CSLM toolkit.
4. Training phase and acoustic model development using five different people.
5. Testing for accuracy, could only achieve roughly 60% due to time limitations for the project submission. I strongly believe this can be improved more this time.

I have communicated with Mrs. Deepa Gopinath (deepagopinath on irc.freenode.net) who is the listed mentor for the project.
My other works can be found in the same GitHub profile.