PROJECT PROPOSAL
Speech-to-text with publicly available deep learning models

Student Details

Name: Birat Datta
Email ID: birat.datta26@gmail.com
Contact Number: +916009344713
GitHub: https://www.linkedin.com/in/biratdatta/
LinkedIn: https://www.linkedin.com/in/biratdatta/
Geographic Location: Bangalore, India
Timezone: IST (GMT + 0530)
Primary Language: English

Student Affiliation

Institution: Jain Deemed-to-be University, Bangalore
Program: Bachelor of Technology in CSE with Specialization in Artificial Intelligence & Machine Learning
Stage: 1st Year (2nd Semester)
Contacts to Verify: Dr. V. Vivek (Program Coordinator) [v.vullikanti@jainuniversity.ac.in]

Student Bio

My name is Birat and I am a student at Jain Deemed-to-be University in Bangalore, India. I'm now studying Bachelor of Science in Computer Science and Engineering with a focus on AI and ML. I previously competed in the Smart India Hackathon and was nominated for the finals for pitching an App using a material design UI.

Throughout the Hackathon, I worked on the prototype app in Java and created UI mock-ups in Adobe Xd. I'd want to contribute to the Jitsi, and GSoC would be an excellent chance for me to do so. I've been interested in JavaScript and HTML for quite some time. I've also been honing my Java abilities, as well as those of other coding languages like C, Kotlin, and Python.

I have quite a good hold on Adobe Xd, and Adobe Dreamweaver, and a good understanding on making materialistic UI using Angular and Android.

I have acquired a deep interest in UI designing, AIML, Open Source Organizations and its tools would assist me a lot during my path, it would be my joy if I could study and progress under your valuable direction for an acclaimed company like The Jitsi.

I have hands-on expertise with UI design and its potential. I have previously worked on data visualisation, data transformation, and designing simple and straightforward UI/UX for ordinary people to grasp. As an AIML student, I am interested in Deep Learning and Statistical Mathematics.

Schedule Conflicts:

I have no other commitments for the summer other than GSoC, therefore there should be no conflicts in my calendar, and I guarantee availability for the needed number of hours in GSoC 2022. I can devote more than 30 hours per week to the project's completion.

Project Info

Project Title: Speech-to-text with publicly available deep learning models

Project Synopsis:

The speech-to-text feature for the Jitsi Meet aims now to create an in-house solution for the earlier Google or IBM's Cloud based Speech to text API's.

This is done with the understanding that the processing power necessary to train the models and the requisite engineering to provide transcriptions in a Jitsi Meet Conference Call will be limited.

URL of Project Idea Page:

https://github.com/jitsi/gsoc-ideas/blob/master/2022/speech-to-text.md

Mentors:

Nik Vaeseen

Have you ever been in touch with your mentor? When and How?

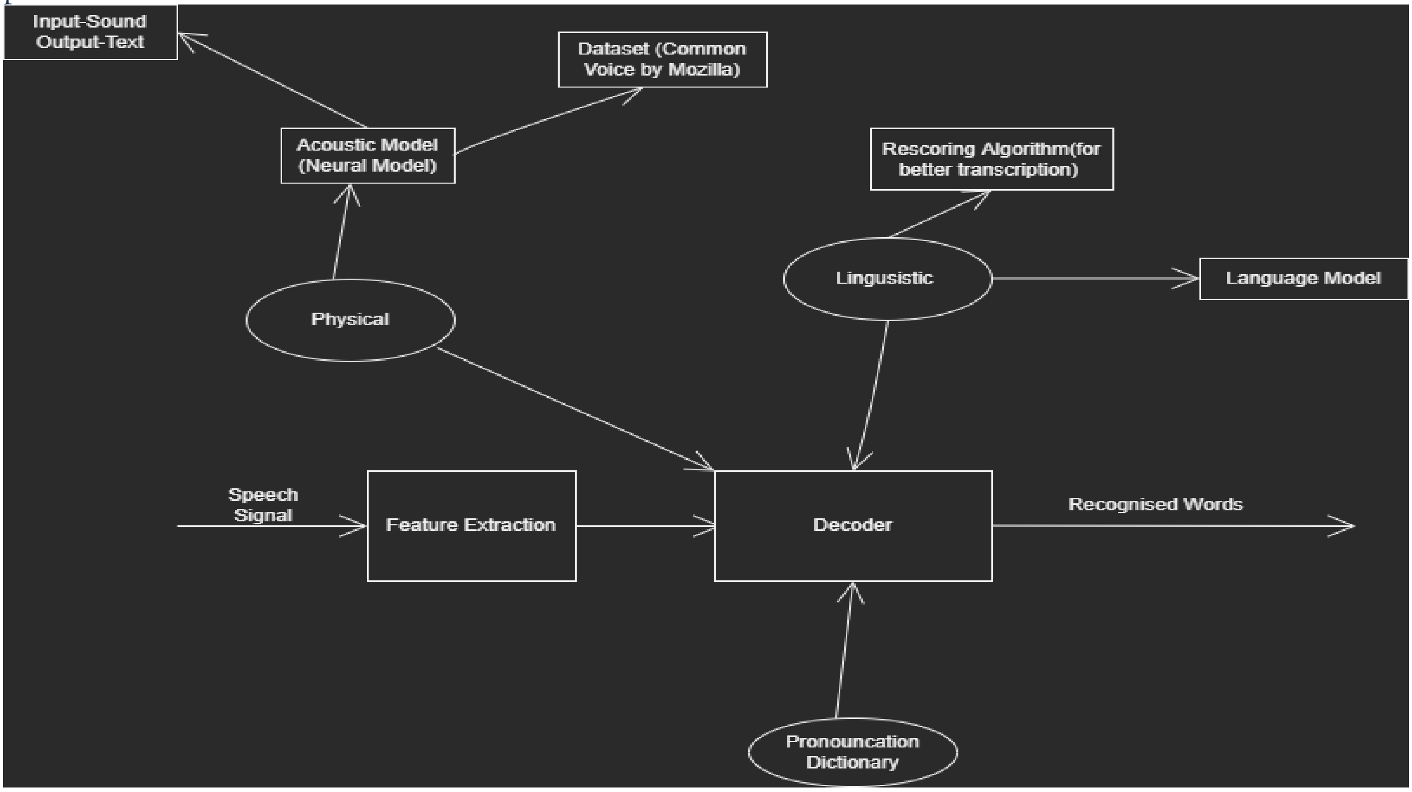
Currently No.

Project Details

Introduction

Speech(noun. the expression of or the ability to express thoughts and feelings by articulate sounds) is different in humans ranging from gender, age, accents or pronunciations and speaking in different styles and at different rates and in different emotional states. The presence of environmental noise, reverberation, different types of microphones and recording devices can results in additional variability. The main goal of the project is to implement a open-source speech-to-text model and to implement the communication between the sever and Jitsi Meet.

There is a use case diagram that will properly depict the way we will solve this problem.



Creating A Neural Network

Acoustic Model

Speech is a naturally occurring Time Sequence , for that we need a neural network that can process sequential data. **Recurrent Neural Networks (RNN)** are lightweight and can process sequential data for smaller network size. For speech data set we are using **Common Voice by Mozilla** for the nuances of speech to counter the age, gender, types of microphone etc .

Linguistic Model

The output from the Acoustic Model will have linguistic mistakes, which will be solved by language model by building a probability distribution over sequence of words it will be trained on. For our usage we are using **KenLM** , and for a rescoring model we are using **ctcdecode** that will rescore the outputs for better transcriptions.
https://github.com/kpu/kenlm
https://github.com/huggingface/transformers
https://github.com/parlance/ctcdecode

Benefits To Community

This Model which we will be building will be building during the summer , will be open sourced and will be available to anyone. The Models which I will be using are going to focus on certain key areas:-

- Low Compute Power
- Can be easily implemented in the backend
- Open Sourced Databases

Our Video Conferencing app is open sourced and people can transcribe their audio while speaking in scenarios such as office meetings, classes etc it will be easier for people to take notes on the go.

Time Line:

Having an organized schedule is the most important factor when it comes to handling a project as it increases productivity and allows us be consistent.
I have divided the project timeline into 5 Phases, Each phase contains **2 weeks** each and it's respective deadlines.
I understand the valuable time that the mentor's volunteer for the project, hence the weekly discussions and the evaluations that would be required will be done as and when they are most suited and comfortable according to them.

Community Bonding:

May 20, 2022 - June 12 , 2022
print("Hello World to Jitsi Community)

- I would use this time frame to get to know more about the Jitsi Community & it's culture.
- Interact with my peers & mentors and trying to understand different work environments
- Get ready for the project by bridging up for any required skill gap.

Phase I:

June 13, 2022 - June 27, 2022

- Data processing pipelines to transform Audio waves into mel spectograms as features to feed it into the neural network
- Transform our character labels into integer indexes. Our Neural Model will output characters instead of words.
- Augment our Data to effectively have a bigger dataset.

Phase 2:

June 27, 2022 - July 11, 2022

- CNN to extract data from the mel spectogram and also reduce the time dimensions of the data
- For RNN we are using Long Short Term Memory(LSTM) Neural Network takes the data from the previous layer and step-by-step produces.
- AI Dense Layer with Softmax Activation will act as a classifier that takes the mn's output snf predicts character probabilities for each time step.

Phase 3:
July 11, 2022 - July 25, 2022

- Adding Layer Normalization and Gelu activation between each layers
- Dropout added after each layer to ensure the network is more generalizable and robust to real world layer.

Phase 4:
July 25, 2022 - August 8, 2022

- Training the Model we will use pytorch learning and CTC loss Function to train speech recognition Model
- Training the Model for the rest of the days to get better results

Phase 5:
August 8 , 2022 - August 22, 2022

- Implementing it to the server to run with Jitsi Meet
- Testing for the rest of the model and checks before pushing it into the master repo

Wrapping Up:
August 22, 2022 - September 12, 2022

- Finalizing the Code Checks and Documentations.
- Getting Final reviews and checks from fellow peers
- Submit to google Team.

Related Work

Since I'm from a computer Science background , I started researching on unfamiliar terms and concepts related to Deep Learning and Neural Networks. I have read 4 related research papers and studied some similar projects on the web.
I researched on potential Linguistic Models along with Packages which will allow us to easily obtain the required materials to work with. I also acquired some knowledge on the Dataset which we will use along with some dataset which we will use to make the algorithm stronger.

Why Choose Me?

As mentioned earlier, I have experienced with working on complex datasets and python programing language and I have a deep hold on statistical mathematics which includes but not limited to:-

- Probability definitions and properties
- Common discrete and continuous distributions
- Bivariate distributions

- Conditional probability,Combinatorics and basic set theory notation
- Random variables, expectation, variance
- Univariate and bivariate transformations
- Convergence of random variables: in probability, in distribution, almost sure
- Central Limit Theorem, Laws of Large Numbers
- Estimation: bias, MSE, consistency, sufficiency, maximum likelihood, method of moments, UMVUE, Rao-Blackwell Theorem, Fisher Information
- Hypothesis testing: significance level and power, Neyman-Pearson lemma, Likelihood ratio tests
- Confidence Intervals: definitions, duality with hypothesis tests.

After GSoC

GSoC would be an initial part of the journey, but I believe open source is a never ending wonderful & insightful journey.
I will continue maintaining and developing what we would have started and also continue upgrading the limited dataset support to different varieties of data which would involve all of the data around the world & it's different categories. I'll also be looking for more efficient deep learning algorithms for upgrading our computing skills over time as it is a rapid developing field.